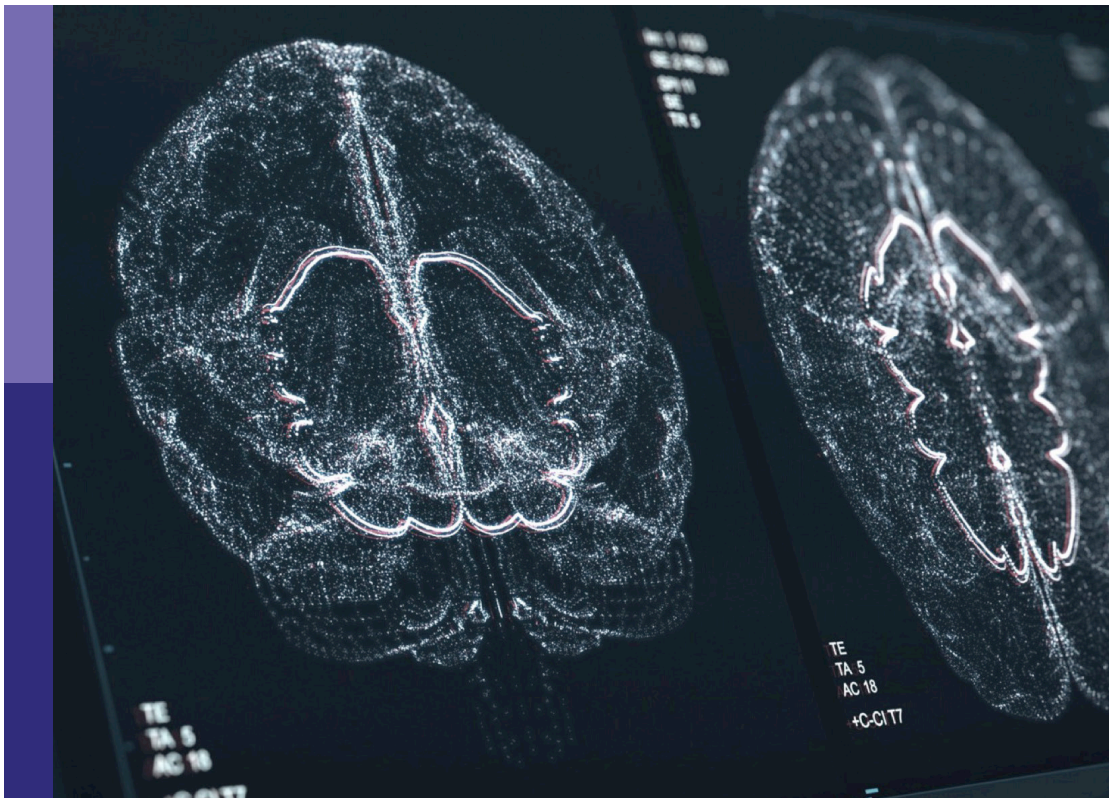
frontiers | Research Topics

Navigating the landscape of FAIR data sharing and reuse: Repositories, standards, and resources

Edited by
Maaike M. H. Van Swieten and Christian Haselgrove

Published in
Frontiers in Neuroinformatics



ISSN 1664-8714
ISBN 978-2-8325-4593-5
DOI 10.3389/978-2-8325-4593-5

About Frontiers

Frontiers is more than just an open access publisher of scholarly articles: it is a pioneering approach to the world of academia, radically improving the way scholarly research is managed. The grand vision of Frontiers is a world where all people have an equal opportunity to seek, share and generate knowledge. Frontiers provides immediate and permanent online open access to all its publications, but this alone is not enough to realize our grand goals.

Frontiers journal series

The Frontiers journal series is a multi-tier and interdisciplinary set of open-access, online journals, promising a paradigm shift from the current review, selection and dissemination processes in academic publishing. All Frontiers journals are driven by researchers for researchers; therefore, they constitute a service to the scholarly community. At the same time, the *Frontiers journal series* operates on a revolutionary invention, the tiered publishing system, initially addressing specific communities of scholars, and gradually climbing up to broader public understanding, thus serving the interests of the lay society, too.

Dedication to quality

Each Frontiers article is a landmark of the highest quality, thanks to genuinely collaborative interactions between authors and review editors, who include some of the world's best academicians. Research must be certified by peers before entering a stream of knowledge that may eventually reach the public - and shape society; therefore, Frontiers only applies the most rigorous and unbiased reviews. Frontiers revolutionizes research publishing by freely delivering the most outstanding research, evaluated with no bias from both the academic and social point of view. By applying the most advanced information technologies, Frontiers is catapulting scholarly publishing into a new generation.

What are Frontiers Research Topics?

Frontiers Research Topics are very popular trademarks of the *Frontiers journals series*: they are collections of at least ten articles, all centered on a particular subject. With their unique mix of varied contributions from Original Research to Review Articles, Frontiers Research Topics unify the most influential researchers, the latest key findings and historical advances in a hot research area.

Find out more on how to host your own Frontiers Research Topic or contribute to one as an author by contacting the Frontiers editorial office: frontiersin.org/about/contact

Navigating the landscape of FAIR data sharing and reuse: Repositories, standards, and resources

Topic editors

Maaike M. H. Van Swieten — Netherlands Comprehensive Cancer Organisation (IKNL), Netherlands

Christian Haselgrove — UMass Chan Medical School, United States

Citation

Van Swieten, M. M. H., Haselgrove, C., eds. (2024). *Navigating the landscape of FAIR data sharing and reuse: Repositories, standards, and resources*. Lausanne: Frontiers Media SA. doi: 10.3389/978-2-8325-4593-5

Table of contents

frontiers | Frontiers in Neuroinformatics

TYPE Editorial
PUBLISHED 01 March 2024
DOI 10.3389/fninf.2024.1387758

OPEN ACCESS

EDITED AND REVIEWED BY
Sean L. Hill,
Krembil Centre for Neuroinformatics, CAMH, Canada

*CORRESPONDENCE
Maaike M. H. van Swieten
mvanswieten@outlook.com

RECEIVED 18 February 2024
ACCEPTED 20 February 2024
PUBLISHED 01 March 2024

CITATION
van Swieten MMH and Haselgrove C (2024) Editorial: Navigating the landscape of FAIR data sharing and reuse: repositories, standards, and resources.
Front. Neuroinform. 18:1387758.
doi: 10.3389/fninf.2024.1387758

Editorial: Navigating the landscape of FAIR data sharing and reuse: repositories, standards, and resources

Maaike M. H. van Swieten[1]* and Christian Haselgrove[2]

[1]Netherlands Comprehensive Cancer Organization (IKNL), Utrecht, Netherlands, [2]UMass Chan Medical School, Worcester, MA, United States

KEYWORDS

FAIR principles, neuroinformatics, data sharing, data reuse, repositories, standards

Editorial on the Research Topic
Navigating the landscape of FAIR data sharing and reuse: repositories, standards, and resources

In response to the expanding landscape of neuroscience data and the diverse array of formats emerging from various research communities, the scientific community faces a pressing challenge in traditional data management, sharing, and mining methods. The push for data sharing mandates and the increasing demand for open data utilization (e.g., National Institutes of Health, 2023) have prompted the evolution of sophisticated methodologies and tools. These advancements aim to empower researchers in effectively exploring, mining, and integrating datasets. However, the growing number of resources in this rapidly evolving field poses a substantial hurdle for researchers attempting to navigate this complex landscape.

As the scientific community strives to uphold data sharing mandates and embrace open data principles, it becomes imperative to equip researchers with the awareness and knowledge necessary to navigate this landscape successfully. This Frontiers in Neuroinformatics Research Topic was designed to showcase exciting recent developments in the field and offer a nuanced overview of available resources, with a focus on ensuring that data are findable, accessible, interoperable, and reusable—adhering to the FAIR principles (Wilkinson et al., 2016).

The Research Topic reflects the broad extent of the FAIR landscape. While the FAIR principles apply directly to data, repositories, and standards such as ontologies, satisfying the full intent of the FAIR principles often requires more diverse considerations, as exemplified here: from atlases and software to workflows and even data governance.

A comprehensive overview of the components and practices required to achieve FAIR in neuroscience along with the perspectives on the past, present and future of a FAIR infrastructure for neuroscience, are provided in the review article by Martone. This article also compares large next-generation neuroscience infrastructures, including EBRAINS, CONP, SPARC, DANDI, Open Neuro, and BRAIN/Minds.

This Research Topic also features four articles about FAIR repositories, namely BrainCODE for general neuroscience data, COINSTAC Vaults and Image and Data Archive (IDA) for neuroimaging data, and GAAIN, DPUK, ADDI for Alzheimer's and dementia-related data. Each article delves into the challenges and solutions related to making the

repository FAIR, the governance and sovereignty concerns, and the steps taken to enhance the user experience. These repositories mainly vary in their technical implementations and mechanisms for managing data governance requirements and sovereignty, particularly for datasets containing sensitive or personal data, which require specific permissions.

COINSTAC, for example, addresses challenges through federated analysis, enabling researchers to analyze datasets without public data sharing (Martin et al.). The introduction of COINSTAC Vaults (CVs) enhances this capability by providing standardized, persistent datasets that seamlessly integrate with COINSTAC's federated analysis. CVs offer a user-friendly interface, promoting self-service analysis and filling a crucial gap in the data sharing ecosystem.

Other platforms like DPUK, GAAIN, and ADDI rely on two core design principles, such as "trust-by-design" and "data federation", actively developing a range of innovative solutions to enhance large-scale data access (Toga et al.). This includes simplifying stakeholder involvement through streamlined data sharing agreements, introducing decentralized data sharing solutions, and establishing universally accessible analysis through workspaces and containerized software.

The IDA, run by the Laboratory of Neuro Imaging, presents an alternative approach to managing and reusing multi-center data (Neu et al.). Serving as a central hub for collaborative groups, it facilitates data transfers and offers a suite of informatics tools. These tools are designed to support in various tasks, including de-identifying, integrating, searching, visualizing, and sharing a diverse range of neuroscience data. Researchers maintain full control over the data stored in the IDA, benefiting from a reliable infrastructure that safeguards and preserves research data.

Brain-CODE, a large-scale neuroinformatics platform, supports the collection, storage, federation, sharing and analysis of different data types across different types of brain disorders. Behan et al. discuss the data sharing processes on Brain-CODE, aligning them with the FAIR principles. Brain-CODE not only provides extensive metadata for interactive searches and the ability to generate subsets of data, but also focuses on mechanisms and services that facilitate interoperability and the combination of data using advanced privacy preserving record linking and homomorphic encryption. Sensitive data can be accessed within a secure workspace on Brain-CODE, and public datasets can be exported to a locally device.

Currently, repositories predominantly address data governance concerning data derived from human subjects. However, there is a noticeable absence of regulatory frameworks for non-human data, despite divergent legal and ethical principles across countries about the generation of animal data. Eke et al. advocate for the establishment of animal data governance, proposing to delineate and collect metadata related to ethical considerations. This proposal aims to enhance data transparency and promote the FAIR principles within the context of animal research.

Despite the growing number of datasets on repositories mentioned above, a considerable amount of data remains underutilized and inaccessible, especially smaller-sized datasets. This is often attributed to the quality of the associated metadata and the degree of annotations. The NeuroBridge platform (Wang et al.) and the NeuroBridge Ontology (Sahoo et al.) offer innovative approaches for extracting metadata related to study design and data collection from full-text papers through ontology developments and machine-learning-based natural-language processing. By harnessing the search capabilities of the NeuroBridge platform, researchers can pinpoint neuroimaging datasets tailored to their specific research questions, thereby promoting data reuse.

Queder et al. propose an alternative method for standardizing and annotating neuroimaging datasets. Neuroimaging datasets are typically organized with the Brain Imaging Data Structure (BIDS) (Gorgolewski et al., 2015), which, while useful for file-naming and controlling directory structures, does not support querying across datasets. To address this, Queder et al. introduce NIDM-Terms, a formal set of user-friendly terminology management tools, and associated software to annotate BIDS datasets with a Neuroimaging Data Model (NIDM) semantic web representation.

Standardization of metadata is not only crucial for neuroimaging data, but also for anatomical studies that heavily rely on brain atlases. Kleven et al. provide a guide on the interpretation, navigation, spatial registration, data visualization, and transparent reporting of findings using different types of murine brain atlases. In addition, Blixhavn et al. provide a workflow defining the anatomical location of data elements in rodent brains as geometric objects based on atlas coordinates, which can be stored in a standardized file format. Using this method, disparate multimodal and multilevel neuroscience data can be co-visualized in three-dimensional digital brain atlases, enabling spatial data queries.

Even when data are shared, data accessibility, interoperability and reusability can be hindered by the use of proprietary data formats, especially when accompanying software becomes unavailable or unsupported. For proprietary electrophysiological data recorded with the DAPSYS software, Konradi et al. designed PyDapsys to enable direct opening of recorded files in Python and save them as NIX files, commonly used for open research in electrophysiology. This software promotes transparency and long-term accessibility in neuroscience research.

In this Research Topic, researchers describe various challenges and solutions surrounding FAIR data sharing and reuse in neuroscience. Their insights cover best practices for achieving data interoperability, the development of tools supporting scientists in data management and annotation, and the formulation of workflows to enhance the value of current and future data. We anticipate that the repositories, standards, and resources discussed in this Research Topic will not only simplify data sharing but also elevate reproducibility and foster widespread reuse of valuable neuroscience data. This collective effort holds the potential to significantly advance collaborative neuroscientific research.

Author contributions

MvS: Writing—original draft, Writing—review & editing. CH: Writing—review & editing.

Funding

The author(s) declare that no financial support was received for the research, authorship, and/or publication of this article.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

References

Gorgolewski, K., Tibor, A., Calhoun, V., Cameron Craddock, R., Samir, D., Duff, E., et al. (2015). The brain imaging data structure: a standard for organizing and describing outputs of neuroimaging experiments. *bioRxiv* [Preprint]. doi: 10.1101/034561

National Institutes of Health (2023). *NIH Data Management & Sharing Policy*. Available online at: https://sharing.nih.gov/data-management-and-sharing-policy/about-data-management-and-sharing-policies/data-management-and-sharing-policy-overview (accessed January 30, 2024).

Wilkinson, M., Dumontier, M., Aalbersberg, I., Appleton, G., Axton, M., Baak, A., et al. (2016). The FAIR Guiding Principles for scientific data management and stewardship. *Sci. Data* 3:160018. doi: 10.1038/sdata.2016.18

frontiers | Frontiers in Neuroinformatics

TYPE Perspective
PUBLISHED 09 March 2023
DOI 10.3389/fninf.2023.1154080

OPEN ACCESS

EDITED BY
Maaike M. H. Van Swieten,
Integral Cancer Center Netherlands (IKNL), Netherlands

REVIEWED BY
Yongsoo Kim,
Penn State Health Milton S. Hershey Medical Center, United States
Jonathan Robert Whitlock,
Kavli Institute for Systems Neuroscience, Norway

*CORRESPONDENCE
Ingvild E. Bjerke
✉ i.e.bjerke@medisin.uio.no

RECEIVED 30 January 2023
ACCEPTED 21 February 2023
PUBLISHED 09 March 2023

CITATION
Kleven H, Reiten I, Blixhavn CH, Schlegel U, Øvsthus M, Papp EA, Puchades MA, Bjaalie JG, Leergaard TB and Bjerke IE (2023) A neuroscientist's guide to using murine brain atlases for efficient analysis and transparent reporting.
Front. Neuroinform. 17:1154080.
doi: 10.3389/fninf.2023.1154080

A neuroscientist's guide to using murine brain atlases for efficient analysis and transparent reporting

Heidi Kleven, Ingrid Reiten, Camilla H. Blixhavn, Ulrike Schlegel, Martin Øvsthus, Eszter A. Papp, Maja A. Puchades, Jan G. Bjaalie, Trygve B. Leergaard and Ingvild E. Bjerke*

Neural Systems Laboratory, Institute of Basic Medical Sciences, University of Oslo, Oslo, Norway

Brain atlases are widely used in neuroscience as resources for conducting experimental studies, and for integrating, analyzing, and reporting data from animal models. A variety of atlases are available, and it may be challenging to find the optimal atlas for a given purpose and to perform efficient atlas-based data analyses. Comparing findings reported using different atlases is also not trivial, and represents a barrier to reproducible science. With this perspective article, we provide a guide to how mouse and rat brain atlases can be used for analyzing and reporting data in accordance with the FAIR principles that advocate for data to be findable, accessible, interoperable, and re-usable. We first introduce how atlases can be interpreted and used for navigating to brain locations, before discussing how they can be used for different analytic purposes, including spatial registration and data visualization. We provide guidance on how neuroscientists can compare data mapped to different atlases and ensure transparent reporting of findings. Finally, we summarize key considerations when choosing an atlas and give an outlook on the relevance of increased uptake of atlas-based tools and workflows for FAIR data sharing.

KEYWORDS

brain atlases, FAIR data, reporting practices, spatial registration, rat brain, mouse brain, brain-wide analysis, neuroinformatics

Introduction

Converting the increasing amounts of multifaceted neuroscience data into knowledge about the healthy and diseased brain requires that relevant data are accumulated and combined in a common context. The FAIR principles set forward by Wilkinson et al. (2016), stating that data should be findable, accessible, interoperable, and re-useable, facilitate such data integration. Practical implementation of these principles in neuroscience can be achieved by using brain atlases as a common framework, equipping the data with metadata describing their location in the brain. Brain atlases contain standardized references to brain locations, and their utility for integrating neuroscience data is already well-established (Toga and Thompson, 2001; Zaslavsky et al., 2014; Bjerke et al., 2018b).

Neuroscientists use atlases at several stages of a research project, from planning and conducting studies to analyzing data and publishing results. A variety of atlases exist, revealing different features of rat and mouse (collectively referred to as murine) neuroanatomy. However, different atlases use various traditions for defining and naming brain regions, hampering interpretation, and comparison of data from locations specified

using different atlases. Thus, while atlases provide common frameworks for neuroscience data integration, researchers might find it challenging to know which atlas to choose and how to use it. This makes it difficult for researchers to efficiently interpret and analyze their data using atlases, and for reporting and sharing data in accordance with the FAIR principles. Here, we provide a guide to using murine brain atlases for efficient analysis, reporting and comparison of data, offering the perspective that open volumetric brain atlases are essential for these purposes.

Finding brain locations by navigating and interpreting atlases

There are two types of murine brain atlases: traditional two-dimensional (2D) atlases with serial section images (e.g., Paxinos and Watson, 2013; Swanson, 2018) and digital volumetric (3D) atlases (e.g., Papp et al., 2014; Barrière et al., 2019; Wang et al., 2020). The traditional atlases rank among the most cited neuroscience publications. However, they are limited by the distance between section images and the fixed plane(s) of orientation. They are also poorly suited for automated whole-brain analysis and digital workflows, and reuse of atlas images in publications may require permission from the publisher. The digital volumetric atlases are typically shared openly, and they allow data analysis independent of the plane of sectioning. The most detailed and commonly used volumetric atlas for the mouse is the Allen Mouse Brain Common Coordinate Framework (Allen Mouse Brain CCF; Wang et al., 2020), which has been instrumental for the acquisition and sharing of the Allen Institute's large data collections (Lein et al., 2007; Oh et al., 2014; Tasic et al., 2016). For the rat, the most detailed volumetric atlas is the Waxholm Space atlas of the Sprague Dawley rat brain (WHS rat brain atlas; RRID:SCR_017124; Papp et al., 2014; Kjonigsen et al., 2015; Osen et al., 2019; Kleven et al., 2023a). Other murine brain atlases are also available [see summary by Barrière et al. (2019)]. Regardless of the 2D or 3D format, murine brain atlases can be navigated and interpreted using the spatial, visual, and semantic reference space (**Figure 1A**; Kleven et al., 2023a).

The *spatial reference* consists of a coordinate system and a reference image. The reference image of an atlas may originate from a single specimen (Papp et al., 2014) or represent a population average (Wang et al., 2020) of multiple specimens, with different brain region characteristics (e.g., cyto- or chemoarchitecture, and gene expression) visible depending on the modality. The reference image is made measurable through the coordinate system. Most brain atlases use a 3D Cartesian coordinate system with a defined origin and each of the x, y, z axes oriented in one of the standard anatomical planes. Atlases typically follow the neurological orientation of axes described by the right-anterior-superior (RAS) scheme, where the x-axis is oriented toward the right (R), the y toward anterior (A), and the z toward superior (S)[1]. The origin may be defined by skull features (stereotaxic coordinate system; Paxinos and Watson, 2013), internal landmarks (Waxholm Space; Papp et al., 2014), or the physical limits of the reference image such as the corner of a volume (Wang et al., 2020).

The *visual reference* consists of the reference image and a set of boundaries of brain regions (annotations), defined using criteria-based interpretations (e.g., differences in gene expression patterns, and changes in cyto-, myelo-, or chemoarchitecture). Easily recognizable features that are consistent across individuals are often used as landmarks when positioning an experimental image in an atlas (Sergejeva et al., 2015). For example, the beginning and end of easily distinguished brain regions, such as the caudoputamen or hippocampus (**Figure 1D**), are highly useful for orientation. Such landmarks are particularly useful for guiding and assessing the quality of the spatial registration of experimental section images to an atlas (Puchades et al., 2019; see section on analysis below), as well as for detecting abnormal anatomical features in the images. A selection of useful murine brain landmarks are given by Bjerke et al. (2023).

The *semantic reference* consists of the brain region annotations and their names. Regions, areas, and nuclei of the brain may be named after the person who first defined them, or after distinct features, such as their architecture or relative position within a broader region. While murine brain atlas terminologies often combine terms from different conventions, most atlases present white matter regions with a lower case first letter and gray matter regions with a capital first letter. Digital atlases may also use color coding schemes to indicate relationships between region annotations, e.g., using the same color for all white matter regions or for regions at the same level of the hierarchy of gray matter regions (Wang et al., 2020).

Analyzing data using atlas-based tools and workflows

Atlas coordinates provide spatial reference in machine-readable units. When coupled to the atlas terminology, they enable automated analysis of data registered to that atlas. A broad range of software incorporating atlases, here called atlas-based tools, are available to perform various digital analyses of brain image data. Atlas-based analyses rely on spatial registration, here defined as the process of assigning anatomical location to each pixel or voxel of the data (**Figure 2A**). This is achieved through aligning 2D and/or 3D data with the reference image of the atlas.

Several computational methods for registration of 2D image data to atlases have been developed. However, implementations are typically tailored to specific data types (e.g., fluorescent images or 3D data) and may require coding skills. Thus, tools with a graphical user interface that are applicable to a broad range of data types have also been developed, often incorporated as part of analytic workflows (Tappan et al., 2019; Ueda et al., 2020; BICCN Data Ecosystem Collaboration et al., 2022; Tyson and Margrie, 2022). An example of a standalone tool for spatial registration of histological sections to volumetric atlases is QuickNII (RRID:SCR_017978; Puchades et al., 2019). QuickNII is available with the WHS rat brain atlas (v2, v3, and v4) and the Allen Mouse Brain CCF (v3-2015 and v3-2017). Manual alignment of individual section images is relatively time-consuming, and can greatly benefit from a machine learning-based approach for section alignment, such as implemented in DeepSlice for

1 https://nipy.org/nibabel/neuro_radio_conventions.html

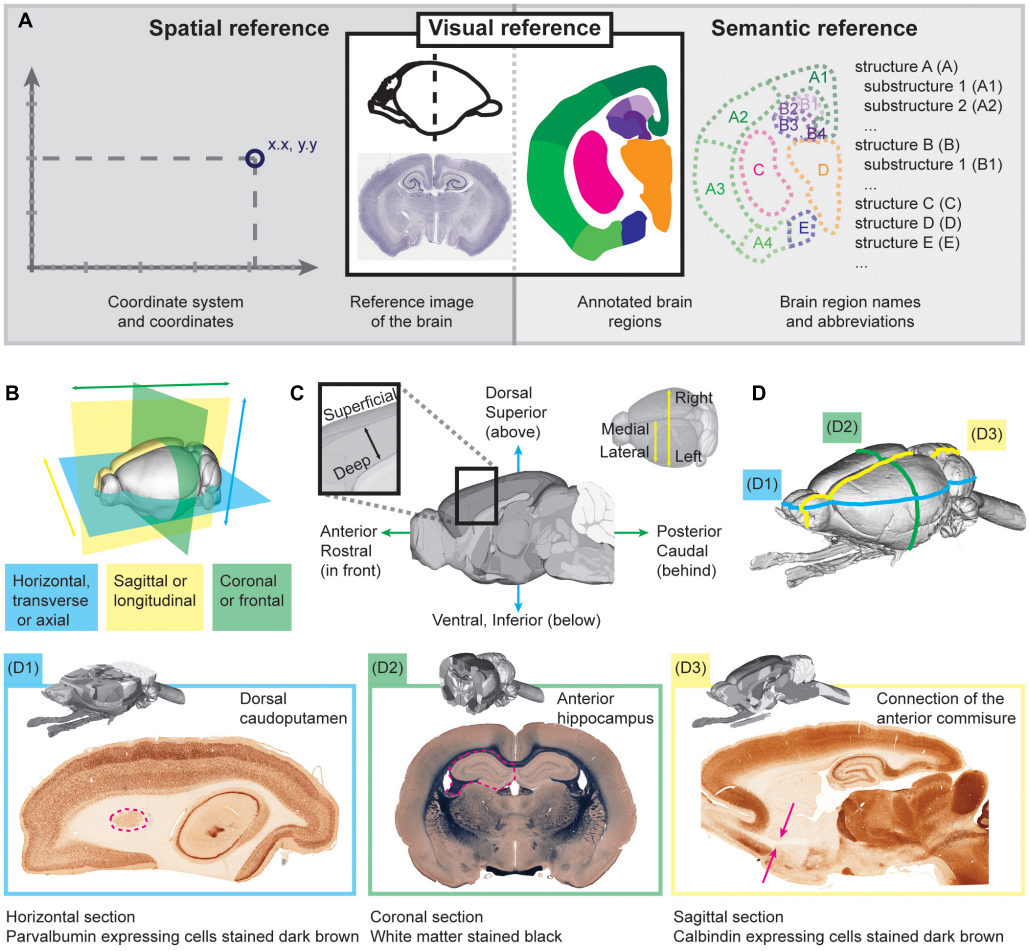


FIGURE 1
Navigating brain atlases to find anatomical locations. **(A)** Simplified version of the brain atlas ontology model (AtOM, Kleven et al., 2023a). The main elements of an atlas include the coordinate system, the reference image (here exemplified with a coronal platypus brain section; Mikula et al., 2007), the annotated brain regions, and the brain region names. The elements provide different entry points for navigating the atlas, through a spatial, semantic or visual reference. **(B)** Illustration of the three standard planes (horizontal, blue; sagittal, yellow; coronal, green) typically used to cut brain sections. **(C)** Illustration of the essential terminology typically used for indicating positions in the brain (e.g., the terms "rostral" and "caudal" to refer to positions towards the front and back of the brain, respectively). **(D)** Illustration of useful landmark regions in the murine brain [adapted from Bjerke et al. (2023)], with examples from the horizontal **(D1)**, coronal **(D2**; Leergaard et al., 2018), and sagittal **(D3)** planes.

coronal rat and mouse brain sections[2] (Carey et al., 2022). While these tools rely on linear registration methods, murine brains show variability (Badea et al., 2007; Scholz et al., 2016) that cannot always be compensated for by using linear transformations. Histological brain sections are also prone to physical damage and deformities caused by tissue processing (Simmons and Swanson, 2009). To amend this, non-linear adaptations of linearly registered murine images can be achieved using VisuAlign (RRID:SCR_017978).

Murine brain research increasingly includes 3D imaging data acquired by magnetic resonance or diffusion tensor imaging (Gesnik et al., 2017), serial two-photon imaging (Oh et al., 2014) or light sheet microscopy (Ueda et al., 2020). As these data are spatially coherent and avoid the deformities and damage seen in histological sections, they lend themselves well to volume-to-volume registration with 3D reference atlases. Several groups have developed computational methods for this type of alignment [see review by BICCN Data Ecosystem Collaboration et al. (2022)[3] and Tyson and Margrie (2022)], most often toward the Allen Mouse Brain CCF. The Elastix toolbox (Klein et al., 2010) also offers a collection of algorithms that can be used for 3D image registration.

Spatially registered image data can be used in analytic workflows for region-based annotation, quantification, and reconstruction of features in and across images. Such workflows typically entail three steps: (1) registration of image data (2D or 3D) to an atlas, (2) feature extraction, and (3) quantification and/or visualization of extracted features (**Figure 2B**). Several authors have demonstrated how such workflows can be used to quantify features of the brain (Kim et al., 2017; Pallast et al., 2019; Newmaster et al., 2020). Although many use custom code, workflows based on both commercial and open source tools exist.

2 www.deepslice.org

3 https://www.biorxiv.org/content/10.1101/2022.10.26.513573v1

For example, NeuroInfo from MBF Bioscience (Tappan et al., 2019) supports reconstruction of sections into a volume and registration to an atlas with automatic image segmentation and quantification. Alternatively, the free and open source QUINT workflow (Yates et al., 2019) aligns histological section images to atlas, and applies the same alignment to segmented images where a given feature (e.g., labeled cell bodies) is represented with a single color using the Nutil tool (Groeneboom et al., 2020, RRID:SCR_017183). For registration of electrode positions or viral expression, the HERBS software (Fuglstad et al., 2023) offers integrated spatial registration and feature extraction, where results can be directly visualized in 3D.

Visualization of atlases and image data

Spatial metadata makes it possible to view and interact with atlases and image data in several online atlas viewers. The Scalable Brain Atlas Composer[4] (SBA; Bakker et al., 2015) is capable of viewing 2D or 3D images of a range of different formats. In addition, the SBA can view spatial metadata (e.g., from QuickNII or DeepSlice) together with .png images of histological sections. Another online tool is the EBRAINS interactive atlas viewer[5], which is available for all versions of the WHS rat brain atlas and the Allen Mouse Brain CCF. This viewer also allows upload of user-defined data. For example, the user can drag-and-drop a .nii volume to view it in the three standard planes and slice it in arbitrary angles, with region annotations available as an overlay. Additionally, 3D rendering of coordinate-based data such as point clouds representing tracer distributions or cell bodies can be achieved online via MeshView (RRID:SCR_017222). MeshView allows slicing of volumes containing point clouds in user-defined planes for inspection and analysis of topographical patterns (see e.g., Tocco et al., 2022).

Customizing brain atlases for analysis and visualization

Open access digital brain atlases allow researchers to customize the anatomical annotations, reference images, or terminology in the atlas for specific analyses. Several tools have taken advantage of this, and enable the user to customize the atlas in an interactive way through a user interface. For example, QCAlign (RRID:SCR_023088; Gurdon et al., in preparation) allows interactive exploration of the hierarchy and grouping of brain region names that can subsequently be used in the QUINT workflow to merge brain regions into broader, custom regions for analysis. This may for example be used to merge and rename regions to make them compatible with a different naming convention, e.g., to enable cross-species comparison where atlases for different species must be harmonized (**Figure 2C**; Bjerke et al., 2021). Merging regions can also facilitate teaching by introducing students to macrostructure before revealing details. A more advanced use case is to modify or create new brain region annotations. For this purpose, the open access segmentation software like e.g., ITK-SNAP (Yushkevich et al., 2006) is useful for viewing and editing volumetric files across a range of different formats.

Comparing atlases and data mapped to different atlases

A major challenge across atlases is the variety of brain region annotations and terminologies (Swanson, 2000; Bohland et al., 2009). When different names are used to refer to the same brain region, or when similar names are used for partly overlapping ones, confusion is inevitable (Van De Werd and Uylings, 2014; Bjerke et al., 2020). Unequivocal referencing (see "Citing atlases and anatomical locations") can mitigate some of this, but the challenge remains that different terminologies often reflect differences in criteria for annotating brain regions. Differences in the brain region annotations across atlases and their versions make it difficult to compare data where locations are reported using different atlases. To amend this, Khan et al. (2018) performed a co-registration between versions of the stereotaxic rat brain atlases. They migrated data originally registered to one of Paxinos and Watson's (1986) earliest atlases to its corresponding plate in Swanson's (2018) most recent versions, making the data comparable. It is also possible to migrate legacy data to a volumetric atlas, upon which different datasets can be compared and co-visualized in 3D space (**Figure 2D**). To support such efforts, we have spatially registered several versions of the traditional stereotaxic atlases to the WHS rat brain atlas and Allen Mouse Brain CCF (Bjerke et al., 2020). The co-registration data are available for download through the EBRAINS Knowledge Graph[6] in QuickNII compatible format (see, e.g., Bjerke et al., 2018b and the related EBRAINS project on the web portal)[7]. The open access Swanson atlases are also available in an interactive viewer. Thus, the variety of atlases available and the fact that different data will be referenced using different atlases, while a challenge, can be mitigated by mapping atlases to each other.

Citing atlases and anatomical locations

Brain locations may be specified by names or coordinates, but to be reproducible a specific citation of the atlas used is required. A challenge is that researchers often report the name of a brain region that they are familiar with, and not the name recorded in the brain atlas they have used (Bjerke et al., 2020). For example, a researcher may use "striatum" to refer to the dorsal part of the striatal complex called "caudoputamen" in most atlases. While the researcher may see these names as interchangeable, a reader may

4 https://scalablebrainatlas.incf.org/composer/

5 https://interactive-viewer.apps.hbp.eu

6 https://search.kg.ebrains.eu/

7 https://search.kg.ebrains.eu/?category=Dataset&q=swanson#e2a1f65d-41fa-4bb1-ba48-93b36174a405

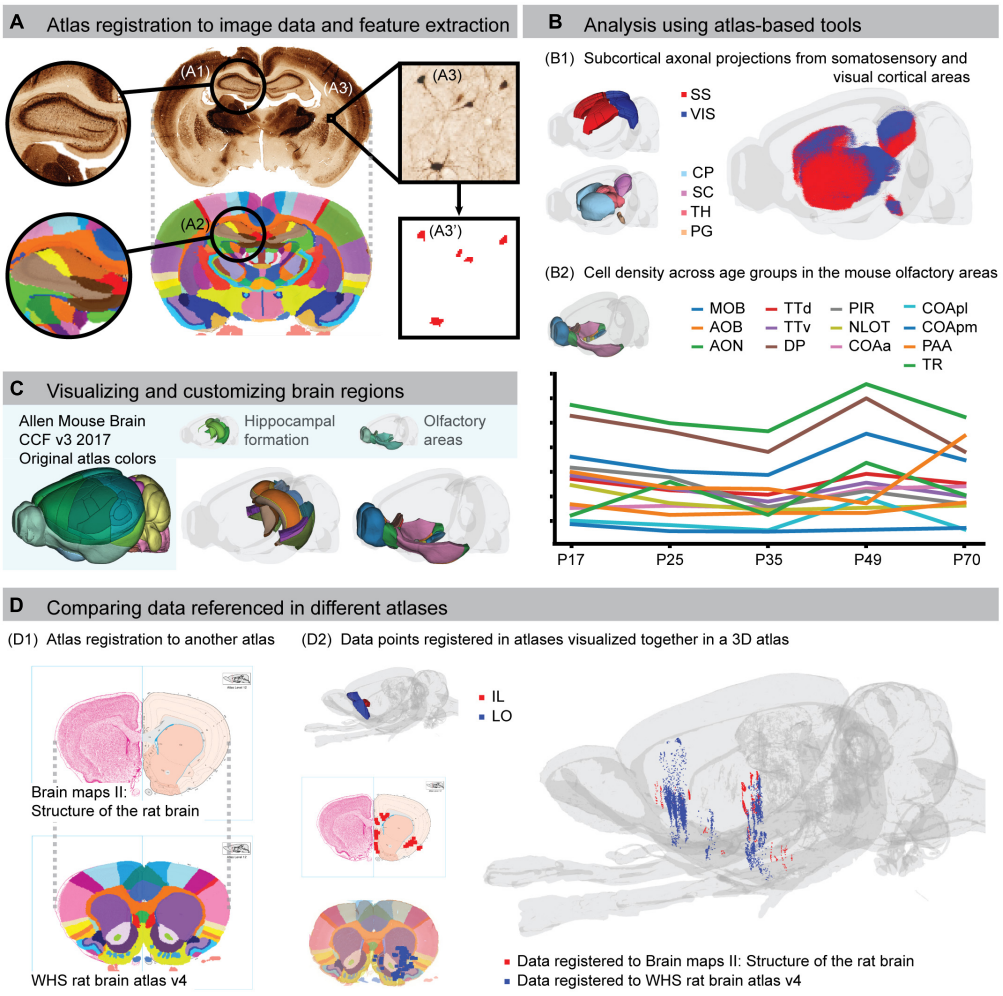


FIGURE 2
Using brain atlases for spatial registration, analysis, visualization and comparison of data. **(A)** Example of spatial registration of a histological section to the Waxholm Space (WHS) rat brain atlas. Landmark regions, such as the hippocampus **(A1,A2)**, are used to find corresponding positions between the image and the atlas. Features in the images **(A3)**, with spatial metadata from the registration, can be extracted **(A3')**. The example in **(A)** shows a histological image stained for parvalbumin neurons registered to the WHS atlas. **(B)** The principal workflow of combining atlas registration with extracted features illustrated in **(A)** can be used for different types of atlas-based analyses. **(B1)** 3D dot map visualization of corticostriatal, corticotectal, and corticopontine axonal projections originating from the primary somatosensory cortex (SS, red) and visual cortex (VIS, blue) cortical areas, extracted from anterograde tract tracing data (Oh et al., 2014) registered to the Allen mouse brain CCFv3-2017 (Ovsthus et al., 2022). **(B2)** Analysis of dopamine 1 receptor positive cell densities in olfactory regions of the mouse brain across five postnatal day (P) age groups [y axis values not shown, preliminary data extracted from images provided by Bjerke et al. (2022)]. **(C)** Visualization of customized regions from the Allen mouse brain CCFv3-2017. The left panel shows the entire atlas with the default color scheme. The middle panel shows a transparent view of the brain with regions of the hippocampal formation color coded to their corresponding region in the WHS rat brain atlas, facilitating cross-species comparisons. In the right panel, to better visualize the extent of individual regions, they are coded with contrasting colors, whereas the original atlas uses the same or highly similar colors. **(D)** Example of how co-registration of brain atlases supports comparison of data referenced in different atlases. The stereotaxic atlas by Swanson (1998) has been spatially registered to the WHS rat brain atlas **(D1)**. Data that have been extracted and mapped to the two atlases can therefore be co-visualized in the same 3D space **(D2)**. In this example, the red points are extracted from a previous study where retrograde projections from injections in the infralimbic cortex were represented with schematic drawing of terminal fields onto atlas plates from the Swanson atlas (Figure 8, data mirrored for comparison; Hoover and Vertes, 2007). The blue points are extracted from a public dataset showing the anterograde projections originating from the lateral orbitofrontal cortex (case F1 BDA; Kondo et al., 2022). AOB, accessory olfactory bulb; AON, anterior olfactory nucleus; CP, caudoputamen; COAa, cortical amygdalar area, anterior part; COApl, cortical amygdalar area, posterior part, lateral zone; COApm, cortical amygdalar area, posterior part, medial zone; DP, dorsal peduncular area; IL, infralimbic cortex; LO, lateral orbitofrontal cortex; MOB, main olfactory bulb; NLOT, nucleus of the lateral olfactory tract; PAA, piriform-amygdalar area; PG, pontine gray; PIR, piriform area; SC, superior colliculus; SS, somatosensory area; TH, thalamus; TR, postpiriform transition area; TTd, taenia tecta dorsal part; TTv, taenia tecta ventral part; VIS, visual area.

consider "striatum" to include the nucleus accumbens, which is also a common convention. This creates a source of confusion even when citing an atlas. We have previously put forward a set of recommendations to unambiguously refer to anatomical locations in the murine brain (Bjerke et al., 2018a), e.g., highlighting the importance of using terms as they appear in the atlas, or otherwise specifying how the terms used relate to those in the atlas.

Citation of an atlas should include the version. This is easy with traditional atlases following a linear versioning track

(Paxinos and Watson, 2007; Swanson, 2018). However, volumetric digital atlases are often provided with several files that may be versioned separately. To facilitate correct citation, Kleven et al. (2023a) proposed an Atlas ontology model and an overview of the versioning of the two most commonly used volumetric murine brain atlases, the WHS rat brain atlas and the Allen Mouse Brain CCF. Beyond consistent and correct citation of atlases, any customizations (see "Customizing brain atlases for analysis and visualization") should be clearly documented (Rodarie et al., 2021).

When using atlas-based software, it is important to be aware that software versioning is often independent of the atlas versioning. Thus, the software and atlas versions will have separate citation policies (usually along with separate RRIDs; Bandrowski and Martone, 2016), and should be named and cited accordingly when reporting data acquired using atlas-based software. Multiple atlases may be available in the same tool, in which case it is critical to record which atlas and version was used.

How to choose a brain atlas?

With several atlases available, it is challenging to know what sets different atlases apart and choosing the most appropriate brain atlas depends on its intended purpose. First, reproducibility and availability should be considered. In most laboratories, there are 2D book atlases on the shelf. While the mere physical availability of book atlases makes them convenient to use during experimental work, many of them are challenging to use for transparent reporting due to restrictive licenses and high costs for reproducing figures. Choosing an open access atlas makes it easier to communicate findings transparently and ensure their replicability. Second, reference images differ among atlases and should ideally match the experimental data at hand. For example, different strains are used across available rat brain atlases, with Wistar used in the Paxinos atlases (Paxinos and Watson, 2013) and Sprague Dawley used in Swanson's atlases (Swanson, 2018) and in the Waxholm Space rat brain atlas (Papp et al., 2014). Other characteristics, such as age category, sex, wild type or transgenic specimens, and data modality will also influence how well the atlas can be applied to experimental data. In general, the more characteristics match between the subjects used in an experiment and the reference atlas, the better the atlas will represent the data, an essential consideration for analyses. A third important feature of an atlas is its interoperability with other atlases and related analysis software. A researcher intending to analyze data based on an atlas will benefit from a digital 3D atlas incorporated in digital tools and workflows. Whether the atlas has been used in a similar study or is part of a data integration effort may also be relevant (Oh et al., 2014; Bjerke et al., 2018b; Erö et al., 2018), as this will facilitate comparison of findings with published data and enable similar comparisons in the future.

The evolution of brain atlases

Brain atlases are continuously created and refined to reflect researchers' needs for appropriate references for subjects of different ages (developmental or aging) or strains, or for data acquired with various imaging modalities, to mention some. In particular, there has been an increasing focus on the need for a common coordinate framework to map data across different developmental stages. A challenge with these resources is that they either do not cover early postnatal and embryonic stages (Newmaster et al., 2020), or have delineations that are not readily compatible with adult atlases (Young et al., 2021). Additionally, there is need for brain atlases capturing the fine details of brain regions distinguished by e.g., topographical organization of connections (Zingg et al., 2014; Hintiryan et al., 2016). For example, Chon et al. (2019) created an atlas with highly granular annotations of the mouse caudoputamen by using cortico- and thalamo-striatal connectivity data. By combining delineations from Allen Mouse Brain CCF and the Franklin and Paxinos atlases, this atlas also helps alleviating some of the inconsistencies in nomenclature (Chon et al., 2019). As these examples show, several atlases are required to cater to current needs, and future methodologies and findings will add further possibilities and needs for continued development and refinement of atlases. For such new atlases to enable researchers to cite, (re-)analyze, and compare data independently of the original atlas used, it is essential that they are openly shared and properly documented (Kleven et al., 2023a).

Conclusion and outlook: Open atlases help make data FAIR

In this perspective, we have provided a guide to murine brain atlases with a focus on how to use them for spatial registration, efficient analysis, and transparent reporting of data. Powerful analytic pipelines will hopefully incentivize more researchers to spatially register their data to atlases. We anticipate that the increasing availability and automation of atlas-based software with graphical user interfaces will fundamentally change how neuroscience will be performed in the future and lead to a major increase in the amount of more easily interpretable neuroscience data. For the field to benefit maximally from this shift, it is crucial that datasets and spatial metadata are openly shared in a public repository. This can be achieved with open access volumetric atlases as essential resources for making the wealth of multifaceted neuroscience data FAIR.

Data availability statement

The original contributions presented in this study are included in the article/supplementary material, further inquiries can be directed to the corresponding author.

Author contributions

HK and IB conceptualized and wrote the manuscript with input from TL. HK, IR, MØ, and IB prepared figures. All authors

contributed to the development of concepts and resources described in the manuscript, and to manuscript revision.

Funding

This work was funded by the European Union's Horizon 2020 Framework Program for Research and Innovation under the Specific Grant Agreement No. 945539 (Human Brain Project SGA3) and the Research Council of Norway under Grant Agreement No. 269774 (INCF Norwegian Node).

Acknowledgments

We thank Gergely Csucs and Dmitri Darine for their expert technical assistance.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

References

Badea, A., Ali-Sharief, A., and Johnson, G. (2007). Morphometric analysis of the C57BL/6J mouse brain. *Neuroimage* 37, 683–693. doi: 10.1016/j.neuroimage.2007.05.046

Bakker, R., Tiesinga, P., and Kötter, R. (2015). The scalable brain atlas: instant web-based access to public brain atlases and related content. *Neuroinformatics* 13, 353–366. doi: 10.1007/s12021-014-9258-x

Bandrowski, A., and Martone, M. (2016). RRIDs: a simple step toward improving reproducibility through rigor and transparency of experimental methods. *Neuron* 90, 434–436. doi: 10.1016/j.neuron.2016.04.030

Barrière, D., Magalhães, R., Novais, A., Marques, P., Selingue, E., Geffroy, F., et al. (2019). The SIGMA rat brain templates and atlases for multimodal MRI data analysis and visualization. *Nat. Commun.* 10, 1–13. doi: 10.1038/s41467-019-13575-7

BICCN Data Ecosystem Collaboration, Hawrylycz, M. J., Martone, M. E., Hof, P. R., Lein, E. S., Regev, A., et al. (2022). The BRAIN initiative cell census network data ecosystem: a user' s guide. *bioRxiv* doi: 10.1101/2022.10.26.513573 [Preprint].

Bjerke, I., Cullity, E., Kjelsberg, K., Charan, K., Leergaard, T., and Kim, J. (2022). DOPAMAP, high-resolution images of dopamine 1 and 2 receptor expression in developing and adult mouse brains. *Sci. Data* 9, 1–11. doi: 10.1038/s41597-022-01268-8

Bjerke, I., Øvsthus, M., Andersson, K., Blixhavn, C., Kleven, H., Yates, S., et al. (2018a). Navigating the murine brain: toward best practices for determining and documenting neuroanatomical locations in experimental studies. *Front. Neuroanat.* 12:82. doi: 10.3389/FNANA.2018.00082

Bjerke, I., Øvsthus, M., Papp, E., Yates, S., Silvestri, L., Fiorilli, J., et al. (2018b). Data integration through brain atlasing: human Brain Project tools and strategies. *Eur. Psychiatry* 50, 70–76. doi: 10.1016/j.eurpsy.2018.02.004

Bjerke, I., Ovsthus, M., Checinska, M., and Leergaard, T. (2023). An illustrated guide to landmarks in histological rat and mouse brain images. *Zenodo*. doi: 10.5281/zenodo.7575515

Bjerke, I., Puchades, M., Bjaalie, J., and Leergaard, T. (2020). Database of literature derived cellular measurements from the murine basal ganglia. *Sci. Data* 7:211. doi: 10.1038/s41597-020-0550-3

Bjerke, I., Yates, S., Laja, A., Witter, M., Puchades, M., Bjaalie, J., et al. (2021). Densities and numbers of calbindin and parvalbumin positive neurons across the rat and mouse brain. *iScience* 24:101906. doi: 10.1016/j.isci.2020.101906

Bohland, J., Bokil, H., Allen, C., and Mitra, P. (2009). The brain atlas concordance problem: quantitative comparison of anatomical parcellations. *PLoS One* 4:e0007200. doi: 10.1371/journal.pone.0007200

Carey, H., Pegios, M., Martin, L., Saleeba, C., Turner, A., Everett, N., et al. (2022). DeepSlice: rapid fully automatic registration of mouse brain imaging to a volumetric atlas. *bioRxiv* doi: 10.1101/2022.04.28.489953v1 [Preprint].

Chon, U., Vanselow, D., Cheng, K., and Kim, Y. (2019). Enhanced and unified anatomical labeling for a common mouse brain atlas. *Nat. Commun.* 10:5067. doi: 10.1038/s41467-019-13057-w

Erö, C., Gewaltig, M., Keller, D., and Markram, H. (2018). A cell atlas for the mouse brain. *Front. Neuroinform.* 12:84. doi: 10.3389/fninf.2018.00084

Fuglstad, J., Saldanha, P., Paglia, J., and Whitlock, J. (2023). HERBS: histological e-data registration in rodent brain spaces. *eLife* doi: 10.1101/2021.10.01.462770 [Epub ahead of print].

Gesnik, M., Blaize, K., Deffieux, T., Gennisson, J. L., Sahel, J. A., Fink, M., et al. (2017). 3D functional ultrasound imaging of the cerebral visual system in rodents. *Neuroimage* 149, 267–274. doi: 10.1016/j.neuroimage.2017.01.071

Groeneboom, N., Yates, S., Puchades, M., and Bjaalie, J. (2020). Nutil: a pre- and post-processing toolbox for histological rodent brain section images. *Front. Neuroinform.* 14:37. doi: 10.3389/fninf.2020.00037

Gurdon, B., Yates, S. C., Csucs, G., Groeneboom, N. E., Hadad, N., Telpoukhovskaia, M., et al. (in preparation). *Detecting the effect of genetic diversity on brain-wide cellular and pathological changes in a novel Alzheimer's disease mouse model.* Manuscript in preparation.

Hintiryan, H., Foster, N., Bowman, I., Bay, M., Song, M., Gou, L., et al. (2016). The mouse cortico-striatal projectome. *Nat. Neurosci.* 19, 1100–1114. doi: 10.1038/nn.4332

Hoover, W. B., and Vertes, R. P. (2007). Anatomical analysis of afferent projections to the medial prefrontal cortex in the rat. *Brain Struct. Funct.* 212, 149–179. doi: 10.1007/s00429-007-0150-4

Khan, A., Perez, J., Wells, C., and Fuentes, O. (2018). Computer vision evidence supporting craniometric alignment of rat brain atlases to streamline expert-guided, first-order migration of hypothalamic spatial datasets related to behavioral control. *Front. Syst. Neurosci.* 12:7. doi: 10.3389/fnsys.2018.00007

Kim, Y., Yang, G., Pradhan, K., Venkataraju, K., Bota, M., García del Molino, L., et al. (2017). Brain-wide maps reveal stereotyped cell-type-based cortical architecture and subcortical sexual dimorphism. *Cell* 171, 456–469. doi: 10.1016/j.cell.2017.09.020

Kjonigsen, L., Lillehaug, S., Bjaalie, J., Witter, M., and Leergaard, T. (2015). Waxholm Space atlas of the rat brain hippocampal region: three-dimensional delineations based on magnetic resonance and diffusion tensor imaging. *Neuroimage* 108, 441–449. doi: 10.1016/j.neuroimage.2014.12.080

Klein, S., Staring, M., Murphy, K., Viergever, M. A., and Pluim, J. P. W. (2010). Elastix: a toolbox for intensity-based medical image registration. *IEEE Trans. Med. Imaging* 29, 196–205. doi: 10.1109/TMI.2009.2035616

Kleven, H., Gillespie, T. H., Zehl, L., Dickscheid, T., and Bjaalie, J. G. (2023a). AtOM, an ontology model for standardizing use of brain atlases in tools, workflows, and data infrastructures. *bioRxiv* doi: 10.1101/2023.01.22.525049 [Preprint].

Kleven, H., Bjerke, I., Clascá, F., Groenewegen, H., Bjaalie, J. and Leergaard, T. (2023b). Waxholm Space atlas of the rat brain: A 3D atlas supporting data analysis and integration. [Preprint].

Kondo, H., Olsen, G., Gianatti, M., Monterotti, B., Sakshaug, T., and Witter, M. (2022). *Anterograde visualization of projections from orbitofrontal cortex in rat (v1.1).* EBRAINS doi: 10.25493/2MX9-3XF

Leergaard, T. B., Lillehaug, S., Dale, A., and Bjaalie, J. G. (2018). *Atlas of normal rat brain cyto- and myeloarchitecture [Data set].* Human Brain Project Neuroinformatics Platform. doi: 10.25493/C63A-FEY

Lein, E., Hawrylycz, M., Ao, N., Ayres, M., Bensinger, A., Bernard, A., et al. (2007). Genome-wide atlas of gene expression in the adult mouse brain. *Nature* 445, 168–176. doi: 10.1038/nature05453

Mikula, S., Trotts, I., Stone, J., and Jones, E. (2007). Internet-enabled high-resolution brain mapping and virtual microscopy. *Neuroimage* 35, 9–15. doi: 10.1016/j.neuroimage.2006.11.053

Newmaster, K., Nolan, Z., Chon, U., Vanselow, D., Weit, A., Tabbaa, M., et al. (2020). Quantitative cellular-resolution map of the oxytocin receptor in postnatally developing mouse brains. *Nat. Commun.* 11, 1–12. doi: 10.1038/s41467-020-15659-1

Oh, S., Harris, J., Ng, L., Winslow, B., Cain, N., Mihalas, S., et al. (2014). A mesoscale connectome of the mouse brain. *Nature* 508, 207–214. doi: 10.1038/nature13186

Osen, K., Imad, J., Wennberg, A., Papp, E., and Leergaard, T. (2019). Waxholm Space atlas of the rat brain auditory system: three-dimensional delineations based on structural and diffusion tensor magnetic resonance imaging. *Neuroimage* 199, 38–56. doi: 10.1016/j.neuroimage.2019.05.016

Ovsthus, M., Van Swieten, M. M. H., Bjaalie, J. G., and Leergaard, T. B. (2022). *Point coordinate data showing spatial distribution of corticostriatal, corticothalamic, corticocollicular, and corticopontine projections in wild type mice.* EBRAINS. doi: 10.25493/QT31-PJS

Pallast, N., Wieters, F., Fink, G., and Aswendt, M. (2019). Atlas-based imaging data analysis tool for quantitative mouse brain histology (AIDAhisto). *J. Neurosci. Methods* 326:108394. doi: 10.1016/j.jneumeth.2019.108394

Papp, E., Leergaard, T., Calabrese, E., Johnson, G., and Bjaalie, J. (2014). Waxholm Space atlas of the Sprague Dawley rat brain. *Neuroimage* 97, 374–386. doi: 10.1016/j.neuroimage.2014.04.001

Paxinos, G., and Watson, C. (1986). *The Rat Brain in Stereotaxic Coordinates*, 2nd Edn. Burlington, MA: Academic Press.

Paxinos, G., and Watson, C. (2007). *The Rat Brain in Stereotaxic Coordinates*, 6th Edn. Burlington, MA: Academic Press.

Paxinos, G., and Watson, C. (2013). *The Rat Brain in Stereotaxic Coordinates*, 7th Edn. Burlington, NJ: Elsevier Inc.

Puchades, M., Csucs, G., Ledergerber, D., Leergaard, T., and Bjaalie, J. (2019). Spatial registration of serial microscopic brain images to three-dimensional reference atlases with the QuickNII tool. *PLoS One* 14:e0216796. doi: 10.1371/journal.pone.0216796

Rodarie, D., Verasztó, C., Roussel, Y., Reimann, M., Keller, D., Ramaswamy, S., et al. (2021). A method to estimate the cellular composition of the mouse brain from heterogeneous datasets. *PLoS Comput. Biol.* 18:e01010739. doi: 10.1371/journal.pcbi.1010739

Scholz, J., LaLiberté, C., van Eede, M., Lerch, J., and Henkelman, M. (2016). Variability of brain anatomy for three common mouse strains. *Neuroimage* 142, 656–662. doi: 10.1016/j.neuroimage.2016.03.069

Sergejeva, M., Papp, E. A., Bakker, R., Gaudnek, M. A., Okamura-Oho, Y., Boline, J., et al. (2015). Anatomical landmarks for registration of experimental image data to volumetric rodent brain atlasing templates. *J. Neurosci. Methods* 240, 161–169. doi: 10.1016/j.jneumeth.2014.11.005

Simmons, D., and Swanson, L. (2009). Comparing histological data from different brains: sources of error and strategies for minimizing them. *Brain Res. Rev.* 60, 349–367. doi: 10.1016/j.brainresrev.2009.02.002

Swanson, L. (1998). *Brain Maps II: Structure of the Rat Brain*, 2nd Edn. Amsterdam: Elsevier.

Swanson, L. (2000). What is the brain? *Trends Neurosci.* 23, 519–527. doi: 10.1016/S0166-2236(00)01639-8

Swanson, L. (2018). Brain maps 4.0—Structure of the rat brain: an open access atlas with global nervous system nomenclature ontology and flatmaps. *J. Comp. Neurol.* 526, 935–943. doi: 10.1002/cne.24381

Tappan, S. J., Eastwood, B. S., O'Connor, N., Wang, Q., Ng, L., Feng, D., et al. (2019). Automatic navigation system for the mouse brain. *J. Comp. Neurol.* 527, 2200–2211. doi: 10.1002/cne.24635

Tasic, B., Menon, V., Nguyen, T., Kim, T., Jarsky, T., Yao, Z., et al. (2016). Adult mouse cortical cell taxonomy revealed by single cell transcriptomics. *Nat. Neurosci.* 19, 335–346. doi: 10.1038/nn.4216

Tocco, C., Øvsthus, M., Bjaalie, J., Leergaard, T., and Studer, M. (2022). Topography of corticopontine projections is controlled by postmitotic expression of the area-mapping gene Nr2f1. *Development* 149:26. doi: 10.1242/dev.200026

Toga, A., and Thompson, P. (2001). Maps of the brain. *Anat. Rec.* 265, 37–53. doi: 10.1002/ar.1057

Tyson, A. L., and Margrie, T. W. (2022). Mesoscale microscopy and image analysis tools for understanding the brain. *Prog. Biophys. Mol. Biol.* 168, 81–93. doi: 10.1016/j.pbiomolbio.2021.06.013

Ueda, H. R., Dodt, H. U., Osten, P., Economo, M. N., Chandrashekar, J., and Keller, P. J. (2020). Whole-brain profiling of cells and circuits in mammals by tissue clearing and light-sheet microscopy. *Neuron* 106, 369–387. doi: 10.1016/j.neuron.2020.03.004

Van De Werd, H., and Uylings, H. (2014). Comparison of (stereotactic) parcellations in mouse prefrontal cortex. *Brain Struct. Funct.* 219, 433–459. doi: 10.1007/s00429-013-0630-7

Wang, Q., Ding, S., Li, Y., Royall, J., Feng, D., Lesnar, P., et al. (2020). The allen mouse brain common coordinate framework: a 3D reference atlas. *Cell* 181, 1–18. doi: 10.1016/j.cell.2020.04.007

Wilkinson, M., Dumontier, M., Aalbersberg, I., Appleton, G., Axton, M., Baak, A., et al. (2016). The FAIR guiding principles for scientific data management and stewardship. *Sci. Data* 3, 1–9. doi: 10.1038/sdata.2016.18

Yates, S., Groeneboom, N., Coello, C., Lichtenthaler, S., Kuhn, P., Demuth, H., et al. (2019). QUINT: workflow for quantification and spatial analysis of features in histological images from rodent brain. *Front. Neuroinform.* 13:75. doi: 10.3389/fninf.2019.00075

Young, D., Darbandi, S. F., Schwartz, G., Bonzell, Z., Yuruk, D., Nojima, M., et al. (2021). Constructing and optimizing 3D atlases from 2D data with application to the developing mouse brain. *eLife* 10:e61408.

Yushkevich, P., Piven, J., Hazlett, H., Smith, R., Ho, S., Gee, J., et al. (2006). User-guided 3D active contour segmentation of anatomical structures: significantly improved efficiency and reliability. *Neuroimage* 31, 1116–1128. doi: 10.1016/j.neuroimage.2006.01.015

Zaslavsky, I., Baldock, R., and Boline, J. (2014). Cyberinfrastructure for the digital brain: spatial standards for integrating rodent brain atlases. *Front. Neuroinform.* 8:74. doi: 10.3389/fninf.2014.00074

Zingg, B., Hintiryan, H., Gou, L., Song, M., Bay, M., Bienkowski, M., et al. (2014). Neural networks of the mouse neocortex. *Cell* 156, 1096–1111. doi: 10.1016/j.cell.2014.02.023

frontiers | Frontiers in Neuroinformatics

TYPE Brief Research Report
PUBLISHED 26 April 2023
DOI 10.3389/fninf.2023.1173623

OPEN ACCESS

EDITED BY
Maaike M. H. Van Swieten,
Integral Cancer Center Netherlands (IKNL), Netherlands

REVIEWED BY
Maja Puchades,
University of Oslo, Norway
Brendan Behan,
Ontario Brain Institute, Canada

*CORRESPONDENCE
Arthur W. Toga
✉ toga@loni.usc.edu

RECEIVED 24 February 2023
ACCEPTED 12 April 2023
PUBLISHED 26 April 2023

CITATION
Neu SC, Crawford KL and Toga AW (2023) The image and data archive at the laboratory of neuro imaging.
Front. Neuroinform. 17:1173623.
doi: 10.3389/fninf.2023.1173623

The image and data archive at the laboratory of neuro imaging

Scott C. Neu, Karen L. Crawford and Arthur W. Toga*

Laboratory of Neuro Imaging, Department of Neurology, USC Mark and Mary Stevens Neuroimaging and Informatics Institute, University of Southern California, Los Angeles, CA, United States

The Image and Data Archive (IDA) is a secure online resource for archiving, exploring, and sharing neuroscience data run by the Laboratory of Neuro Imaging (LONI). The laboratory first started managing neuroimaging data for multi-centered research studies in the late 1990's and since has become a nexus for many multi-site collaborations. By providing management and informatics tools and resources for de-identifying, integrating, searching, visualizing, and sharing a diverse range of neuroscience data, study investigators maintain complete control over data stored in the IDA while benefiting from a robust and reliable infrastructure that protects and preserves research data to maximize data collection investment.

KEYWORDS

IDA, data repository, data sharing, data archive, neuroimaging

1. Introduction

The IDA (Toga et al., 2010; Toga and Crawford, 2015; Crawford et al., 2016) is a global resource for storing and disseminating neuroimaging, clinical, biospecimen, and genetic data for national and international consortia efforts (Redolfi et al., 2022) as well as smaller, single-center studies. Locally developed and managed at LONI, clinical data, imaging data, and analysis results are uploaded to the IDA daily, allowing users to obtain data from multiple studies within a single system. This manuscript summarizes recent improvements and developments within the IDA since our last report (Crawford et al., 2016).

Widespread data sharing is supported by IDA web pages that allow study-designated reviewers to receive, evaluate, and approve/disapprove online data use applications. Studies may define preset collections of data that meet specified criteria so that multiple users can access the same sets of data without first needing to conduct searches of the database. Since the IDA keeps extensive records of download activity, users can avoid downloading the same data twice and can easily locate new data after it arrives. There is no requirement to acknowledge the IDA in publications that use data obtained through the IDA, however, study-designated publication policies presented during data use application may specify acknowledgment requirements.

Image and Data Archive data ownership and access policies are defined so that the data belongs solely to its owners and that all data access decisions remain under their direct control. This functionality is often needed by study managers to control access to the data that is being pooled from multiple sites. Permissions to edit and delete data may also be assigned as needed to support review, tracking, and other data management operations. Quality assessments may be conducted by study owners, independent quality control contractors, or by LONI personnel on newly uploaded neuroimaging files that are hidden from users until quality ratings have been assigned.

2. Research studies

The IDA currently manages data on over 85,000 subjects from more than 140 research studies and 270 institutions and receives new data daily (Figure 1). These studies focus primarily on neurological diseases and conditions, and data has been collected in many different research areas including Alzheimer's disease, Epilepsy, Parkinson's disease, and traumatic brain injury (Table 1). While many studies use the IDA exclusively to archive and share data, there are a few studies that mirror data available in other repositories. Each study may provide a logo, set of colors, and a link to an external website that are used to alter the style of IDA web pages. This allows study coordinators to effectively brand the look and feel of the IDA to match each study's identity.

Originally conceived and developed as an image archive for magnetic resonance imaging (MRI) files, to date the IDA continues to collect neuroimaging scans from multiple modalities (Figure 2). Currently the IDA manages approximately 90% MRI [73% structural, 10% functional, and 17% diffusion tensor imaging (DTI)], 5% positron emission tomography (PET), and 5% other modalities such as computed tomography (CT), single-photon emission computed tomography (SPECT), and electroencephalogram (EEG). Additionally, the IDA stores files from other types of data, including clinical, electronic patient-reported outcomes (ePRO), proteomics, genetic (DNA/RNA), biospecimen analysis [cerebrospinal fluid (CSF), Fibroblast, peripheral blood mononuclear cells (PBMC), plasma, serum, urine, whole blood, cell line/induced pluripotent stem cells (iPSCs)], digital sensor (smart watch/smart phone), metabolomics, and proteomics.

3. Uploading and de-identifying data

The IDA website invites investigators to contact us via email to learn more about using the repository for their study. A welcome package with information about the repository and a form for gathering study details is provided to interested study contacts. Study investigators are asked to complete the form and submit study protocol and informed consent documents for USC IRB review. The completed form is used to assess whether the IDA is a good fit for the study's data, to determine the scope of work needed, and to estimate costs. A DTA/DUA template is available but local DTAs/DUAs can be accepted with USC Compliance Office approval.

Neuroimaging data files are de-identified and uploaded to the IDA using an executable Java jar file that implements the Java FX framework. This provides a graphical user interface (GUI) that guides users through the de-identification and upload steps. For easy installation, separate installers are available and have been code-signed for the Windows, Mac, and Linux operating systems, and each installer provides its own copy of the Java 15 runtime environment. Many neuroimaging data file formats are supported, including ANALYZE, DICOM, ECAT, EDF, FDF, FreeSurfer, GE, INTERFILE, MINC, NIFTI, NRRD, multi-image TIFF files (in regular and "big" TIFF format) and selected variants of MP4 video files. Target files are read and de-identified at each local institution before the de-identified files are sent to the IDA for archiving. Unlike previous IDA uploaders, the current uploader does not require a temporary working directory. The uploader is also self-updating; after installation the latest updates are automatically retrieved each time the uploader is started. After a user logs in, selects an IDA study, and specifies the user's site, the user enters a replacement subject ID and the directory path of files to upload. The file format of each file is automatically identified, and the appropriate de-identification removes patient-identifying information. De-identifications can be customized for the needs of each study, but in general all de-identifications replace the patient's name and ID fields with the user-supplied research identifier, remove all fields that are non-numeric unless otherwise specified, and replace unique identifier fields with hashed values. If required, obfuscation of binary content (e.g., images) is performed by image experts before uploading. A progress bar displays the total number of files that have been de-identified and uploaded to the IDA server along with reports of any files that have been rejected using de-identification criteria specific to each study. After all files have been uploaded, the user is directed to an IDA web page where additional study-specific information is entered, and the upload is finished. Copies of all files archived in the IDA are backed up in the cloud using the Amazon AWS S3 Glacier service.

More advanced users invoke a command line version of the Java uploader to perform batch uploads. This batch process requires a CSV file that must contain at least two columns; one column for the replacement subject ID and one column for the path of the files to upload. Each row of the CSV file identifies a separate upload. The batch uploader processes each row of the CSV file and writes a new CSV file as output. This new progress CSV file contains all the information provided in the first CSV file with additional columns describing the ID assigned to the upload, the LONI UID created for the uploaded files, the numbers of uploaded files and their file formats, and a column that provides the status of each upload. If a study requires additional information for an upload, new empty columns are also created. After users enter missing information and correct any status errors reported in the progress CSV file, they run the command line uploader again with the progress CSV file as input. Additional progress files are created until all uploads have completed. At the end of the batch upload process, the last progress CSV file provides users with a receipt containing detailed information about each upload.

For all neuroimaging data uploads, newly uploaded files are checked against previously uploaded files for duplicates using hashes and image header fields. When a duplicate is detected, the newly uploaded file replaces the previously uploaded file. In practice, many uploaders find duplicate detection and removal essential since in general institutions tend to focus more on image acquisition and less on local data management. As most uploads are files copied directly from imaging scanners, we have not encountered sufficient need to support multiple versions of the same upload. Optionally, non-duplicate neuroimaging data may be placed into one or more download queues for study personnel who wish to receive all newly uploaded data. They typically invoke an IDA download queue API on a nightly basis to get all data uploaded to a study for each day. Users can locate newly uploaded data by searching the IDA for all neuroimaging data they have not yet downloaded. Often after a study begins and de-identified neuroimaging data files have been archived in the IDA, study coordinators make requests for additional image header metadata

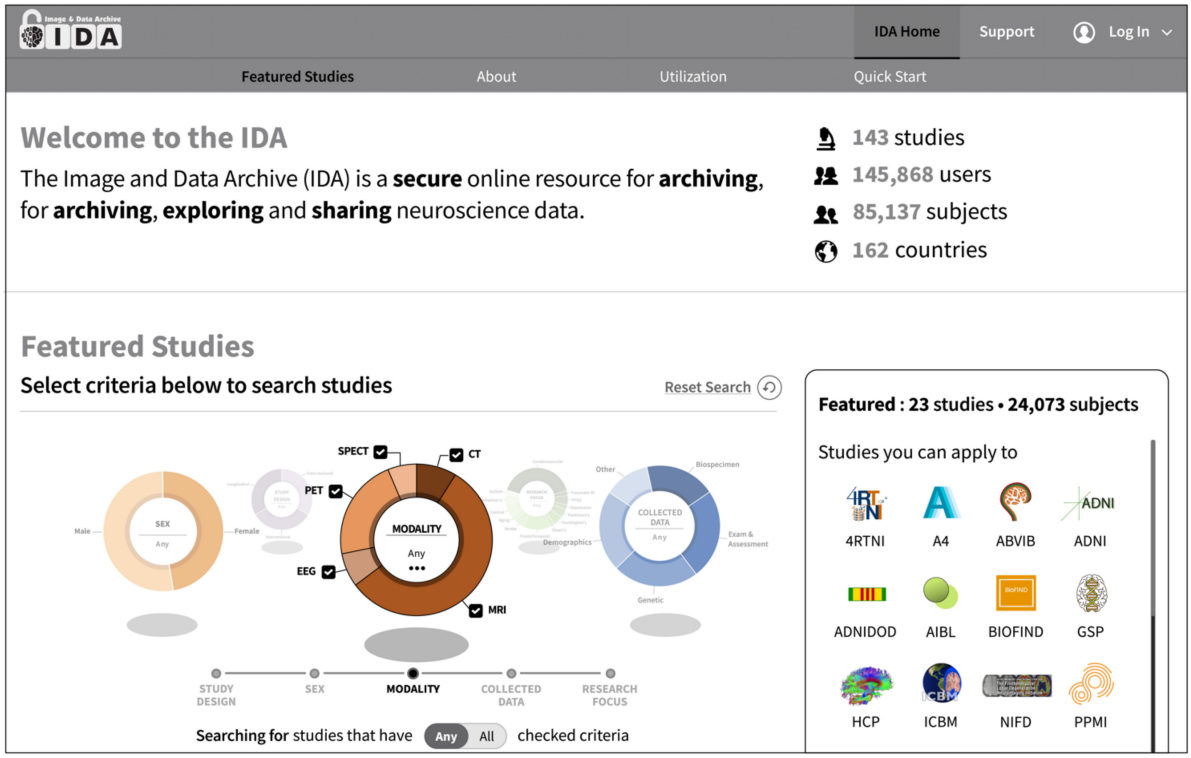


FIGURE 1
The IDA manages more than 85,000 subjects from over 140 research studies and 270 institutions and receives new data daily. Users search for and apply to featured studies on the IDA home page by searching study criteria such as study design and research focus.

changes. For example, new subject ID substitutions or higher levels of de-identification may be required. One common request is to retroactively "shift" all dates so that the time difference between any two dates in every neuroimaging data file is unchanged for the same subject. This involves selecting a date difference for each subject and replacing all neuroimaging image data files with the date-shifted metadata. This post-processing of archived neuroimaging files can be automatically applied to new image uploads or retroactively to all archived files in an IDA study.

4. Data management

Since the IDA functions as a hub for data transfers between collaborating groups (Figure 3), many studies require tabular data uploaded to the IDA to be processed and/or combined with other data before it becomes available for downloading. Data harmonization, quality control processing, and/or further de-identification is conducted in about one half of all studies featured on the IDA home page. To support these data mapping aims, we have developed an SQL-like language to create, edit, and execute transformations on database tables. A Java client provides a command line interface to add and remove "rules" that are executed by the client on data stored in the IDA database. Built on top of MySQL commands, IDA rule commands provide extra functionality to create and execute loops as well as to define variables. Statements in an IDA rule script are indented with white space similar to the Python programming language and are imported and exported from the IDA using the Java client. These rules can be executed manually, as part of cron jobs, or can be triggered after tabular data is uploaded to the IDA. Each line of a rule script may be associated with a variable that represents all output for that line, and output values are referenced by indented lines using the variable. There are five basic rule statements: loops (SQL SELECT statements), updates (SQL INSERT and UPDATE statements), if/else branches, identities (SET @X = 1), and IDA-specific functions (e.g., import REDCap instrument data). Additionally, error catching clauses can be added to execute logic if any rule statements fail. Unlike MySQL stored procedures, IDA rules are executed in two steps. First, all database changes output by the IDA rule are written to a separate working database. In the second step, these changes are copied (i.e., committed) from the working database to the target database. The primary advantage of this two-step paradigm is that IDA rules can be developed and tested without changing the target database data, which is preferable to making a copy of the database each time an IDA rule is tested. IDA rules also support temporary tables that provide temporary storage caches while rules are executing and can be used to import CSV content into a rule.

5. Data sharing and dissemination

Data sharing is primarily supported by study-designated reviewers who approve or disapprove data access requests submitted from IDA data use applications. Applicants

TABLE 1 Representative research studies utilizing the IDA.

Research focus	Study	Study	Institutions	Subjects	Deposit activity	Archived (GB)
Aging	FCDNA	Finance, cognition and default network in aging	1	76	2021–present	325
Aging	HABLE	Health and aging brain among Latino elders study	1	3147	2017–present	3,763
Aging	SLS	Seattle longitudinal study	1	5016	2019–present	113
Alzheimer's disease/dementia	ADSP_NACC	ADSP Phenotype harmonization consortium	1	3431	2022–present	148
Alzheimer's disease/dementia	ADPC	Alzheimer's Disease in primary care	1	483	2019–present	363
Alzheimer's disease/dementia	ADNI	Alzheimer's Disease neuroimaging initiative	71	5123	2005–present	7,135
Alzheimer's disease/dementia	A4	Anti-amyloid treatment in asymptomatic Alzheimer's	68	4486	2019–2021	790
Alzheimer's disease/dementia	VCSGT	Biomarkers of ABCA1 mediated functions in Alzheimer's	1	51	2017–present	259
Alzheimer's disease/dementia	DHA2BRP	DHA delivery to brain pilot study	2	460	2016–present	916
Alzheimer's disease/dementia	DVCID	Diverse VCID	9	239	2022–present	256
Alzheimer's disease/dementia	EEAJ	Estudio de la enfermedad de Alzheimer en Jalisciences	2	181	2016–present	638
Alzheimer's disease/dementia	GS1	Generation study 1	432	435	2022–present	501
Alzheimer's disease/dementia	GS2	Generation study 2	927	2446	2022–present	1,438
Alzheimer's disease/dementia	IDEASHOLD	Imaging dementia–evidence for amyloid scanning	343	10774	2021–present	97
Alzheimer's disease/dementia	LEADS	Longitudinal early-onset Alzheimer's disease study	18	576	2018–present	1,440
Alzheimer's disease/dementia	DVR	Model-based cerebrovascular markers for diagnosing MCI or AD	3	168	2019–present	115
Alzheimer's disease/dementia	SCAN_AL	SCAN legacy	4	420	2021–present	90
Alzheimer's disease/dementia	SCAN	Standardized centralized Alzheimer's neuroimaging	29	2225	2021–present	916
Alzheimer's disease/dementia	VCID	Vascular cognitive impairment and dementia	1	205	2017–present	307
Alzheimer's disease/dementia	VCD	Vascular cohort study	2	186	2015–present	699
Cerebrovascular disease	CHBC	Cardiovascular and HIV/AIDS effects on brain and cognition	4	520	2009–present	835
Cerebrovascular disease	PPG	Vascular contributions to dementia and genetic risk factors	3	452	2016–present	730
COVID-19	CVB	COVID-BRAIN	5	53	2021–present	138
Down syndrome	ABCDSU19	Alzheimer biomarker consortium–down syndrome	8	139	2021–present	196
Down syndrome	ADDS	Biomarkers of AD in adults with down syndrome	1	149	2019–present	436
Down syndrome	NIAD	Neurodegeneration in aging Down syndrome	6	250	2016–present	237
Epilepsy	EPIBIOS4	Epilepsy bioinformatics study for antiepileptogenic therapy	17	307	2016–present	2,221
Frontotemporal lobar degeneration	ALLFTD	ARTFL LEFFTDS longitudinal frontotemporal lobar degeneration	23	892	2020–present	1,258
Frontotemporal lobar degeneration	4RTNI	Four repeat Tauopathy neuroimaging initiative	5	129	2011–2016	147
Frontotemporal lobar degeneration	4RTNI2	Four repeat Tauopathy neuroimaging initiative cycle 2	8	257	2017–present	210

(Continued)

TABLE 1 (Continued)

Research focus	Study	Study	Institutions	Subjects	Deposit activity	Archived (GB)
Frontotemporal lobar degeneration	LEFFTDS	Longitudinal evaluation of familial frontotemporal dementia	18	909	2015–2020	588
Lifestyle intervention (Alzheimer's)	GEMS	Gene, exercise, memory and neurodegeneration in blacks study	2	143	2010–present	49
Lifestyle intervention (Alzheimer's)	LA_FINGERS	LatAm-FINGERS	10	381	2022–present	178
Lifestyle intervention (Alzheimer's)	POINTER	POINTER Imaging	6	809	2020–present	495
Parkinson's disease	BIOFIND	BioFIND	8	232	2012–present	1
Parkinson's disease	DODPD	DOD US army PD cognition longitudinal study	2	38	2018–2021	137
Parkinson's disease	PPMI	Parkinson's progression markers initiative	50	4314	2010–present	2,936
Stroke	SPAN	Stroke preclinical assessment network	7	2981	2020–present	2,117
Traumatic brain injury	ADNIDOD	Effects of TBI and PTSD on Alzheimer's disease in Vietnam vets	19	463	2013–2020	744
Traumatic brain injury	TRACKTBI	Transforming research and clinical knowledge in TBI	18	3569	2014–2022	3296

The total amount of neuroimaging files archived for each study is given in gigabytes (GB).

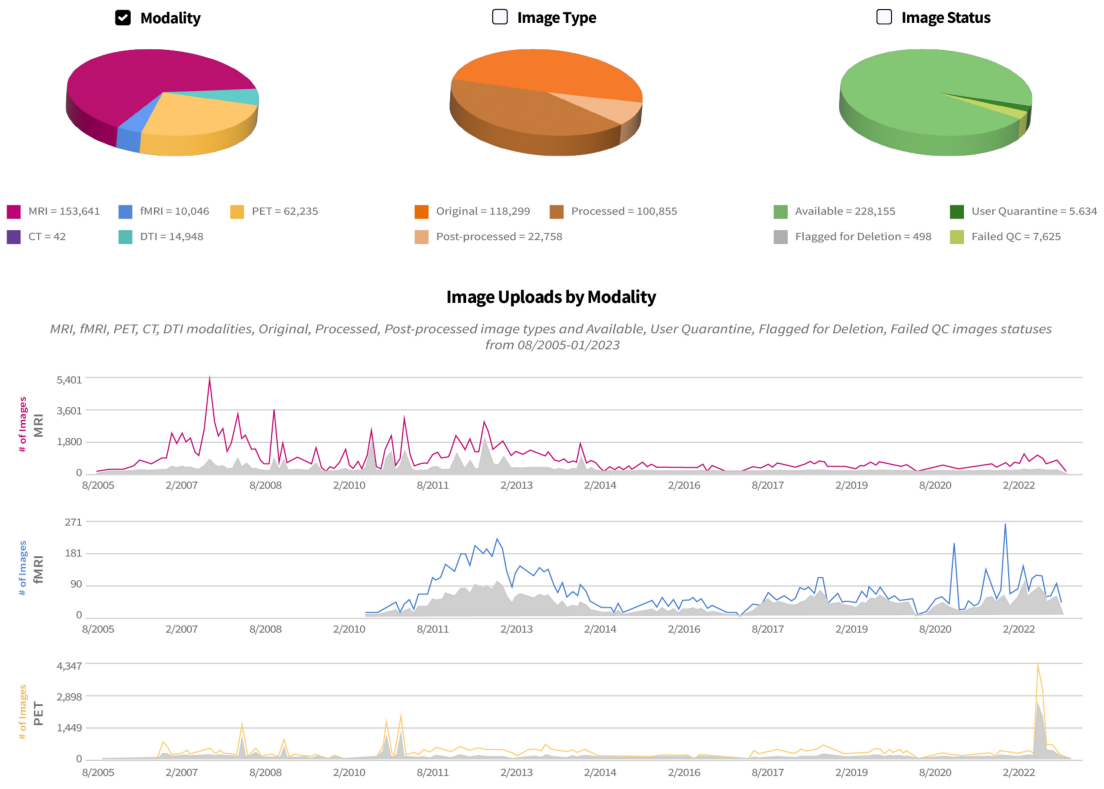


FIGURE 2
IDA study management pages provide a history of uploads and downloads of each neuroimaging modality. Counts of image uploads are categorized as original (images archived directly from scanner), processed (images corrected for artifacts, etc.) quarantined (removed from general view), and failed (not passing quality control).

must agree to the terms of data use agreement for each study and must provide information relevant to their proposed use of the study data. Applicants may also be required to submit any manuscripts they have written using study data to the IDA manuscript submission and review subsystem.

Access to data stored in the IDA for a study can be granted in three ways: (1) access to data from one or more institutions in a

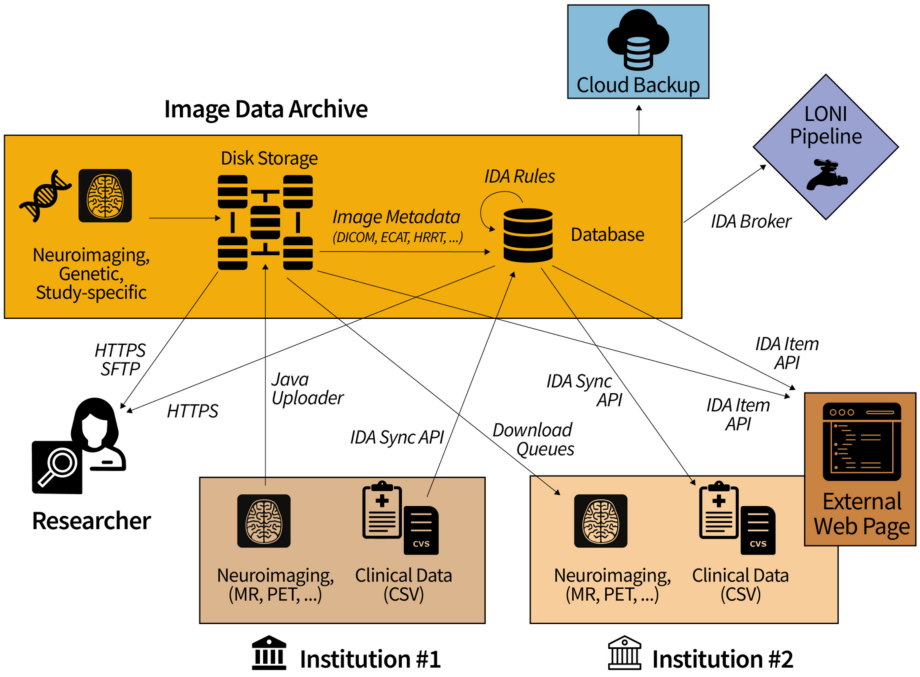


FIGURE 3
Collaborating institutions transfer data to and from the IDA while study committees allow access to approved researchers. In the diagram, Institution #1 uploads locally-acquired neuroimaging scans and clinical data. Download queues and the IDA Sync API make this data readily accessible to other institutions (e.g., Institution #2) for processing and reuploading.

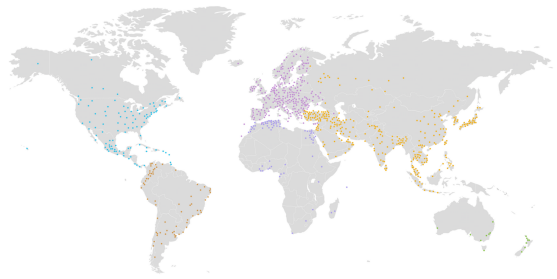


FIGURE 4
Over 145,000 users from 162 countries have downloaded more than 272 million neuroimaging data scans from the IDA.

TABLE 2 Top 10 countries downloading neuroimaging files from the IDA in gigabytes (GB).

Country	Downloaded (GB)
United States of America	866,105
China	238,059
Canada	117,341
Korea	110,357
Germany	87,606
United Kingdom of Great Britain and Northern Ireland	74,512
Japan	72,339
Hong Kong	69,609
India	69,257
Australia	54,252

study can be set in a user management web page, (2) a reviewer can grant guest-level (search and download) access to applicants through semi-automated data application web pages, and (3) all study data can be made publicly accessible to everyone having an IDA user account. Other access levels enable users to upload and download data files acquired at one or more institutions, and data management operations to edit and delete data can be granted.

Downloaders create customized collections of data files from the results of IDA searches and download each collection as a ZIP 64 file. As each archived neuroimaging scan is individually ZIP-compressed and stored with other scans from the same upload using a proprietary IDA "bundle" file format, for each download the requested ZIP-compressed scans are located in their respective bundle files and dynamically assembled into a ZIP 64 stream that is sent to the downloader. This paradigm eliminates the need to create the ZIP 64 file on a local IDA file system before sending and enables support for HTTP range and head requests. The HTTP head request provides the total size of the ZIP 64 file and range options specify a range of bytes to be downloaded from the file. This functionality supports the use of 3rd party download software to establish multiple connections to IDA servers to download different parts of a ZIP 64 file simultaneously, which can decrease the total amount of time needed to download the file. To prevent server overloads, the maximum number of connections allowed for a single user is capped at 10 per IDA server. We have also extended this download paradigm to individual files archived in the

IDA, including downloading tables from the IDA database in the comma-separated values (CSV) file format.

In addition to HTTPS requests, which provide a secure data transfer method used by all web browsers, the IDA also supports SSH File Transfer Protocol (SFTP) data transfers. This can be particularly useful for downloaders who are receiving large data files over poor connections. SFTP runs over SSH, has built-in integrity checks, and in our experience guards against file corruption much better than HTTPS, which (beyond TCP) does not incorporate check sum error checking. The IDA SFTP service provides a "virtual" file system using the open-source Apache MINA SSHD library into which users can log in and retrieve files. For security purposes, each download is assigned a random 36-character code that is used as the SFTP login name and expires after 12 h. Downloaders enter their IDA password as the SFTP password and then execute SFTP commands to download files from the virtual directory.

Tabular data is transferred to and from IDA servers with the IDA Sync API, which is a Representational State Transfer (REST) API that can be invoked by standard HTTPS utilities such as CURL and WGET. Authorization keys with limited lifetimes are obtained using IDA user credentials and are used in subsequent REST API endpoints. Responses can be returned in either XML or JSON, and tabular data is downloaded as CSV files. Flexible permissions for users are defined with regular expressions that identify accessible database tables by their names, and additionally filters can be applied to limit data per table row. IDA Sync API endpoints provide functionality for a user to (a) list all accessible tables in a database, (b) list column properties (e.g., data type) of all accessible tables, (c) download data from multiple tables in CSV format, (d) specify search criteria to filter downloaded data, and (e) upload data from a CSV file to an IDA database table. Data may be uploaded as a "partial sync" that updates existing database tables or as a "full sync" that deletes all data not referenced during the update. Additionally, users may define their own NULL characters and date/time formats.

Web applications integrated with the IDA Item API enable files archived in the IDA to be downloaded from web sites external to the IDA. This allows IDA collaborators to design their own web pages with links that access IDA information. The API provides a listing of the IDs, names, descriptions, and versions of all files in study-specific groups defined internally in the IDA. Download permissions for all IDA files accessed by the API endpoints are the same as if the files were directly downloaded from the IDA. External developers first obtain an authorization key for each user by invoking the API with the user's IDA email address and password. The API provides download links to the latest version of each file as well as older versions. In addition to providing access to files archived in the IDA, tabular data stored in the IDA database may be downloaded as CSV files.

Neuroimaging files archived in the IDA may be downloaded from the IDA using the IDA Java Broker, which can be integrated into external programs written in Java 1.8 or higher. The Broker requires each user's IDA email address and password and provides a list of all neuroimaging collections created by the user in the IDA. Every neuroimaging scan downloaded by the Broker is transferred as a ZIP 64 compressed stream and is automatically decompressed before being written to the target directory.

6. Discussion

To date, the IDA has enabled more than 145,000 users from 162 countries (**Figure 4** and **Table 2**) to download over 272 million neuroimaging data scans. The IDA currently manages 1.5 petabytes of storage, including 79 terabytes of 342 million neuroimaging files. Over 4,900 manuscripts (Weiner et al., 2013, 2015, 2017) have been accepted from IDA studies that require investigators to report their scientific findings. We believe these statistics demonstrate that the IDA functions as an effective repository for data sharing and promotes data reuse.

Data availability statement

The original contributions presented in this study are included in the article/supplementary material, further inquiries can be directed to the corresponding author.

Author contributions

SN, KC, and AT contributed to writing the text of this manuscript. All authors contributed to the article and approved the submitted version.

Funding

This work was supported by the Laboratory of Neuro Imaging Resource (LONIR)/P41/PPG grant 5P41EB015922-25, the Alzheimer's Disease Neuroimaging Initiative (ADNI) (3U01AG024904-10), and the Parkinson's Progression Markers Initiative (PPMI) of the Michael J. Fox Foundation for Parkinson's Research (MJFF-022056).

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

References

Crawford, K. L., Neu, S. C., and Toga, A. W. (2016). The image and data archive at the laboratory of neuro imaging. *Neuroimage* 124, 1080–1083.

Redolfi, A., Archetti, D., De Francesco, S., Crema, C., Tagliavini, F., Lodi, R., et al. (2022). Italian, European, and international neuroinformatics efforts: An overview. *Eur. J. Neurosci.* doi: 10.1111/ejn.15854 [Epub ahead of print].

Toga, A. W., and Crawford, K. L. (2015). The Alzheimer's disease neuroimaging initiative informatics core: A decade in review. *Alzheimers Dement.* 11, 832–839. doi: 10.1016/j.jalz.2015.04.004

Toga, A. W., Crawford, K. L., and Alzheimer's Disease Neuroimaging Initiative. (2010). The informatics core of the Alzheimer's Disease neuroimaging initiative. *Alzheimers Dement.* 6, 247–256.

Weiner, M., Veitch, D., Aisen, P., Beckett, L., Cairns, N., Cedarbaum, J., et al. (2015). Impact of the Alzheimer's disease neuroimaging initiative, 2004 to 2014. *Alzheimers Dement.* 11, 865–884. doi: 10.1016/j.jalz.2015.04.005

Weiner, M., Veitch, D., Aisen, P., Beckett, L., Cairns, N., Green, R., et al. (2013). The Alzheimer's Disease neuroimaging initiative: A review of papers published since its inception. *Alzheimers Dement.* 8(Suppl. 1), S1–S68.

Weiner, M., Veitch, D., Aisen, P., Beckett, L., Cairns, N., Green, R., et al. (2017). Recent publications from the Alzheimer's Disease neuroimaging initiative: Reviewing progress toward improved AD clinical trials. *Alzheimers Dement.* 13, e1–e85. doi: 10.1016/j.jalz.2016.11.007

frontiers | Frontiers in Neuroinformatics

TYPE Perspective
PUBLISHED 18 May 2023
DOI 10.3389/fninf.2023.1158378

OPEN ACCESS

EDITED BY
Maaike M. H. Van Swieten,
Integral Cancer Center Netherlands (IKNL), Netherlands

REVIEWED BY
Anita Sue Jwa,
Stanford University, United States

*CORRESPONDENCE
Brendan Behan
bbehan@braininstitute.ca

RECEIVED 03 February 2023
ACCEPTED 10 April 2023
PUBLISHED 18 May 2023

CITATION
Behan B, Jeanson F, Cheema H, Eng D, Khimji F, Vaccarino AL, Gee T, Evans SG, MacPhee FC, Dong F, Shahnazari S, Sparks A, Martens E, Lasalandra B, Arnott SR, Strother SC, Javadi M, Dharsee M, Evans KR, Nylen K and Mikkelsen T (2023) FAIR in action: Brain-CODE - A neuroscience data sharing platform to accelerate brain research. *Front. Neuroinform.* 17:1158378. doi: 10.3389/fninf.2023.1158378

FAIR in action: Brain-CODE - A neuroscience data sharing platform to accelerate brain research

Brendan Behan[1]*, Francis Jeanson[2], Heena Cheema[1], Derek Eng[1], Fatema Khimji[1], Anthony L. Vaccarino[3], Tom Gee[3], Susan G. Evans[3], F. Chris MacPhee[3], Fan Dong[3], Shahab Shahnazari[3], Alana Sparks[3], Emily Martens[3], Bianca Lasalandra[3], Stephen R. Arnott[4], Stephen C. Strother[4], Mojib Javadi[3], Moyez Dharsee[3], Kenneth R. Evans[3], Kirk Nylen[1,5] and Tom Mikkelsen[1]

[1]Ontario Brain Institute, Toronto, ON, Canada, [2]Datadex, Toronto, ON, Canada, [3]Indoc Research, Toronto, ON, Canada, [4]Rotman Research Institute, Toronto, ON, Canada, [5]Department of Pharmacology and Toxicology, University of Toronto, Toronto, ON, Canada

The effective sharing of health research data within the healthcare ecosystem can have tremendous impact on the advancement of disease understanding, prevention, treatment, and monitoring. By combining and reusing health research data, increasingly rich insights can be made about patients and populations that feed back into the health system resulting in more effective best practices and better patient outcomes. To achieve the promise of a learning health system, data needs to meet the FAIR principles of findability, accessibility, interoperability, and reusability. Since the inception of the Brain-CODE platform and services in 2012, the Ontario Brain Institute (OBI) has pioneered data sharing activities aligned with FAIR principles in neuroscience. Here, we describe how Brain-CODE has operationalized data sharing according to the FAIR principles. Findable—Brain-CODE offers an interactive and itemized approach for requesters to generate data cuts of interest that align with their research questions. Accessible—Brain-CODE offers multiple data access mechanisms. These mechanisms—that distinguish between metadata access, data access within a secure computing environment on Brain-CODE and data access via export will be discussed. Interoperable—Standardization happens at the data capture level and the data release stage to allow integration with similar data elements. Reusable - Brain-CODE implements several quality assurances measures and controls to maximize data value for reusability. We will highlight the successes and challenges of a FAIR-focused neuroinformatics platform that facilitates the widespread collection and sharing of neuroscience research data for learning health systems.

KEYWORDS

neuroinformatics, neuroscience, data sharing, data management, FAIR

Introduction

The sharing of data in the health biosciences domain is a vital component of advancing scientific research and accelerating discovery—with several funding agencies now mandating the sharing of datasets for such purposes (National Institutes of Health, 2023 NIH Data Management and Sharing Policy, Wellcome Trust Data (2017), Software and Materials Management and Sharing Policy, Government of Canada (2021) Tri-Agency Research Data Management Policy). The ability to harness knowledge from such datasets is dependent on there being sufficient information available that document their creation and curation. This is particularly important in a learning health system where research findings can be used to inform clinical care in the future. Four foundational principles of data sharing—Findability, Accessibility, Interoperability, Reusability (FAIR)—have emerged in the last decade as guiding elements on how datasets should be structured, annotated, and packaged to enable maximal reuse (Wilkinson et al., 2016). Within the domain of neuroscience, there has been movement toward greater efforts around data standardization in alignment with the FAIR principles (Poline et al., 2022).

The Ontario Brain Institute (OBI) is a provincially funded, not-for-profit organization founded in 2010 that accelerates discovery and innovation, benefiting both patients and the economy (Stuss, 2014). OBI funds research activities across several neuroscience domains through its Integrated Discovery Program (IDP) model. These pan-Ontario programs take an approach to research that spans several disciplines and brings together a diverse group of stakeholders including researchers, clinicians, industry partners, and patients and their advocates. Programs collect multiple data types including, but not limited to clinical, imaging, and molecular data. Within their studies, the programs have also incorporated standardized consent language to allow for re-use of de-identified datasets by external researchers and organizations (Lefaivre et al., 2019). This consent language was developed, in consultation with provincial research ethics board chairs and the Information and Privacy Commissioner of Ontario, in 2015 to support data sharing both within IDPs, as well as with external researchers and organizations. The consent language also speaks to linkage of data sets with independent databases, how participant information is kept confidential, and how participants can request withdrawal of data from respective studies.

To support the activities of these IDPs, a large-scale neuroinformatics platform—Brain-CODE—was developed to support the collection, storage, federation, sharing and analysis of different data types across several brain disorders. The technical and governance features of Brain-CODE have been previously described (Vaccarino et al., 2018; Lefaivre et al., 2019). This article will focus on the data sharing processes on Brain-CODE and their alignment with the FAIR principles.

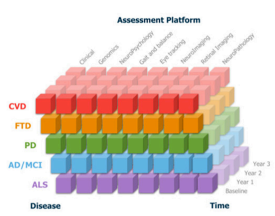
Alignment with FAIR principles—Findability

All available datasets for re-use are highlighted on the Brain-CODE portal.[1] An important element of Findability relates to describing datasets with rich and concise metadata. As such, Brain-CODE presents each data release with an initial description followed by standardized metadata including version #, data release date, # of participants, # of files, and overall dataset size. Other information presented about each data release include conditions of interest standardized to the Medical Dictionary for Regulatory Activities (MedDRA) ontology,[2] imaging scan type including task type described within The Cognitive Atlas knowledge base,[3] as well as data collection timepoints, modalities, and file formats. **Figure 1** highlights how a data release is typically presented to a data requestor.

As part of their activities, IDPs also highlight the availability of their datasets through their respective presentations and publications. Relatedly, OBI has partnered with the Canadian Open Neuroscience Platform (CONP) in the advertising of available datasets on the Brain-CODE portal. The CONP is a national network of Canadian neuroscience research centers committed to collaborating on a series of new open neuroscience initiatives (Harding et al., 2022).[4] All Brain-CODE data releases are registered on the searchable CONP data portal (Poline et al., 2023) and are described according to a customized version of the Data Tags Suite (DATS) metadata model (Alter et al., 2020). Another key feature of the Findability Principle is the assignment of a globally unique and persistent identifier to the respective dataset. Via involvement in CONP, Brain-CODE datasets are assigned an Archival Resource Key (ARK) ID[5]–a persistent identifier for information objects. In the near future, Brain-CODE plans to incorporate Digital Object Identifiers (DOIs) linked to their respective data releases.

Further advances are being planned to enhance Brain-CODE's Findability including expanding upon the current study-specific data release configuration to allow for cross-study query and cohort creation. This will enable data requestors to pool datasets across respective studies for integrative analyses. Additionally, OBI will continue to enhance the findability of datasets on the Brain-CODE portal through the incorporation of standardized metadata schemas (e.g., schema.org) to allow for querying by certain data search engines (e.g., Google Dataset Search). Finally, OBI is a member of the Global Alliance for Genomics and Health (GA4GH)–a policy-framing and technical standards-setting organization, seeking to enable responsible genomic data sharing (Terry, 2014)—and OBI continues to examine the incorporation of tools from various GA4GH driver projects [e.g., tagging of usage restrictions linked to datasets via the GA4GH Data Use Ontology (DUO)] (Lawson et al., 2021).

1 www.braincode.ca
2 https://www.meddra.org/
3 https://www.cognitiveatlas.org/
4 https://conp.ca/
5 https://arks.org/



Request Access

Ontario Neurodegenerative Disease Research Initiative (ONDRI): Foundational Study Baseline Data - Release 1

Dr. Mario Masellis, Dr. Douglas Munoz, Dr. Michael Strong, Dr. Richard Swartz, ONDRI study team

ONDRI's Foundational study includes data from 520 people living with Alzheimer's Disease, Mild Cognitive Impairment, Parkinson's Disease, Amyotrophic Lateral Sclerosis, Frontotemporal Dementia, or Cerebrovascular Disease, along with their care partners...

Show more

Version: 1.0 **Date Released:** 2022-06-08 **ARK ID:** 70798/d7r439px587bx4v6vq

Participants: 520 **Files:** 176 **Size:** 10.85 MB

Conditions: Alzheimer's disease, Cognitive impairment, Parkinson's disease, Amyotrophic lateral sclerosis, Frontotemporal dementia, Unspecified cerebrovascular disease

Timepoints: Baseline, Screening *(additional timepoints coming soon)*

Modalities: Cohort study, clinical, Cognitive, Gait and balance, Eye tracking, Retinal imaging, Genetic markers, Neuroimaging, Neuropsychological assessment, Cross-disease, Cross-platform, Caregiver, Deep pheotyping

Formats: CSV, TXT

For further information

Characteristics of the Ontario Neurodegenerative Disease Research Initiative cohort.
Sunderland KM, Beaton D, Arnott SR, Kleinstiver P, Kwan D, ,Lawrence-Dewar JM, Ramirez J, Tan B, Bartha R & Black SE (2022). Alzheimer's & Dementia.

Additional Resources

ONDRI's Research Initiative | ONDRI's Data Standards | Data Standards and Processes Videos

FIGURE 1
An example of how information about a data release is presented to a data requestor on the Brain-CODE platform—the Ontario neurodegenerative disease research initiative (ONDRI): Foundational study baseline data—Release 1 (Farhan et al., 2017; Sunderland et al., 2023).

Alignment with FAIR principles—Accessibility

There are different data access mechanisms based on the type of data that are being requested. Before data can be requested for access, there is a 12-month exclusivity period during which IDPs maintain exclusive access to their data. Data is then made accessible via either a Public or Controlled Access Mechanism on Brain-CODE. Datasets that have not previously contained Personal Health Information Protection Act [PHIPA] (2004) are made available by the Public Access Mechanism and can be accessed via the Brain-CODE portal without submitting a data access request proposal. Once a Brain-CODE account is created, data requestors utilize interactive dashboards to explore the data and metadata, select packages of interest, and download their respective data cuts. The most recent Public Access data release on Brain-CODE involves a priority setting partnership for epilepsy and seizures that was conducted by OBI, its epilepsy research program, EpLink, and the James Lind Alliance.[6]

Datasets that have previously contained PHI are made available by the Controlled Access Mechanism. A recent Controlled Access Mechanism data release is from the Canadian Biomarker Integration Network in Depression (CAN-BIND) and its foundational study (Lam et al., 2016; **Figure 2**). To prepare datasets to be released through this mechanism, IDPs provide data files, modality specific data dictionaries, and README files which are used to create an interactive Data Release Dashboard through which data requests can be submitted. Data files and supplementary documentation are manually reviewed by the IDP and OBI for direct identifiers as defined in OBI's Brain-CODE Governance policy,[7] such that the data are suitably de-identified (Theyers et al., 2021) prior to being available for external requests. The selection of these direct identifiers stems from legislation governing Brain-CODE activities in Ontario, Canada, notably the Personal Health Information Protection Act [PHIPA] (2004). While PHIPA does provide a definition of de-identification, there is limited guidance on its implementation. As such, OBI looked to methods in other jurisdictions, namely the U.S. Safe Harbor provision of the U.S. Health Information Portability and Accountability Act [HIPAA] (1996), and incorporated and customized such processes to both reduce risk of re-identification while ensuring usability of data sets made available via Brain-CODE.

Data requestors can review available data and metadata, select data packages of interest, and then submit a Data Access Request through the study Data Release Dashboard on the Brain-CODE portal. The data requestor will also be expected to provide further information about their planned analyses, provide documentation that an ethics committee has reviewed the project, and sign a Data Use Agreement. All requests are reviewed by the Brain-CODE Data Access Committee (DAC), composed of representatives from IDPs, experts in data privacy, and community representatives. The DAC provides a recommendation to approve or reject a data access request to the Brain-CODE Steering Committee, which is composed of OBI executive members. Once approved, access to data is granted either within a secure workspace on Brain-CODE or via local download. For the latter option, a data transfer agreement must be executed. These processes ensure that there are sufficient administrative and technical

6 https://www.jla.nihr.ac.uk/news/epilepsy-canada-psp-open-dataset-available/30802

7 https://braininstitute.ca/research-data-sharing/brain-code

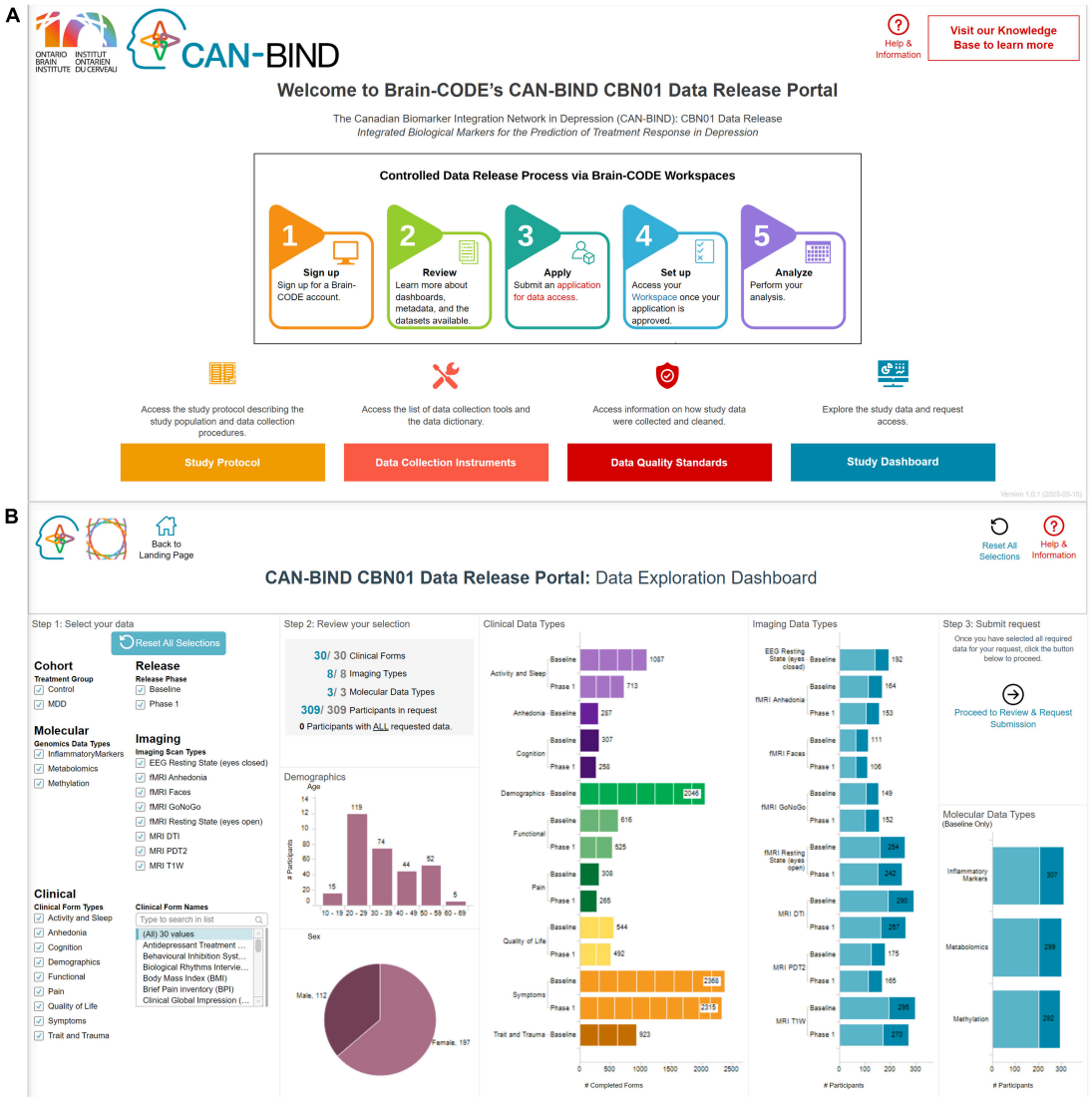


FIGURE 2
An example of how a data release is presented to the data requestor in terms of **(A)** providing supporting documentation [top figure—i.e., study protocol (Orange), data collection instruments (Pink), data quality standards (Red)] and **(B)** how interactive visualizations built using provided metadata can be used to select data packages of interest—CAN-BIND integrated biological markers for the prediction of treatment response in depression data release.

safeguards in making data sets available to data requestors in a secure manner.

Alignment with FAIR principles—Interoperability

Perhaps the greatest challenge in harnessing the full value of data is interoperability. While a dataset alone can have great utility in answering multiple questions, the prospect of combining a single dataset with others can increase its value by order of magnitude. Among the most critical aspect of data interoperability is the adoption of common data types, file formats, common data elements, semantic annotations, as well as common data packaging, metadata description, and indexing. Furthermore, platforms that aim to facilitate interoperability should provide the mechanisms and services that facilitate the combination and linking of data in useful ways.

Brain-CODE was designed to capture clinical, imaging, and molecular/genomics data with the help of three distinct but consistently used electronic data capture (EDC) systems, namely REDCap for demographics and clinical data, SPReD based on the XNAT platform for brain imaging data, and LabKey for molecular and genomics data (Vaccarino et al., 2018). The consistent use of EDC systems facilitates the curation and export of data to standard formats. These formats include comma separated value (CSV) files for tabular data, Neuroimaging Informatics Technology Initiative (NIfTI) files for binary imaging data,[8] the European Data Format

8 https://nifti.nimh.nih.gov/

(EDF +) files for times-series data,[9] and text files for applicable molecular and genomic data. Each of these are common non-proprietary file formats that can be used by numerous software for analysis or further processing.

With respect to common data elements (CDEs), OBI realized the opportunity to identify and adopt a common set of measures for demographic and clinical information collection across its IDPs. Established in 2013, CDEs were developed via a Delphi consensus process through the engagement of the clinical and research neuroscientific community among the IDPs. Identified CDEs span nine sub-domains of inquiry including demographics, socioeconomic status, quality of life, activities of daily living, medical comorbidity, psychiatric comorbidity, depression, anxiety, and sleep, and have been utilized for cross-disorder analysis (Vaccarino et al., 2022).

Datasets on Brain-CODE typically originate at the study level and are organized into distinct packages that can be based on modality type (e.g., clinical, imaging, molecular, etc.), modality subtype (e.g., MRI, EEG), participant cohort, and/or study timepoints. For imaging data, elements of the BIDS data structure standard have been adopted to facilitate interoperability of the data with various research software tools and other BIDS datasets as the adoption of this standard becomes increasingly common amongst the brain imaging research community (Gorgolewski et al., 2016). OBI has worked with the IDPs to collect and document metadata, including study protocol details, data collection processes, preprocessing and data provenance information, and details regarding study contributors, publications, etc. In addition, video presentations regarding the respective study have been published to accompany the most recent releases.[10]

To further support the combining of data, Brain-CODE facilitates the linking of data using advanced privacy preserving record linking (PPRL) via a deterministic El Gamal homomorphic encryption and matching based protocol (Gee et al., 2018). By collecting sensitive direct identifiers (such as provincial health card numbers in Ontario, Canada) in an encrypted format, Brain-CODE can match participant records with other data providers to achieve linkages where new information can be added to data available for the same participants while preserving privacy. With this mechanism, OBI data partners such as the Institute for Clinical Evaluative Sciences (ICES) can link health administrative data with research data on Brain-CODE (Behan et al., 2020; Southwell et al., 2022). As a result, rich participant profiles can be generated to answer deeper research questions.

Alignment with FAIR principles—Reusability

Reusability can only be achieved if the prior three principles are well implemented. Without findability, accessibility, and interoperability, the reuse of data will be challenging. In addition, reusability requires that critical pieces of information accompany datasets and that expert support is provided to data requestors in a consistent and reliable manner with respect to the use of data. While OBI seeks to define upfront standards for datasets, novel data requests may necessitate data to be reformatted, described, or computed in new ways. As a result, OBI seeks to follow key steps for data reuse including simplicity, portability, annotation, and quality reporting.

Simplicity in data formats and packaging help ensure that data requestors will have the ability to interpret and ingest the data using well established processing and analysis tools. As discussed above, common data formats that are non-proprietary, capture essential information, and have the flexibility to be extended are preferred on Brain-CODE. Combining data files with metadata files in a simple data package hierarchy that are consistently implemented across Brain-CODE also plays a role in achieving this.

Portability of data and metadata is essential for nimble reuse of data under various circumstances. For example, data administrators may need to load the data in a secure computing enclave that meets specific analysis protocols. Data may also need to be transformed to match target analytics model needs. If the data were scattered across multiple databases and file systems, it would require significant resources to manage, as well as being more susceptible to errors. By adopting a standard packaging approach based on a hierarchical file system, datasets are portable and can be processed as required to meet the target use case.

Rich annotation is important to generate during data collection and curation to better support data reuse. Fields including their labels, values, and units should be semantically coded according to standard control vocabularies and ontologies to maximize the opportunities for remodeling of data and their structure. For example, a data requestor may require data to match an observation medical outcomes partnership (OMOP) common data model (CDM) to combine with other OMOP CDM data. Without suitable annotations, a mapping from a source structure and labels to OMOP, or to another required data model, cannot be achieved resulting in a failure of data reuse. While rich annotations are important, the process can lead to over-engineered data that negatively impacts simplicity. As a result, Brain-CODE selectively adopts rich semantic vocabularies and data structures on an as-needed basis or as clear standards emerge within a particular domain of research. This is an active area of development for data on Brain-CODE.

Data quality reports provide data requestors with key information on the characteristics of data which helps with the identification of data origin, data completeness, and data integrity. This is a critical element to help create trust in the data and in the veracity of results from their reuse. Brain-CODE generates visual dashboards with rich data characteristics that data requestors benefit from when interpreting the data as we continue to work on improving this quality reporting.

Finally, ongoing support by the Brain-CODE informatics team helps ensure that data requestors understand the steps required for access, in what manner they can access the data (such as download to their local machine or via the use of a computing workspace), and to answer any further questions related to the origin and characteristics of the data. Without a human-in-the-loop, requests can stall and opportunities for discovery may not be realized.

9 https://www.edfplus.info/

10 https://www.braincode.ca/content/controlled-data-releases

Discussion

Altogether, Brain-CODE is a functioning example within the neuroscience field as to how re-use of datasets can be supported in alignment with the FAIR principles. To date, Brain-CODE has handled hundreds of data access requests from both academic and non-academic groups globally. This has allowed for greater opportunities for data exploration as well as affording data requestors the opportunity to address research questions of interest without having to initiate large-scale data collection efforts. The Brain-CODE platform continues to be developed to enhance data sharing efforts to allow for greater data discovery and understanding of various brain disorders.

Data availability statement

The original contributions presented in this study are included in the article/supplementary material, further inquiries can be directed to the corresponding author.

Author contributions

BB, FJ, and HC wrote the first draft of the manuscript and prepared the manuscript. All authors contributed to the development of the data sharing processes via Brain-CODE and commented on/revised the manuscript at all stages.

Funding

This OBI funding was provided in part by the Government of Ontario.

Conflict of interest

AV, TG, SE, FM, FD, ShS, AS, EM, BL, MJ, MD, and KE were employed by Indoc Research.

The remaining authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

References

Alter, G., Gonzalez-Beltran, A., Ohno-Machado, L., and Rocca-Serra, P. (2020). The data tags suite (DATS) model for discovering data access and use requirements. *GigaScience* 9:giz165. doi: 10.1093/gigascience/giz165

Behan, B., Gee, T., Evans, S. G., Dharsee, M., Evans, K., Azimaee, M., et al. (2020). Using A privacy preserving record linkage to facilitate an ongoing crosswalk between research and health administrative databases. *Int. J. Popul. Data Sci.* 5. doi: 10.23889/ijpds.v5i5.1630

Farhan, S. M., Bartha, R., Black, S. E., Corbett, D., Finger, E., Freedman, M., et al. (2017). The Ontario neurodegenerative disease research initiative (ONDRI). *Can. J. Neurol. Sci.* 44, 196–202. doi: 10.1017/cjn.2016.415

Gee, T., Behan, B., Lefaivre, S., Azimaee, M., Dharsee, M., El Emam, K., et al. (2018). Designing and Implementing a privacy preserving record linkage protocol. *Int. J. Popul. Data Sci.* 3. doi: 10.23889/ijpds.v3i4.831

Gorgolewski, K. J., Auer, T., Calhoun, V. D., Craddock, R. C., Das, S., Duff, E. P., et al. (2016). The brain imaging data structure, a format for organizing and describing outputs of neuroimaging experiments. *Sci. Data* 3:160044. doi: 10.1038/sdata.2016.44

Government of Canada, (2021). *Tri-Agency Research Data Management Policy*. Available online at: https://science.gc.ca/site/science/en/interagency-research-funding/policies-and-guidelines/research-data-management/tri-agency-research-data-management-policy (accessed February 1, 2023).

Harding, R. J., Bermudez, P., Beauvais, M., Bellec, P., Hill, S., Knoppers, B. M., et al. (2022). The canadian open neuroscience platform – an open science framework for the neuroscience community. *[Preprint]*. doi: 10.31219/osf.io/eh349

Health Information Portability and Accountability Act [HIPAA], (1996). Guidance Regarding Methods for De-Identification of Protected Health Information in Accordance with the Health Insurance Portability and Accountability Act (HIPAA) Privacy Rule. Available online at: https://www.hhs.gov/hipaa/for-professionals/privacy/special-topics/de-identification/index.html (accessed February 1, 2023).

Lam, R. W., Milev, R., Rotzinger, S., Andreazza, A. C., Blier, P., Brenner, C., et al. (2016). Discovering biomarkers for antidepressant response: protocol from the Canadian biomarker integration network in depression (CAN-BIND) and clinical characteristics of the first patient cohort. *BMC Psychiatry* 16:105. doi: 10.1186/s12888-016-0785-x

Lawson, J., Cabili, M. N., Kerry, G., Boughtwood, T., Thorogood, A., Alper, P., et al. (2021). The data use ontology to streamline responsible access to human biomedical datasets. *Cell Genomics* 1:100028. doi: 10.1016/j.xgen.2021.100028

Lefaivre, S., Behan, B., Vaccarino, A., Evans, K., Dharsee, M., Gee, T., et al. (2019). Big data needs big governance: best practices from brain-CODE, the Ontario-brain institute's neuroinformatics platform. *Front. Genet.* 10:191. doi: 10.3389/fgene.2019.00191

National Institutes of Health (2023). *NIH Data Management & Sharing Policy*. Available online at: https://sharing.nih.gov/data-management-and-sharing-policy/about-data-management-and-sharing-policies/data-management-and-sharing-policy-overview (accessed February 1, 2023).

Personal Health Information Protection Act [PHIPA] (2004). *S.O. 2004, c. 3, Sched. A*. Available online at: https://www.ontario.ca/laws/statute/04p03

Poline, J. B., Das, S., Glatard, T., Madjar, C., Dickie, E., Lecours, X., et al. (2023). Data and tools integration in the canadian open neuroscience platform. *Sci. Data* 10:189. doi: 10.1038/s41597-023-01946-1

Poline, J. B., Kennedy, D. N., Sommer, F. T., Ascoli, G. A., Van Essen, D. C., Ferguson, A. R., et al. (2022). Is neuroscience FAIR? A call for collaborative standardisation of neuroscience data. *Neuroinformatics* 20, 507–512. doi: 10.1007/s12021-021-09557-0

Southwell, A., Bronskill, S., Gee, T., Behan, B., Evans, S., Mikkelsen, T., et al. (2022). Validating a novel deterministic privacy-preserving record linkage between administrative & clinical data: applications in stroke research. *Int. J. Popul. Data Sci.* 7. doi: 10.23889/ijpds.v7i4.1755

Stuss, D. T. (2014). The Ontario Brain Institute: completing the circle. *Can. J. Neurol. Sci.* 41, 683–693. doi: 10.1017/cjn.2014.36

Sunderland, K. M., Beaton, D., Arnott, S. R., Kleinstiver, P., Kwan, D., Lawrence-Dewar, J. M., et al. (2023). Characteristics of the Ontario neurodegenerative disease research initiative cohort. *Alzheimer's Dement.* 19, 226–243. doi: 10.1002/alz.12632

Terry, S. F. (2014). The global alliance for genomics & health. *Genet. Test. Mol. Biomark.* 18, 375–376. doi: 10.1089/gtmb.2014.1555

Theyers, A. E., Zamyadi, M., O'Reilly, M., Bartha, R., Symons, S., MacQueen, G. M., et al. (2021). Multisite comparison of MRI defacing software across multiple cohorts. *Front. Psychiatry* 12:617997. doi: 10.3389/fpsyt.2021.617997

Vaccarino, A. L., Beaton, D., Black, S. E., Blier, P., Farzan, F., Finger, E., et al. (2022). Common data elements to facilitate sharing and re-use of participant-level data: assessment of psychiatric comorbidity across brain disorders. *Front. Psychiatry* 13:816465. doi: 10.3389/fpsyt.2022.816465

Vaccarino, A. L., Dharsee, M., Strother, S., Aldridge, D., Arnott, S. R., Behan, B., et al. (2018). Brain-CODE: a secure neuroinformatics platform for management, federation, sharing and analysis of multi-dimensional neuroscience data. *Front. Neuroinformat.* 12:28. doi: 10.3389/fninf.2018.00028

Wellcome Trust Data (2017). *Software and Materials Management and Sharing Policy*. Available online at: https://wellcome.org/grant-funding/guidance/data-software-materials-management-and-sharing-policy (accessed February 1, 2023).

Wilkinson, M. D., Dumontier, M., Aalbersberg, I. J., Appleton, G., Axton, M., Baak, A., et al. (2016). The fair guiding principles for scientific data management and stewardship. *Sci. Data* 3:160018. doi: 10.1038/sdata.2016.18

frontiers | Frontiers in Neuroinformatics

TYPE Perspective
PUBLISHED 25 May 2023
DOI 10.3389/fninf.2023.1175689

OPEN ACCESS

EDITED BY
Maaike M. H. Van Swieten,
Integral Cancer Center Netherlands (IKNL), Netherlands

REVIEWED BY
Neil P. Oxtoby,
University College London, United Kingdom

*CORRESPONDENCE
Arthur W. Toga
toga@loni.usc.edu
Mukta Phatak
mukta.phatak@alzheimersdata.org
John Gallacher
john.gallacher@psych.ox.ac.uk

†These authors have contributed equally to this work

‡Senior author

RECEIVED 27 February 2023
ACCEPTED 02 May 2023
PUBLISHED 25 May 2023

CITATION
Toga AW, Phatak M, Pappas I, Thompson S, McHugh CP, Clement MHS, Bauermeister S, Maruyama T and Gallacher J (2023) The pursuit of approaches to federate data to accelerate Alzheimer's disease and related dementia research: GAAIN, DPUK, and ADDI. *Front. Neuroinform.* 17:1175689. doi: 10.3389/fninf.2023.1175689

The pursuit of approaches to federate data to accelerate Alzheimer's disease and related dementia research: GAAIN, DPUK, and ADDI

Arthur W. Toga[1*†], Mukta Phatak[2*†], Ioannis Pappas[1], Simon Thompson[3], Caitlin P. McHugh[2], Matthew H. S. Clement[2], Sarah Bauermeister[3], Tetsuyuki Maruyama[2] and John Gallacher[3*‡]

[1]Laboratory of Neuro Imaging, USC Stevens Neuroimaging and Informatics Institute, Keck School of Medicine of USC, University of Southern California, Los Angeles, CA, United States, [2]Alzheimer's Disease Data Initiative, Kirkland, WA, United States, [3]Department of Psychiatry, Warneford Hospital, University of Oxford, Oxford, United Kingdom

There is common consensus that data sharing accelerates science. Data sharing enhances the utility of data and promotes the creation and competition of scientific ideas. Within the Alzheimer's disease and related dementias (ADRD) community, data types and modalities are spread across many organizations, geographies, and governance structures. The ADRD community is not alone in facing these challenges, however, the problem is even more difficult because of the need to share complex biomarker data from centers around the world. Heavy-handed data sharing mandates have, to date, been met with limited success and often outright resistance. Interest in making data Findable, Accessible, Interoperable, and Reusable (FAIR) has often resulted in centralized platforms. However, when data governance and sovereignty structures do not allow the movement of data, other methods, such as federation, must be pursued. Implementation of fully federated data approaches are not without their challenges. The user experience may become more complicated, and federated analysis of unstructured data types remains challenging. Advancement in federated data sharing should be accompanied by improvement in federated learning methodologies so that federated data sharing becomes functionally equivalent to direct access to record level data. In this article, we discuss federated data sharing approaches implemented by three data platforms in the ADRD field: Dementia's Platform UK (DPUK) in 2014, the Global Alzheimer's Association Interactive Network (GAAIN) in 2012, and the Alzheimer's Disease Data Initiative (ADDI) in 2020. We conclude by addressing open questions that the research community needs to solve together.

KEYWORDS

federated data access, Alzheimer's disease and neurodegeneration, data sharing, dementia–Alzheimer's disease, remote data

Introduction

Science is a data-driven economy. Access to high-quality data is the *sine qua non* of creating knowledge and deriving benefit. Data sharing mandates from journals and funders are now requiring studies to make data accessible (Nature Methods, 2023). However, in the health sciences, data access is challenging. For example, in a survey of 3,556 articles from 333 open access biomedical journals, only 7% of corresponding authors responded positively to a data access request, even when their intention to share data was explicitly stated. In this experiment, the revealed preference of 93% of authors was to not share (Gabelica et al., 2022).

For the Alzheimer's disease and related dementias (ADRD) community, these challenges have been recognized for some time, resulting in the development of several data sharing platforms. The first of these was the Laboratory of Neuro Imaging (LONI) (Rex et al., 2003), which supported many data sharing projects including Alzheimer's Disease Neuroimaging Initiative (ADNI) (Mueller et al., 2005) in the Image and Data Archive (IDA) network (Crawford et al., 2015). However, data relevant to the ADRD community goes beyond imaging to include a broad mix of population and clinical cohort data, genetics, experimental medicine data, and randomized trials data. These study designs cover a range of data modalities varying in scale, complexity, and sensitivity, including –omics, imaging, and electronic health records. Platforms for different data modalities include CPAD for trials data, NACC and ALZ-NET for clinical data, NIAGADS for genetic data and FinGen for electronic health record data (**Table 1**). These datasets also vary in governance requirements with some being freely available (open access), others requiring specific permissions (restricted access) and some only being available to the data controller (closed access). More recently multi-modal, multi-cohort repository and analysis platforms have developed to reflect the complexity of Alzheimer's disease. These include the Global Alzheimer's Association Interactive Network (GAAIN) (Neu et al., 2016), the Dementias Platform UK (DPUK) (Bauermeister et al., 2020), and the Alzheimer's Disease Data Initiative (ADDI) (Alzheimer's disease Workbench, 2020). Collaboration between these initiatives has made it apparent that a more general data sharing infrastructure is required to simplify and streamline data access across platforms.

Current solutions are constrained by (i) increasingly complicated data sharing requirements with barriers stemming from institutional, ethical, or legal obligations, (ii) a trend toward bespoke institution-specific platforms that are not designed for data sharing across institutions, and (iii) variability in workflows for the same research question across platforms that can introduce unwanted variation into the findings. These barriers pose significant challenges for data sharing and collaboration. GAAIN, DPUK and ADDI are actively developing a set of innovative solutions that enable data access at scale and pace across platforms to alleviate some of the barriers. Specifically, we offer solutions that:

(1) Resolve the complex pattern of the stakeholder involvement by providing streamlined data sharing agreements designed for use with multi-lateral collaborations.
(2) Provide decentralized data sharing solutions that can operate globally across platforms whether they be institute-specific or institute-agnostic.
(3) Establish universally accessible analysis using workspaces and containerized software that allow the use of standard workflows across platforms.

The two key design principles that underpin delivery of these solutions by these platforms are trust-by-design and data federation (**Figure 1**).

Trust-by-design

Trust-by-design involves a shift away from bespoke bilateral agreements toward trust in a system. Trust is the implicit operating principle underlying collaborative science, simplifying otherwise complex and unpredictable environments, identifying points of certainty around which to organize, and agreeing on a culture where key uncertainties are removed on the basis of mutual agreement, respect, and ethical codes. By facilitating better prediction of the likely reciprocal behavior of others in sharing costs and benefits, trust fosters collaboration. Trust is also foundational to the provenance of data and technologies. As datasets grow in size, complexity, and sensitivity, confidence in the provenance chain becomes increasingly central to the viability of an analysis. Although informal trust-based solutions work well for bilateral collaboration, emerging research questions frequently require multi-lateral collaboration involving large numbers of diverse stakeholders. Multiple actors with multiple interests involving multiple data-sources generate complexity leading to potentially prohibitive transaction costs. Trust-by-design solutions provide the information necessary for accurate and rapid judgments of trustworthiness and scientific value. Here, legal, privacy, security and scientific requirements are embedded within technical and organizational workflows that are explicit, transparent, and fully auditable. This enables systematic streamlining, standardization, and automation. It is trust in systems that underpins federated analysis at scale and pace.

Data federation

Data federation is a mechanism allowing researchers to remotely query a dataset residing at source without ever seeing record-level entries. Within the ADRD community, the data assets that are available are a mix of open, restricted, and closed repositories, with more sensitive data generally being held in restricted or closed environments. Federation requires research questions to be formed into a computation task to be submitted, where the results returned from the submitted task(s), subject to disclosure control where applicable. This pragmatic solution allows data contributors to share datasets that would have otherwise remained inaccessible due to data transfer being undesirable or prohibited due to governance constraints, or unfeasible due to scale. Limitations in federation include data discovery, the wide range of variables, and the diversity of data models used across datasets. Addressing these requires high levels of forward planning

TABLE 1 Summary of dementia related data platforms.

Database	URL	Description	Datasets	Subjects	Reach	Access	Functionality				
							Data discovery	Data analysis	Data federation	Data transfer	Data model
The Global Alzheimer's Association Interactive Network	https://gaaindata.org/	AD-related data platform, federated access to ~n = 500,000 data	61	533,218	Global	Platform application	✓	✓	✓	X	X
Alzheimer's Disease Neuroimaging Initiative	https://ida.loni.usc.edu/login.jsp	Centralized clinical and imaging data from 23 studies	3	3,208	Global	Cohort application	✓	X	X	X	X
Alzheimer's Disease Data Initiative	https://www.alzheimersdata.org/	Centralized clinical and imaging data from 48 studies, enables workspace analysis	48	ongoing	Global	Platform application	✓	✓	✓	✓	C-Surv
Dementias Platform UK	https://www.dementiasplatform.uk/	Centralized clinical and imaging data from 60 studies, enables workspace analysis	60	3.6 m	Global	Platform application	✓	✓	✓	X	C-Surv
The National Institute on Aging Genetics of Alzheimer's Disease Data Storage	https://www.niagads.org/	Genetic data	95	172,701	Global	Cohort application	✓	X	X	X	RS number
National Alzheimer's Coordinating Center	https://naccdata.org/	Centralized ADRC data	1	45,923	United States	Platform application	✓	X	X	✓	X
Critical Path for Alzheimer's Disease	https://c-path.org/programs/cpad/	Clinical Trial data from industry	41	6,500	United States	Platform application	✓	X	X	✓	X
Alzheimer's Network Alzheimer's Association	https://www.alz-net.org/	Real world clinical and imaging data	ongoing	ongoing	Global	Application	✓	✓	X	X	X
Alzheimer's Disease Knowledge Portal	https://adknowledgeportal.synapse.org/	Centralized access to data, provides workspace	12	n/a	Global	Application	X	X	X	X	X
FinnGen	https://www.finngen.fi/en	Centralized data repository	1	589,000	Finland	Application	✓	X	X	X	X
The EU Joint Programme Neurodegenerative Disease Research	https://neurodegenerationresearch.eu/	AD-related worldwide data platform	175	120,000	Global	Cohort application	✓	X	✓	X	X
Integrative analysis of Longitudinal Studies on Aging	https://www.maelstrom-research.org/network/ialsa	Centralized dataset search platform	25	70,000	Global	Cohort application	✓	X	X	X	X
European Progression of Neurological Disease	http://europond.eu/	Models and tool platform	n/a	n/a	Europe	Tool development	X	X	X	X	X
NeuGrid	https://www.neugrid2.eu/	Models and tool platform	n/a	n/a	Europe	Tool development	X	X	X	X	X
European Platform for Neurodegenerative Diseases	https://epnd.org/	AD-related worldwide data platform,	60	120,000	Europe	Platform application	✓	X	X	X	X
Dementias Platform Australia	https://www.dementiasplatform.com.au/	Centralized clinical and imaging data	ongoing	ongoing	Global	Platform application	✓	✓	✓	X	C-Surv
Dementias Platform Korea	https://kdrc.re.kr/eng/about/vision.aspx	Centralized clinical and imaging data	ongoing	ongoing	Republic of Korea	Platform application	✓	✓	✓	X	X
Alzheimer's Disease Data Viewer	https://adata.scai.fraunhofer.de/	Centralized dataset search platform	20	72,372	Global	Platform application	✓	X	X	X	Multiple

10.3389/fninf.2023.1175689

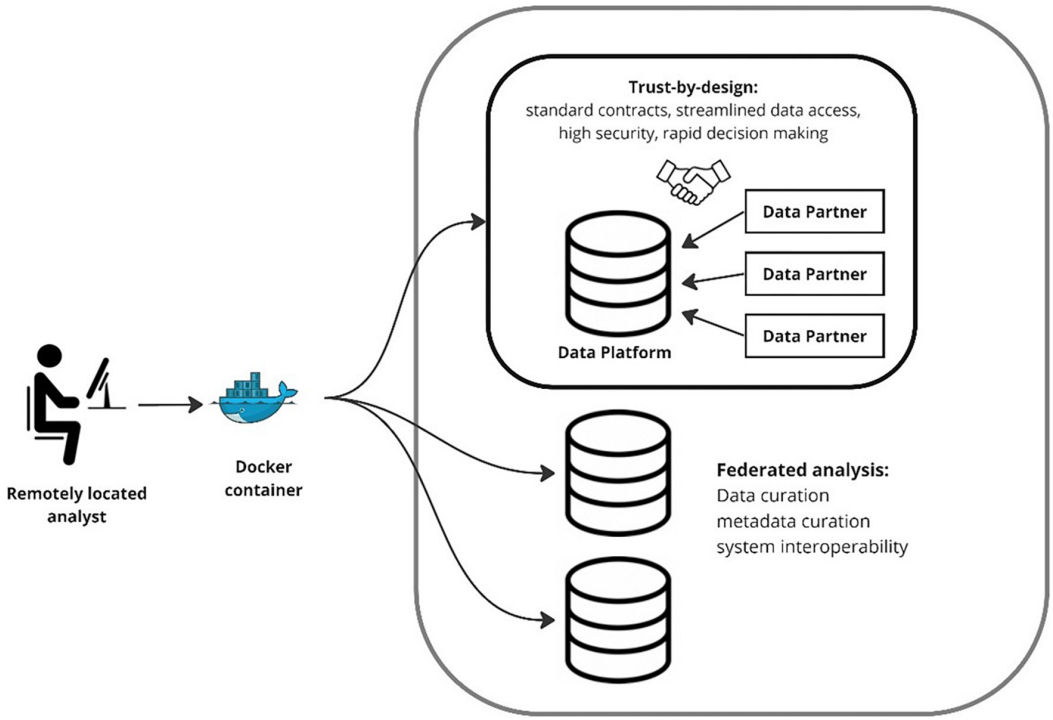


FIGURE 1
Trust by design and data federation.

and coordination. Although these limitations can be mitigated by high quality metadata and the use of standard data models, they remain a challenge for most federated analysis. Limitations apart, by adopting pragmatic strategies and respecting local legal, consent, privacy, and compute concerns, data federation is an increasingly used analysis strategy.

This article describes trust-by-design and federation solutions implemented by GAAIN, DPUK, and ADDI. We conclude by addressing open questions that the research community needs to solve together and by inviting others to join the data sharing movement.

The Global Alzheimer's Association Interactive Network (GAAIN)

Background

Building on the pioneering work of LONI and learnings from ADNI and over 100 other multi-site, multi-modality, cross-sectional and longitudinal studies, GAAIN was the first platform to facilitate data discovery, access, and analysis for ADRD research data. Whilst LONI and ADNI offered centralized imaging data storage, and access (upon approval) to researchers around the world, GAAIN extended the notion of collaborative research in ADRD to a federation model where data can be accessed remotely while preserving data ownership and local, distributed archives. GAAIN addressed concerns of the scientific community regarding data ownership by allowing disease-related data stored in independently operated repositories to be accessed remotely. This enabled data partners to maintain data ownership while providing federated access to users with minimal disruption to the data owners' systems. Since its inception in 2012, the breadth and depth of data accessible through GAAIN has grown to host more than 60 data partners and 500,000 subjects' data from around the world.

Data utilities

Global Alzheimer's Association Interactive Network is optimized for federated analysis and supports exploratory analysis prior to the submission of a formal access request. By bringing together data discovery tools and contact details, GAAIN simplifies the selection of, and access to, datasets. Distinctive features include:

(1) Federated data access and processing wherein data and data repositories of the different data partners stay within their respective infrastructure.
(2) A secure platform of data sharing that is not disruptive to the data partners' systems.
(3) Protection of the policies and ownership of the data.
(4) Directly coupled data exploration and analytics within GAAIN, enabling multi-subject, multi-project, and multi-institutional data aggregation.
(5) Integration of brain imaging metrics via execution of federated processing pipelines.
(6) Federated access to other data platforms that can be connected to DPUK and ADDI.
(7) Harmonization between variables that allows pooling and analysis of different data sets.

Informatics architecture

The GAAIN system architecture comprises a central server that communicates with multiple client applications (Data Partner Clients or DPCs) that are installed in the data partner sites. The DPC is a Java jar file that contains both a light-weight webserver and database (H2 database) that does not disrupt existing systems. This allows GAAIN to remotely connect to data partners without ever having the data stored centrally (federated) unless the data partner decides to do so. The DPC allows remote access to tabular data and brain imaging data. Upon connection to the different DPCs, the data partners appear in the web interface and investigators can explore the available data.

The GAAIN Interrogator (GAAIN, 2017) is the main infrastructure by which investigators can inspect and interact with data in GAAIN. It differs from other data browsing interfaces by allowing dynamic data exploration and visualization through the definition of cohorts. Charts and line graphs make selections visually intuitive where the investigator can choose characteristics from a search range and thus dynamically adjust the cohort definitions. Users can also conduct analysis on the browser using the available data and cohort definitions.

The cohorts of interest can be further used to initiate brain pipelines (for example, executing a brain volumetry analysis on subjects with a certain MMSE range). These brain pipelines are executed remotely on the data partner's site via the DPC in the form of a containerized software (Docker). The results of these pipelines are returned to the interrogator as new variables and can be further analyzed.

Each data partner has complete ownership and control of the data, and data transfer is not required. The data partner signs a non-legal binding Memorandum of Understanding (MOU) before joining GAAIN that formalizes GAAIN's data sharing policies and other terms and conditions of GAAIN participation. Data partners have complete control over data access and display and can disable their DPC at any time thus removing connection to GAAIN.

Summary

Global Alzheimer's Association Interactive Network's data sharing policies and systems are tailored to provide an intuitive and voluntary integration of multiple Alzheimer's disease data repositories within a common sharing network. Combining data from different data partners requires infrastructure like that implemented in GAAIN but also appropriate ontologies to enable cross-cohort search and data aggregation.

Dementias Platform UK (DPUK)

Background

At inception, the DPUK Data Portal was designed to facilitate access to UK population and clinical cohort data. It has since developed to provide an end-to-end data management service for cohorts, clinical studies, trials, and systematic reviews. Currently it facilitates access to 60 cohorts representing individual-level data for 3.6 million participants.

Data utilities

The DPUK Data portal is optimized for multi-modal pooled analysis enabling epidemiologic, imaging, genetic, proteomic, and clinical data to be combined. However, it also has federated analysis capability. Distinctive features include:

(1) Curation of data to research readiness using common standards according to modality (Bauermeister et al., 2023).
(2) A suite of data discovery tools.
(3) Centralized management of access requests.
(4) Personal analysis space with a wide range of standard and specialist software packages.
(5) Data hubs for specialist research groups and consortia.
(6) Synthetic data for preliminary model testing.
(7) Federated access to other data platforms including GAAIN, ADDI, DPAU (Dementias Platform Australia), and Korea Dementia Research Center (Korea Dementia Research Center [KDRC]).

Informatics architecture

The Data Portal operates within the UK Secure eResearch Platform (SeRP, 2006) environment according to ISO 27001 (SeRP, 2006) as a data processor according to General Data Protection Regulation [GDPR] (2016) and Legislation.gov.uk (2018). Data may be accessed remotely for *in situ* analyses but not downloaded to third-party sites. Data-use approval remains with the cohort research teams who retain control over data access. Preparing datasets for third-party researchers and providing suitable documentation is resource intensive.

Summary

The preparing of datasets by data contributors for third-party researchers is resource intensive. The Data Portal reduces this burden through the management of access requests on behalf of cohort research teams, the use of a common data model, and the development of standard documentation for data stored within the Data Portal. This obviates the need for repeated data transfer and pre-processing. The UKSeRP environment has been designed for use with linked electronic health records and is a suitable environment for the onward sharing of linked data.

Alzheimer's Disease Data Initiative (ADDI)

Background

Alzheimer's Disease Data Initiative's mission is to accelerate AD research by enabling collaborative data sharing and analysis. ADDI's trust-by-design solution is the Alzheimer's disease (AD) Workbench (Alzheimer's disease Workbench, 2020). The AD Workbench delivers access to key datasets around the world across public and private sectors using a secure cloud-based data platform. The AD Workbench provides data contributors with flexible data sharing options that preserve their branding and maintains their control and autonomy over the data through configurable data access requests and approval workflows. Along with storing some datasets locally, ADDI has achieved interoperability with DPUK, EPND (Bose et al., 2022), GAAIN and Vivli (Vivli, 2013). To make federated solutions accessible, ADDI has developed the Federated Data Sharing Appliance (FDSA), an option that offers both data providers and researchers a streamlined interface to access, maintain and query data where it resides.

Data utilities

The AD Workbench is optimized for federated analysis. Distinctive features include:

(1) The FDSA is agnostic to data type. Currently querying is available on structural data.
(2) The FDSA is a stand-alone Linux application installable on local data provider's IT environment and deployable on any infrastructure.
(3) Data contributors determine the level of permissioned access to the record-level data that is granted to researchers.
(4) The administrative module gives data contributors a point-and-click interface from which they can manage researcher access, data contributors:
    - Can review and approve Data Access Requests from users.
    - Have visibility of all research query tasks and task status.
    - Can verify, after execution, that the results do not include record-level data.
(5) Submission (upon approval) of container-based analyses across multiple platforms.

Informatics architecture

The federated dataset must be stored as a PostgreSQL database. FDSA seamlessly connects to the datasets. Docker is used to execute user-submitted tasks. FDSA includes a GUI Administrative module running on an onboard web server and a back-end service that manages and executes admin and end user tasks. A common set of research APIs can be used to access the data. FDSA requires minimal infrastructure for installation: 2 CPUs, 8GB of memory, and 100GB of storage.

Summary

With this solution, ADDI has enabled sharing for data contributors who previously were unable to make their data available to researchers. FDSA is installable on-premise and is suited for a diverse range of datasets, data contributor, and data consumer needs. Under active development, FDSA will continue to add features, such as (1) trusted containers that do not necessitate manual review from data contributors, (2) a secure way for FDSA instances to communicate for combined analyses of federated datasets, and (3) a way for users to share models and analyses via containers with the community. ADDI's federated solution removes another barrier to permissioned data access and further enables the research community to make novel discoveries by unlocking access to previously unreachable datasets.

The wider environment

There are many variations on the theme of data sharing and any attempt to compile a comprehensive list will certainly be incomplete. Nonetheless, we endeavored to summarize Alzheimer's-related data initiatives and identify similarities and differences to support analysts in considering which platforms are most relevant to their research question (**Table 1**). Overall, platforms follow a centralized model with various access tiers (open, restricted, closed) and high-level data discovery tools. Few platforms, however, provide data curation/metadata curation or access to computing resources.

The way ahead

These platforms have several commonalities. They are working together to support the FAIR principles of data management (Findable, Accessible, Interoperable and Reusable) (Wilkinson et al., 2016), and are working toward the Dublin Core metadata specification (Dublin Core, 1994). For federated computation, all platforms support the GA4GH Task Execution Standard (GA4GH, 2013) as a suitable candidate for the containerization of analyses. Their collaboration also allows for the automated creation of containers to support standard analyses. Nevertheless, each platform provides distinctive data access options, recognizing that insisting on a single data platform for all use cases would stifle innovation, whilst agreement on commonly used standards facilitates collaboration. The use of federation alongside standardized analysis can further render the data reusable with researchers continuously accessing and processing the data from multiple studies in a similar way. This can have tremendous impact on ADRD scientific discovery where previously unseen

relationships can now emerge via a unified access and analysis model (Neu et al., 2017).

However, federated data sharing is not without challenges. Further integration across platforms is focused on widening the availability of data for federated approaches and work has begun on a framework where datasets from each of the three platforms can be discovered from within the others. GAAIN datasets can be discovered from ADDI and, upon approval, data can be transferred to the AD Workbench. Datasets from DPUK are requestable from the AD Workbench for access at federated level. Handling of multimodal data is a key challenge for a comprehensive federated model. Integration of clinical, neuroimaging and genetics data is essential. To give researchers access to multimodal data, GAAIN has made efforts toward this direction by establishing external connections to the IDA network (Crawford et al., 2015) that hosts a variety of studies. This effort can be augmented by ADDI's AD Workbench tools that are agnostic to any data format. In addition, DPUK has already established ontologies and multi-modal analysis that can be further integrated with ADDI's products. Key to progress is increasing cross-platform interoperability through data standards, efficient data access, and distribution of computational workload.

Implementing data standards across platforms would be transformative. A common ontology (data model) alleviates the data preparation burden for researchers and developers. A recent study comparing data preparation times for 25 variables in two cohorts found that using the 'bespoke' cohort designed data model required 5–6 h per cohort, whilst using a standard data model reduced this time to 30 min per cohort (Bauermeister et al., 2023). Standard ontologies also simplify the building of data discovery tools for developers, as standard metadata models enable tools such as data dictionaries to have broader application across datasets and be harmonized across platforms. However, data standards require consensus, and this will vary according to data modality. GAAIN, DPUK, and ADDI are working together to identify, develop, and test potential data models according to data modality. For example, the ontology implemented by DPUK can be integrated with GAAIN and ADDI.

Analyses conducted in federated settings pose unique opportunities and challenges for data access. Federated approaches increase the potential base of data enabling the design of purpose-specific cohorts, i.e., using existing data to create cohorts designed to address specific research questions. An example of this is the GAAIN Interrogator tool (GAAIN, 2017), a web-based application that allows users to query and explore distributed datasets related to Alzheimer's disease and other neurodegenerative disorders. These cohorts can be characterized using persistent unique identifiers enabling rapid replication and re-purposing. A challenge, however, is the efficient running of models across diverse datasets and informatics architectures. A solution under development within the consortium is the creation of synthetic datasets (Muniz-terrera et al., 2021) that model the characteristics of the original data. These can be used to develop task-specific 'boilerplate' code that is known to operate successfully across platforms and to test the functionality of models across platforms prior to a formal analysis. For higher-order data (imaging and genomics) this approach is time and computationally efficient. By running and verifying models on simulated data, researchers can spend less time submitting federated queries and data providers may have fewer queries to review.

The federated analysis framework has increasing potential for federated learning applications. In federated learning, different data partners/clients train their neural networks, and a central model aggregates the parameters of that model (Rieke et al., 2020). This approach allows the training of large-scale neural network models without the need to access centralized data (Stripelis et al., 2022).

We hope in the future to release a skeleton of the underlying federated analysis framework from these platforms. By doing so, outside researchers can build their own federated methods and models and those can potentially be integrated with the proposed platforms.

Computational workload becomes an increasingly important resource constraint as the scale of datasets grows with a commensurate increase in the complexity of research questions. For federated analyses there is a requirement to establish models of 'smart' federation where requests and computational load can be efficiently managed. Computational and labor-intensive burdens on data providers lead to bottlenecks and longer turnaround times to review and fulfill data access requests. Additionally, extracting information from large cohorts of interest requires increased computational resources. GAAIN, DPUK and ADDI provide distinctive solutions to this challenge, each based around the functionality of its trust-by-design architecture. ADDI's AD Workbench cohort information can be utilized within a user's workspace to create analysis only for this cohort. To increase efficiency and transparency, requests and datasets need to be in specific format before computational resources are allocated. GAAIN is also working on identifying a specific format for how these requests can be more efficient in terms of how they allocate resources. In DPUK, the analysis plan determines the computational resource that is allocated to the project. All platforms are working toward modality specific formats that can be integrated within docker containers.

A further challenge is the scrutiny of findings to ensure that data or personally identifiable information is not observed or downloaded. Safeguards that prevent this from happening can include presentation of summary statistics rather than single data points or preventing analysis of cohorts consisting of few subjects (Neu et al., 2016). Currently this is an arduous task on all platforms; a solution that does not scale and is vulnerable to error. Machine learning provides a potential solution for automating the management of data leakage risk (Shabtai et al., 2012).

Data availability statement

The original contributions presented in this study are included in the article/supplementary material, further inquiries can be directed to the corresponding authors.

Author contributions

AT, MP, and JG contributed to conceptualization and initial draft of the manuscript. All authors contributed to the writing, reviewing, and editing and approved the submitted version.

Funding

SB was supported by Dementias Platform UK (DPUK) funded by Medical Research Council (MRC) MR/T0333771. This work was supported by the Global Alzheimer's Association Interactive Network Initiative of the Alzheimer's Association (SG-20-678486-GAAIN2).

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

References

Alzheimer's disease Workbench (2020). *Alzheimer's disease data initiative*. Available online at: https://www.alzheimersdata.org/ (accessed April 1, 2023).

Bauermeister, S., Bauermeister, J., Bridgman, R., Felici, C., Newbury, M., and North, L. (2023). Research-ready data: the C-Surv data model. *Eur. J. Epidemiol.* 38, 179–187. doi: 10.1007/s10654-022-00916-y

Bauermeister, S., Orton, C., Thompson, S., Barker, R., Bauermeister, J., Ben-Shlomo, Y., et al. (2020). The dementias platform UK (DPUK) data portal. *Eur. J. Epidemiol.* 35, 601–611. doi: 10.1007/s10654-020-00633-4

Bose, N., Brookes, A., Scordis, P., and Visser, P. (2022). Data and sample sharing as an enabler for large-scale biomarker research and development: the EPND perspective. *Front. Neurol.* 13:1031091. doi: 10.3389/fneur.2022.1031091

Crawford, K. L., Neu, S. C., and Toga, A. W. (2015). The image and data archive at the laboratory of neuro imaging. *Neuroimage* 124(Pt B), 1080–1083. doi: 10.1016/j.neuroimage.2015.04.067

Dementias Platform Australia *Dementias platform Australia*. Available online at: https://www.dementiasplatform.com.au/ (accessed April 1, 2023).

Dublin Core (1994). *Dublin core metadata initiative*. Available online at: https://www.dublincore.org/ (accessed April 1, 2023).

GA4GH (2013). *Task execution standard*. Available online at: https://www.ga4gh.org (accessed April 1, 2023).

GAAIN (2017). *The global Alzheimer's association interactive network (GAAIN)*. Available online at: https://www.gaaindata.org/data3/explore/login.jsp (accessed April 1, 2023).

Gabelica, M., Bojčić, R., and Puljak, L. (2022). Many researchers were not compliant with their published data sharing statement: a mixed-methods study. *J. Clin. Epidemiol.* 150, 33–41. doi: 10.1016/j.jclinepi.2022.05.019

General Data Protection Regulation [GDPR] (2016). *General data protection regulation (GDPR)*. Available online at: https://gdpr.eu/tag/gdpr/ (accessed April 1, 2023).

Korea Dementia Research Center [KDRC] *Finding the best way for preventing and treating dementia: through standardized data-gathering, high-quality research, and efficient utilization*. Available online at: https://kdrc.re.kr/eng/main/main.aspx (accessed April 1, 2023).

Legislation.gov.uk (2018). *UK data protection act*. Available online at: https://www.legislation.gov.uk/ukpga/2018/12/contents/enacted (accessed April 1, 2023).

Mueller, S., Weiner, M., Thal, L., Petersen, R., Jack, C., Jagust, W., et al. (2005). Ways toward an early diagnosis in Alzheimer's disease: the Alzheimer's disease neuroimaging initiative (ADNI). *Alzheimer's Dement.* 1, 55–66. doi: 10.1016/j.jalz.2005.06.003

Muniz-terrera, G., Mendelevitch, O., Barnes, R., and Lesh, M. (2021). Virtual cohorts and synthetic data in dementia: an illustration of their potential to advance research. *Front. Artif. Intell.* 4:613956. doi: 10.3389/frai.2021.613956

Nature Methods (2023). Data sharing is the future. *Nat. Methods* 20:471. doi: 10.1038/s41592-023-01865-4

Neu, S., Crawford, K., and Toga, A. (2016). Sharing data in the global Alzheimer's association interactive network. *Neuroimage* 124(Pt B), 1168–1174. doi: 10.1016/j.neuroimage.2015.05.082

Neu, S., Pa, J., Kukull, W., Beekly, D., Kuzma, A., Gangadharan, P., et al. (2017). Apolipoprotein E genotype and sex risk factors for alzheimer disease: a meta-analysis. *JAMA Neurol.* 74, 1178–1189. doi: 10.1001/jamaneurol.2017.2188

Rex, D., Ma, J., and Toga, A. (2003). The LONI pipeline processing environment. *Neuroimage* 19, 1033–1048. doi: 10.1016/s1053-8119(03)00185-x

Rieke, N., Hancox, J., Li, W., Milletarì, F., Roth, H., and Albarqouni, S. (2020). The future of digital health with federated learning. *Digit. Med.* 3:119. doi: 10.1038/s41746-020-00323-1

SeRP (2006). *SeRP | UK*. Available online at: https://serp.ac.uk/serp-uk/ (accessed April 1, 2023).

Shabtai, A., Elovici, Y., and Rokach, L. (2012). *A survey of data leakage detection and prevention solutions*, 1st Edn. New York, NY: Springer.

Stripelis, D., Gupta, U., Saleem, H., Dhinagar, N., Ghai, T., Sanchez, R., et al. (2022). Secure Federated Learning for Neuroimaging. *arXiv* [preprint]. doi: 10.48550/arXiv.2205.05249

Vivli (2013). *A global clinical research data sharing platform*. Available online at: https://vivli.org (accessed April 1, 2023).

Wilkinson, M. D., Dumontier, M., Aalbersberg, I. J., Appleton, G., Axton, M., Baak, A., et al. (2016). The FAIR guiding principles for scientific data management and stewardship. *Sci. Data* 3:160018. doi: 10.1038/sdata.2016.18

frontiers | Frontiers in Neuroinformatics

TYPE Technology and Code
PUBLISHED 19 June 2023
DOI 10.3389/fninf.2023.1207721

OPEN ACCESS

EDITED BY
Christian Haselgrove,
UMass Chan Medical School, United States

REVIEWED BY
David Haynor,
University of Washington, United States
Bo-yong Park,
Inha University, Republic of Korea

*CORRESPONDENCE
Dylan Martin
✉ dmartin99@gsu.edu

RECEIVED 18 April 2023
ACCEPTED 02 June 2023
PUBLISHED 19 June 2023

CITATION
Martin D, Basodi S, Panta S, Rootes-Murdy K, Prae P, Sarwate AD, Kelly R, Romero J, Baker BT, Gazula H, Bockholt J, Turner JA, Esper NB, Franco AR, Plis S and Calhoun VD (2023) Enhancing collaborative neuroimaging research: introducing COINSTAC Vaults for federated analysis and reproducibility. *Front. Neuroinform.* 17:1207721. doi: 10.3389/fninf.2023.1207721

Enhancing collaborative neuroimaging research: introducing COINSTAC Vaults for federated analysis and reproducibility

Dylan Martin[1]*, Sunitha Basodi[1], Sandeep Panta[1], Kelly Rootes-Murdy[1], Paul Prae[1], Anand D. Sarwate[1,2], Ross Kelly[1], Javier Romero[1], Bradley T. Baker[1], Harshvardhan Gazula[3], Jeremy Bockholt[1], Jessica A. Turner[1], Nathalia B. Esper[4], Alexandre R. Franco[4,5,6], Sergey Plis[1] and Vince D. Calhoun[1]

[1]Tri-institutional Center for Translational Research in Neuroimaging and Data Science, Georgia State, Georgia Tech, Emory, Atlanta, GA, United States, [2]Department of Electrical and Computer Engineering, Rutgers University–New Brunswick, Piscataway, NJ, United States, [3]Athinoula A. Martinos Center for Biomedical Imaging, Massachusetts General Hospital and Harvard Medical School, Boston, MA, United States, [4]Center for the Developing Brain, Child Mind Institute, New York, NY, United States, [5]Center for Brain Imaging and Neuromodulation, Nathan Kline Institute for Psychiatric Research, Orangeburg, NY, United States, [6]Department of Psychiatry, NYU Grossman School of Medicine, New York, NY, United States

Collaborative neuroimaging research is often hindered by technological, policy, administrative, and methodological barriers, despite the abundance of available data. COINSTAC (The Collaborative Informatics and Neuroimaging Suite Toolkit for Anonymous Computation) is a platform that successfully tackles these challenges through federated analysis, allowing researchers to analyze datasets without publicly sharing their data. This paper presents a significant enhancement to the COINSTAC platform: COINSTAC Vaults (CVs). CVs are designed to further reduce barriers by hosting standardized, persistent, and highly-available datasets, while seamlessly integrating with COINSTAC's federated analysis capabilities. CVs offer a user-friendly interface for self-service analysis, streamlining collaboration, and eliminating the need for manual coordination with data owners. Importantly, CVs can also be used in conjunction with open data as well, by simply creating a CV hosting the open data one would like to include in the analysis, thus filling an important gap in the data sharing ecosystem. We demonstrate the impact of CVs through several functional and structural neuroimaging studies utilizing federated analysis showcasing their potential to improve the reproducibility of research and increase sample sizes in neuroimaging studies.

KEYWORDS

COINSTAC, neuroimaging, federated learning, reproducibility, open science, datasets, privacy, collaborative analysis

1. Introduction

In recent years, neuroimaging has seen a growing emphasis on data sharing and collaborative research, as evidenced by the development of new standards [e.g., Brain Imaging Data Structure (BIDS), Gorgolewski et al., 2016], open-source software tools, and data repositories. Neuroinformatics consortia such as Enhancing NeuroImaging Genetics through Meta-Analysis consortium (ENIGMA) (Thompson et al., 2014), and data repositories such as OpenNeuro (Markiewicz et al., 2021) and National Institutes of Health Data Archive,[1] were created to facilitate analysis of data and combining data from multiple sites. Pooling data from many studies allows for larger sample sizes that produce more statistical power (Biswal et al., 2010; Andrade, 2020). Though the quantity of neuroimaging data is increasing, there are still barriers to collaboration in the form of technological, policy, administrative, and methodological constraints that can negatively affect data accessibility.

In this section, we discuss in detail some of the challenges associated with collaborative analysis, particularly in centralized approaches, where the data need to be pooled in one location to perform an analysis. We also discuss COINSTAC, a tool built on the principles of federated analysis to enable analysis without the need to centralize data.

1.1. Technological challenges

Technological constraints, such as storage space, download speed, and processing power, play a significant role in the feasibility of performing collaborative analyses on large datasets (Homer et al., 2008; McGuire et al., 2011) such as neuroimaging data. Existing data repositories can contain high-resolution neuroimaging files covering tens of thousands of subjects, with sizes ranging from megabytes to multiple petabytes. Downloading the MPI-Leipzig Mind-Brain-Body dataset (Babayan et al., 2022) (369.78 GB) at the global median download speed of 76.32 Mbps[2] onto a modern MacBook Pro with 512 GB of storage[3] would take 11 h and 33 min, consuming 72.2 percent of the machine's total storage space. The requirements for storage space and download time can increase when an analysis involves aggregating multiple large datasets. Additionally, processing power may be a limiting factor for performing computations, particularly when certain types of analyses are designed to run on specific hardware like GPUs, which can demand resources beyond the capacity of smaller research groups or institutions with limited budgets.

1.2. Policy and privacy challenges

Due to the potentially sensitive nature of neuroimaging datasets, their use in collaborative analysis is often restricted by policies intended to preserve privacy. Collaboration methods include aggregating data in a centralized repository or using Data Usage Agreements (DUAs) (Thompson et al., 2014, 2017). These methods can be cumbersome and, in some cases, insufficient. DUAs may take months or even years to approve without any guarantee of the data's utility. Data sharing might be limited by law, policy, or proprietary restrictions, largely driven by re-identification concerns. In situations where only summary data can be shared, differences in analysis methodology may result in inconsistent measures for meta-analysis (Rootes-Murdy et al., 2022).

1.3. Administrative challenges

Administrative challenges can arise when collaborating on an analysis, as various steps demand researchers' time and attention. These steps may include communicating between agencies, formulating and signing data-sharing agreements, agreeing on data preparation and analysis processes, procuring technical resources, monitoring and auditing processes, performing data transfer, initiating computations, disseminating results of analyses, and so on.

The efficiency of collaborative analysis is influenced by how quickly these manual steps are executed. Synchronized availability of researchers can present a barrier to the collaboration process. When researchers work asynchronously, each step in a serial process requiring manual interaction introduces potential delays. This can be particularly challenging when researchers are distributed across multiple time zones or have limited time to perform manual tasks. Furthermore, researchers' availability may be constrained by the need for expertise and authority, such as having the authority to sign a data-sharing agreement or the technical expertise to run the appropriate Python script against a dataset. Often, these manual steps must be executed for each new analysis, which can slow down and even impede collaborative analysis. By addressing these administrative barriers, research teams can more effectively collaborate and streamline their analysis processes, ultimately contributing to the advancement of neuroimaging research.

1.4. Methodological differences

Variability in methodological approaches to data processing and analysis can make reproducing studies challenging (Vogt, 2023). To validate results, researchers must adhere to the exact methodology used in the original study, which necessitates clear communication of the specific methods employed. However, as methods are often chosen on a case-by-case basis, replicating studies can be time-consuming and difficult (Esteban et al., 2019), and sometimes even impossible. Moreover, when multiple studies adopt different methodologies, combining their results meaningfully becomes challenging, hindering the execution of meta-analyses.

1 https://nda.nih.gov/

2 https://www.speedtest.net/global-index

3 https://www.apple.com/macbook-pro-14-and-16/specs/

To overcome these barriers, we introduce COINSTAC,[4] a tool that supports federated analysis for neuroimaging data.

1.5. Federated analysis using COINSTAC

Federated analysis (also federated learning, or decentralized analysis) (Plis et al., 2016; Kairouz et al., 2021; Rootes-Murdy et al., 2022) allows for multiple datasets to be used in analyses without source data being directly shared. Instead, data holders run computations on their local data and only share the outputs, which are often group-level data derivatives or summary statistics. For example, sites may compute an average or other summary on their local data and send that information. Typically, these summaries are much smaller, meaning that the source data are not shared, thereby removing the technical challenges associated with dataset transfer. A second potential benefit is additional privacy guarantees for the data holders. From a purely policy perspective, datasets are analyzed without being moved from their original location and data holders can determine which computations are and are not allowed on their data. From a technical perspective, strong end-to-end encryption can prevent third parties from acquiring the data derivatives. Depending on the trust model, additional privacy protections are possible, including emerging technologies like secure multiparty computation and differential privacy (Dwork and Roth, 2013; Bonawitz et al., 2016, 2017; Heikkilä et al., 2020; Imtiaz et al., 2021; Senanayake et al., 2022).

The Collaborative Informatics and Neuroimaging Suite Toolkit for Anonymous Computation (COINSTAC) (see text footnote 4) (Plis et al., 2016; Ming et al., 2017; Gazula et al., 2020, 2023; Turner et al., 2022) is a tool developed to support federated analysis specifically for neuroimaging data by overcoming the aforementioned barriers to collaboration through the use of federated analysis and standardization of collaboration methods. COINSTAC enables researchers to run decentralized neuroimaging analyses to perform larger collaborative studies (Rootes-Murdy et al., 2022; Turner et al., 2022). As of now, COINSTAC has attracted 115 users and has been downloaded 2,386 times, showcasing its growing reach and impact within the research community.

The COINSTAC desktop application provides an easy-to-use graphical user interface (GUI) for coordinating and executing federated analysis pipelines among multiple collaborators. Image preprocessing and a variety of univariate and multivariate approaches (e.g., VBM regression, group ICA) can be completed within the app.

For a comprehensive understanding of COINSTAC, its functionalities, and usage, readers are encouraged to refer to the following papers (Plis et al., 2016; Ming et al., 2017; Gazula et al., 2020, 2023; Turner et al., 2022).

One limitation of the original implementation of COINSTAC is that it requires synchronized coordination (Jwa and Poldrack, 2022), users have to coordinate among data owners to confirm their systems are online, that the data are organized within the same structure and that the data are mapped properly within the COINSTAC system. The need for a centralized coordinator can delay contingent analyses.

In this paper, we address this limitation by showcasing a method for hosting both private and public datasets where the datasets are persistently accessible for analysis using COINSTAC without the need for synchronized effort from data owners. Analysis of public datasets is made more accessible by removing the need to find, download, preprocess, and prepare datasets for analysis. We provide curated data vaults for various openly available neuroimaging data which COINSTAC users can simply include in their analyses. Access to private datasets can be restricted to a list of computations approved by the vault owner. Standardizing access to data vaults in the COINSTAC system simplifies analysis, optimizes computational performance, and promotes the reusability of neuroimaging datasets.

2. Method

In this section, we discuss COINSTAC and the extension of the COINSTAC framework with the addition of vaults, their architecture, and various use-cases they enable. All code for COINSTAC and COINSTAC Vaults can be found in the COINSTAC Github repository.[5]

2.1. COINSTAC

To understand how Vaults improve the workflow of federated analysis in COINSTAC, we will describe the COINSTAC system and how it is used.

The main components of the COINSTAC system are: the desktop application, the central server, and computation containers. The desktop application provides a graphical user interface (GUI) and manages local computation containers used to participate in federated analyses. The central server manages the central database and runs the containers that act as the inner node in federated analyses.

In the COINSTAC desktop application, users join collections of users called "consortia" to collaborate on an analysis pipeline. A consortium is a group formed by individual COINSTAC users, each with their machine that is capable of being a node in a federated analysis pipeline. Each member within a consortium will act as a node in the federated analysis group by running local computations inside of a container on their system.

The following is how a researcher would use the COINSTAC user interface to create a consortium and run a federated analysis pipeline:

- Log in as a user
- Join (as a member) or create (as an owner) a consortium
- Configure a set of computations (a pipeline) to be performed by a consortium
- Map their local data to the pipeline
- Initiate the pipeline (a run)
- View the results of the pipeline run.

4 https://coinstac.org/

5 https://github.com/trendscenter/coinstac

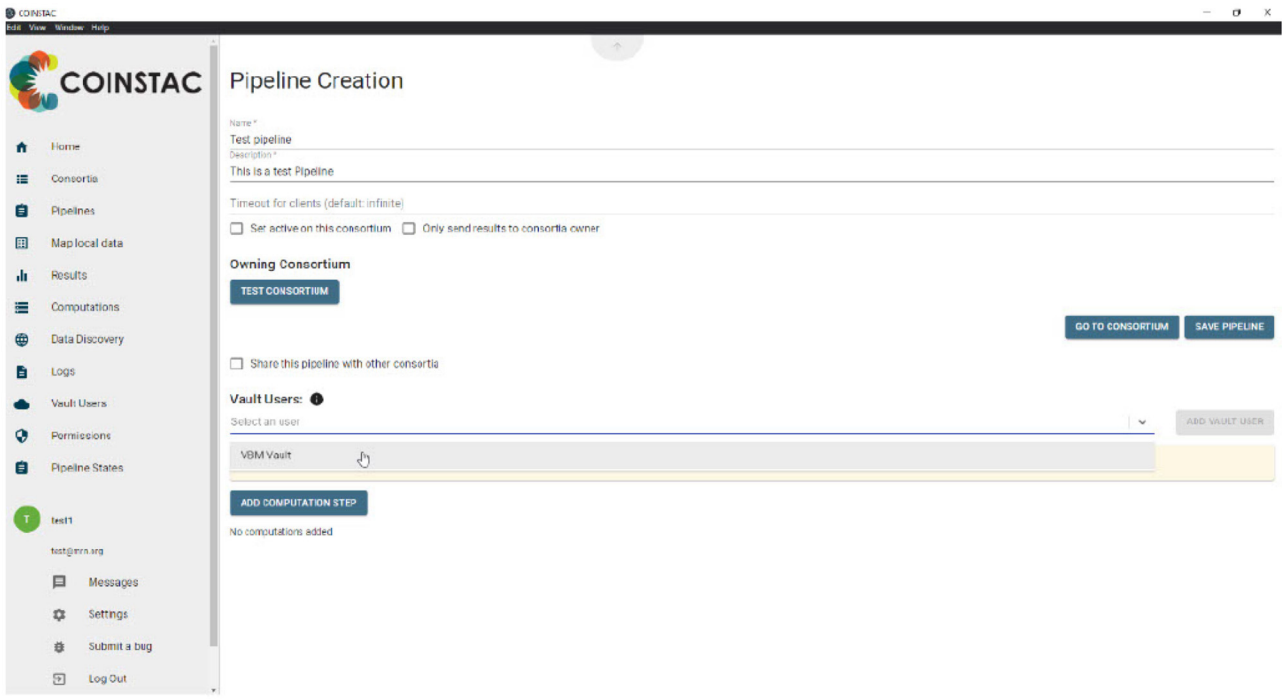


FIGURE 1
Adding vault data to an analysis pipeline.

2.2. COINSTAC Vaults

2.2.1. Purpose and high level overview

The Vaults system is an extension of the COINSTAC platform that allows datasets to be persistently available for participation in federated analyses without requiring manual action from data owners apart from the initial setup. COINSTAC consortium owners can independently add Vaults members to their consortia, allowing vault datasets to participate in federated analyses without the need for coordination between consortia owners and Vault data owners. The Vault client allows datasets to be made available to the larger COINSTAC ecosystem, giving the ability for others to run pipelines using the Vault's data without it ever leaving its respective system.

2.2.2. Using the GUI to add a Vault to a consortium and run an analysis

Vault clients can be added to a consortium by a consortium owner without any action required from the owner of the Vault data, as shown in Figure 1.

2.2.3. Hosting Vaults

Making datasets available for federated analysis through COINSTAC is simple using Vaults. Vaults can be hosted in a variety of compute environments such as: on personal machines, on-premises servers, on a cluster of compute nodes, or in a virtual cloud. Both publicly available datasets and private datasets can be made available to the COINSTAC platform via Vaults. COINSTAC consortia can include any combination of diverse types of data: public and private datasets, data hosted on local machines, Vaults hosted by the Tri-institutional Center for Translational Research in Neuroimaging and Data Science (TReNDS), and third-party Vaults connected to COINSTAC as shown in Figure 2.

In addition to TReNDS-hosted vaults, data owners are able to host their own (public or private) data as Vaults (Figure 3) by using the coinstac-vault-client software package at https://www.npmjs.com/package/coinstac-vault-client.

The process for hosting a dataset in a Vault is described below:

- Install the Vault client: The user installs the Vault client on their host machine.
- Request Vault integration: The user submits a request to the COINSTAC team for integrating the Vault into the COINSTAC ecosystem.
- Receive API keys: The COINSTAC team provides the user with the necessary API keys for the user's Vault client.
- Configure dataset directory: The user specifies the local directory containing the dataset in the Vault client configuration.
- Select approved computations: The user chooses a list of computations, granting permission for these computations to be executed on their vault data.

After this process, the Vault becomes available for use in the COINSTAC system. Consortium owners can select to include the Vault in their consortium and perform federated analysis using Vault data. Whether the data was downloaded from a public repository or collected privately, the process is the same for both types of data since the source data stays on the user's local machine.

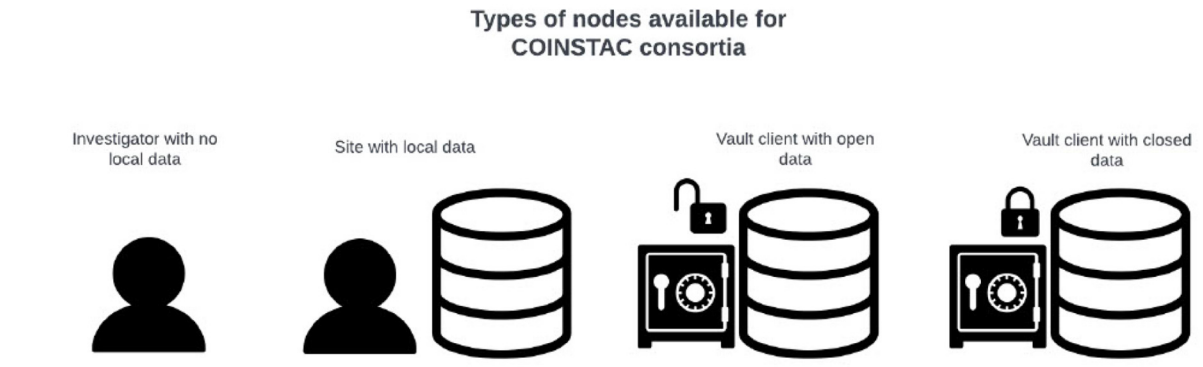


FIGURE 2
Different types of participants interacting with COINSTAC.

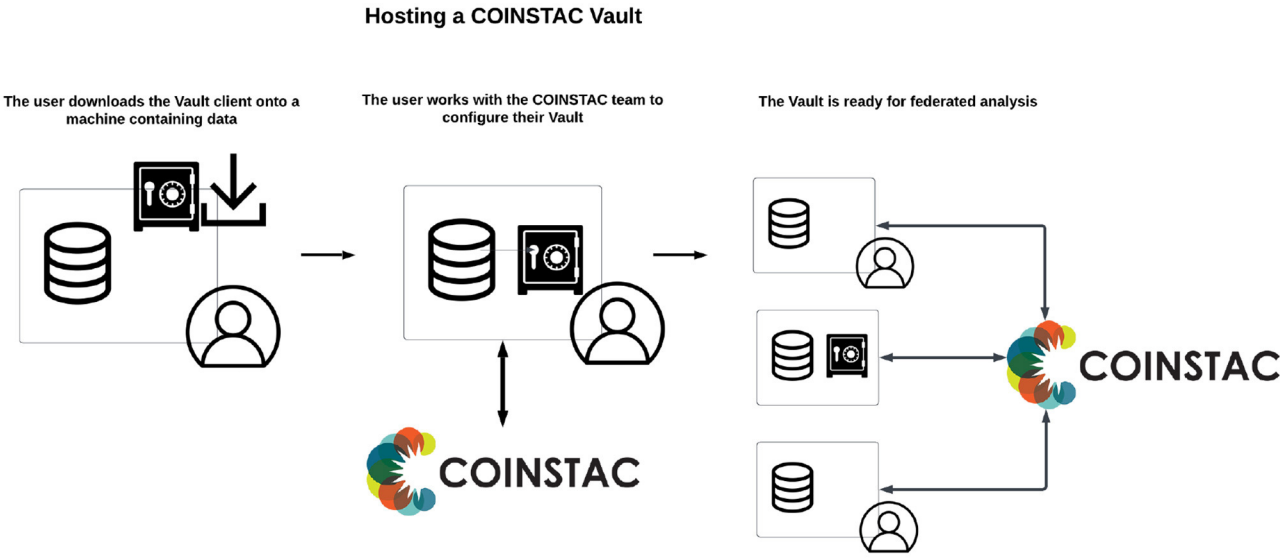


FIGURE 3
Process of creating a vault in COINSTAC.

2.2.4. Vault architecture overview

The Vault client software package is built upon the same core code as the COINSTAC desktop application to manage containers and execute computation pipelines. However, it omits the user interface (UI) component and includes additional code that enables the client to be persistently online and available. The desktop application has been modified to allow consortium owners to add Vault clients to their consortium via the GUI.

The Vault client is a NodeJS server running on the local machine, responsible for maintaining a persistent connection with the COINSTAC system using the coinstac-vault-client package. The server communicates with the COINSTAC central server using websockets and HTTP protocols. It manages the life-cycle of containers (Docker, Singularity) through the coinstac-container-manager package, which is responsible for isolating and executing the computations within the federated analyses. The Vault client also utilizes other core COINSTAC libraries such as coinstac-client-core, coinstac-client-server, coinstac-pipeline, and coinstac-common, all of which are npm packages, to ensure seamless integration with the COINSTAC ecosystem. An overview is shown in Figure 4.

Message passing, which is an integral part of federated analyses, is handled by the Vault client using MQTT (MQ Telemetry Transport) and HTTP protocols. MQTT is a lightweight messaging protocol optimized for high-latency or unreliable networks.

For pipeline runs in consortia that only use Vaults, the result data is uploaded to a secure Amazon S3 bucket, which can then be downloaded by consortium members using the desktop application. This ensures that the results are securely stored and easily accessible by authorized users.

In summary, the Vault architecture in COINSTAC improves the overall efficiency and user experience of performing federated analyses. By maintaining a persistent connection, the Vault client

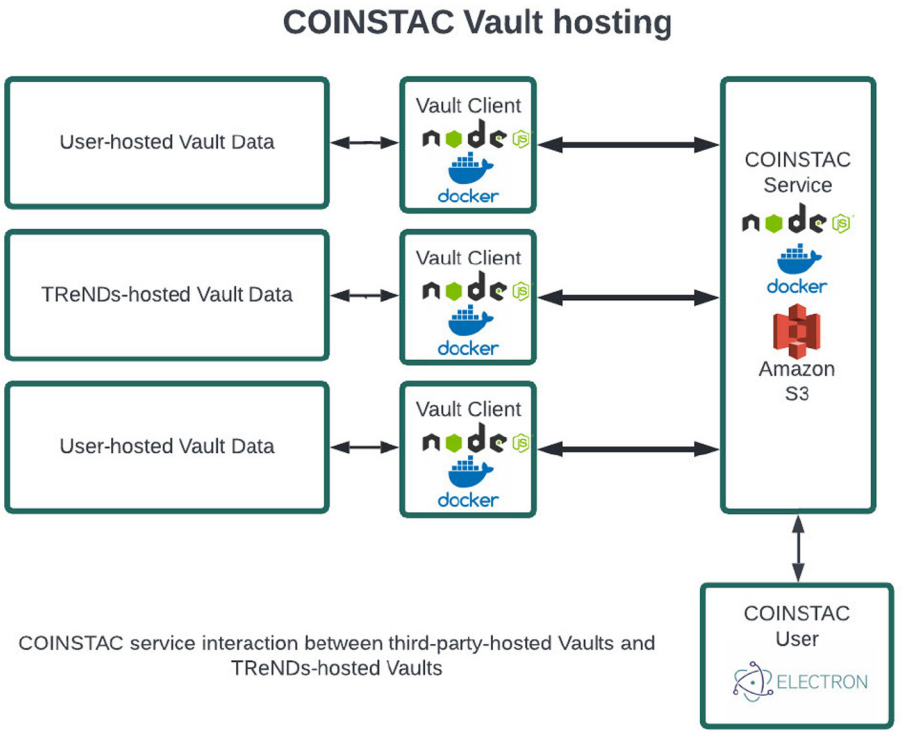


FIGURE 4
Architecture of vaults in COINSTAC.

ensures that datasets are readily available for analysis without the need for manual intervention by data owners. Additionally, the integration of the Vault client within the COINSTAC ecosystem allows for seamless interaction between the desktop application and the Vaults, making it simple for consortium owners to include Vault data in their federated analyses.

2.2.5. Vault use-cases

In this section, we present various use-cases that highlight the benefits and versatility of Vaults in COINSTAC.

2.2.5.1. Curated Vaults

TReNDS actively curates and hosts public datasets, making them readily available for the COINSTAC community through the creation of Vaults. These curated Vaults ensure that the public datasets are vetted, of high quality, and easily accessible. Users can contribute to this initiative by hosting Vaults for other public datasets, further expanding the range of data resources available within COINSTAC.

2.2.5.2. User with local data

A researcher with a local dataset can benefit from incorporating Vault datasets containing relevant variables into their analysis. Integrating multiple datasets is especially advantageous when the researcher's local data is inadequate for conducting a comprehensive analysis. Collaborating with other COINSTAC consortium members and leveraging data from Vaults enables researchers to enhance the sample size and statistical power of their study efficiently while preserving privacy and streamlining the process by eliminating manual collaboration steps.

2.2.5.3. User with no local data

For investigators who do not have their own data but want to analyze existing datasets, Vaults provide a valuable solution. The investigator can create a consortium, add selected Vaults using the COINSTAC UI, and initiate the analysis. This approach enables the investigator to obtain meaningful insights from existing datasets without needing to coordinate with the Vault data owners.

2.2.5.4. User with limited storage/computing resources

Vaults are also advantageous for researchers with limited storage or computing resources. For example, a researcher with a low-powered laptop and minimal storage capacity can still analyze large datasets by creating a consortium and running an analysis using only Vault clients. The data processing occurs on the respective Vault servers, and the results are sent back to the investigator, eliminating the need for high-capacity local hardware.

By addressing these diverse use-cases, COINSTAC Vaults offer a flexible and efficient solution for researchers to access, collaborate, and analyze datasets in a federated environment.

3. Results

In this section, we conduct a series of analyses using multiple Vaults hosted by TRENDS, emphasizing the practical

application and utility of the Vaults feature. We specifically focus on the TReNDS VBM COBRE, TReNDS FreeSurfer COBRE, Child Mind Institute (CMI) VBM, and TReNDS NeuroMark Group-ICA COBRE datasets. These datasets were chosen to be hosted in Vaults based on their relevance to the neuroimaging research community, and their potential to demonstrate the diverse capabilities of COINSTAC Vaults. The hosting decisions were made in coordination with the respective data owners.

Our analyses highlight how the inclusion of Vault data can significantly increase sample size, thereby enhancing the statistical power of results. The diversity of datasets also underscores the flexibility and adaptability of COINSTAC Vaults, demonstrating how they can accommodate a wide range of research contexts and data types.

3.1. TReNDS VBM COBRE

The TReNDS VBM COBRE Vault contains structural MRI images from 152 participants, approximately half healthy volunteers and half individuals diagnosed with schizophrenia, collected as part of the Mind Research Network COBRE study (Aine et al., 2017). The Vault includes gray matter MRI images that have been run through a VBM preprocessing pipeline in the SPM toolbox. In addition, we have demographic information, symptom severity scales, and cognitive measures to select from when building a desired model. Figure 5 shows the beta images from running VBM regression on all the voxels from normalized smoothed gray matter images from the TReNDS COBRE Vault. Age, sex, and diagnosis information were used as covariates in the regression model. Results show decreases in brain volume with age, reduced volume in visual areas and along the gray/white boundary in females, and reduced volume in insular-temporal and medial frontal regions in schizophrenia patients, consistent with previous results.

The following section describes this use-case with 55 participant's structural MRI scans collected under MCIC project (Gollub, 2013). The results from running regression on the normalized smoothed gray matter images from this project are shown in Figure 6.

Using the MCIC dataset, we similarly see widespread reduction in brain volume for age, visual and gray/white boundary reductions in volume in females, and insular-temporal and medial frontal (as well as more wide spread) reductions in schizophrenia patients.

The TReNDS VBM COBRE Vault was combined with the MCIC dataset, allowing for an increased sample size, in the same regression analysis to examine diagnostic effects while accounting for age and sex. The combined dataset was largely consistent with the individual site analysis, with the exception of the male/female effect which shows a more complex pattern of increases and decreases, though still largely conforming to reductions in white/gray matter boundary and primary visual area volumes (Gupta et al., 2015). Results of this study are shown in Figure 7.

3.2. TReNDS FreeSurfer COBRE

This Vault contains data from 152 subjects, approximately half controls and half individuals with chronic schizophrenia, collected as part of the Mind Research Network COBRE study.[6] The Vault includes cortical and sub-cortical volumetric and surface-based measurements from two FreeSurfer atlases, Desikan-Killiany and Destrieux. In addition, we have a total of 11 variables across demographic, cognitive, and substance use to select from when building a desired model.

We ran Ridge regression on the above Vault data on Freesurfer volumetric and surface based measurements on about 500 regions of interest. We noticed the following differences between controls and patients.

Controls have higher values in temporal lobe, as shown in the thickness measurements of tables (Tables 1–5).

3.3. Child Mind Institute (CMI) VBM VAULT

This Vault contains data from 922 children and adolescents (ages 6–22, 603 Male and 319 female), collected as part of the Healthy Brain Network study (Alexander et al., 2017). The Vault includes gray matter segmentation data from an SPM VBM preprocessing pipeline. In addition, we have a total of 11 variables across various demographic, cognitive and substance use domains to select from when building a desired model.

Figure 8 shows the beta images from running regression on all the voxels from normalized smoothed gray matter images from the CMI VBM VAULT. Age and sex were used as covariates in the regression model. Results were largely consistent with those from the MCIC and COBRE analyses, showing widespread volume reductions with age, and reductions along the gray/white matter boundary in females.

3.4. TReNDS NeuroMark Group-ICA COBRE VAULT

Group ICA (Calhoun et al., 2001) is one of the frequently used preprocessing computations for neuroimaging data. Data preprocessed with group ICA can be used to perform different types of analyses. This GICA Vault comprises data from 189 subjects from the COBRE project analyzed with Neuromark template which uses 66 predefined ROIs. This Vault data includes independent component analysis (ICA) maps, Functional network connectivity maps (FNC) data etc. that have been generated using spatially constrained ICA with the `Neuromark_fMRI_1.0 template` (available in the GIFT software)[7,8] including 53 intrinsic networks (components). This Vault data can be readily used for secondary analysis like mancova. In this case, we use the GICA pre-processed data from the Vault to perform univariate regression analysis, the results of which are shown in Figure 9.

6 http://fcon_1000.projects.nitrc.org/indi/retro/cobre.html

7 http://trendscenter.org/software/gift

8 http://trendscenter.org/data

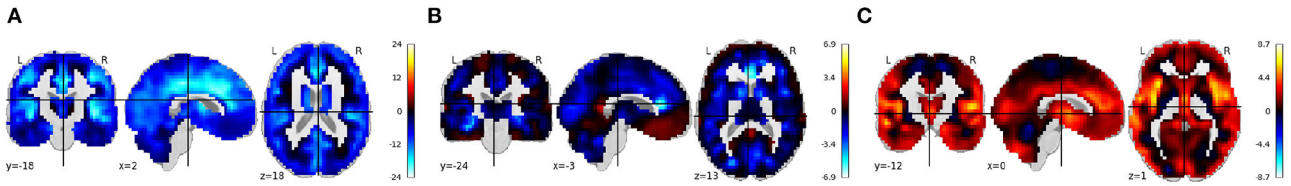


FIGURE 5
(A–C) Rendered images show voxel-wise β values corresponding to the *age*, *sex*, and *diagnosis* covariates using COBRE VBM data Vault in COINSTAC. For age, negative values show that the gray matter volume decreases with age. For sex, positive values indicate male's gray matter volume is greater than female's gray matter volume and vice versa. For diagnosis, positive values indicate control's gray matter volume is greater than patient's gray matter volume and vice versa.

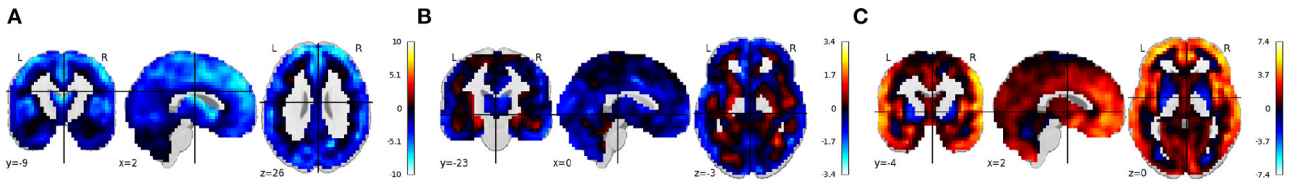


FIGURE 6
(A–C) Rendered images show voxel-wise β values corresponding to the *age*, *sex*, and *diagnosis* covariates using MCIC *s*MRI data in COINSTAC. For age, negative values show that the gray matter volume decreases with age. For sex, positive values indicate male's gray matter volume is greater than female's gray matter volume and vice versa. For diagnosis, positive values indicate control's gray matter volume is greater than patient's gray matter volume and vice versa.

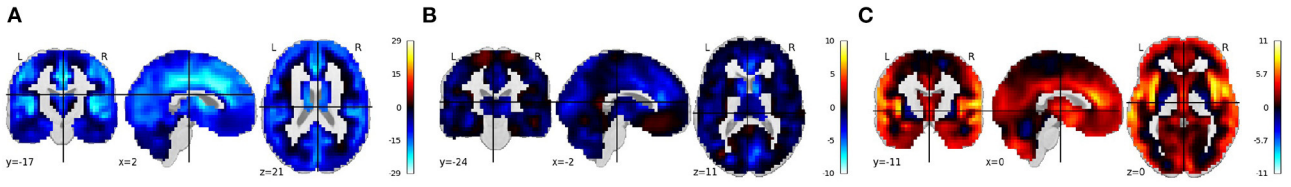


FIGURE 7
(A–C) MCIC+COBRE vault: rendered images show voxel-wise β values corresponding to the *age*, *sex*, and *diagnosis* covariates using MCIC *s*MRI data along with the data in the COBRE Vault in COINSTAC. For age, negative values show that the gray matter volume decreases with age. For sex, positive values indicate male's gray matter volume is greater than female's gray matter volume and vice versa. For diagnosis, positive values indicate control's gray matter volume is greater than patient's gray matter volume and vice versa.

TABLE 1 Global freesurfer stats for *lh_S_temporal_inf_thickness*.

Global stats – lh_S_temporal_inf_thickness	β0 (const)	β1 (age)	β2 (sex)	β3 (isControl_True)
Coefficient	2.5552	-0.0052	0.0323	0.1071
t stat	44.8963	-5.014	1.0642	4.1085
P-value	0	0	0.289	1.00E-04
R squared	0.237444911			
Degrees of freedom	145			

The Neuromark fMRI domains identified in Du et al. Briefly, these seven identified network templates were divided based on anatomical and functional properties (Du et al., 2020). In each subfigures, one color in the composite maps corresponds to an intrinsic connectivity network (ICN). The Neuromark_fMRI_1.0 template is available in the GIFT software (Figure 10).

4. Discussion

In recent decades, data sharing has driven substantial advancements in the field of neuroimaging and expanded opportunities for open science collaboration. Although data sharing has undeniable merits, it also faces inherent

TABLE 2 Global freesurfer stats for *rh_S_oc – temp_lat_thickness*.

Global Stats – rh_S_oc-temp_lat_thickness	β0 (const)	β1 (age)	β2 (sex)	β3 (isControl_True)
Coefficient	2.5331	−0.0036	0.0127	0.1158
t stat	38.8572	−3.0471	0.3663	3.878
P-value	0	0.0027	0.7147	2.00E-04
R squared	0.149884354			
Degrees of freedom	145			

TABLE 3 Global freesurfer stats for *lh_middletemporal_thickness*.

Global Stats – lh_middletemporal_thickness	β0 (const)	β1 (age)	β2 (sex)	β3 (isControl_True)
Coefficient	3.0038	−0.0057	−0.0134	0.0829
t stat	60.2257	−6.3161	−0.5056	3.63
P-value	0	0	0.6139	4.00E-04
R squared	0.275552216			
Degrees of freedom	145			

TABLE 4 Global freesurfer stats for *lh_superiortemporal_thickness*.

Global Stats – lh_superiortemporal_thickness	β0 (const)	β1 (age)	β2 (sex)	β3 (isControl_True)
Coefficient	2.9341	−0.0067	0.0169	0.0682
t stat	52.568	−6.6199	0.5679	2.6659
P-value	0	0	0.571	0.0085
R squared	0.266849477			
Degrees of freedom	145			

TABLE 5 Global freesurfer stats for *Left_Inf_Lat_Vent*.

Global Stats – Left_Inf_Lat_Vent	β0 (const)	β1 (age)	β2 (sex)	β3 (isControl_True)
Coefficient	428.1455	4.8982	−144.8783	−164.4922
t stat	5.1293	3.2264	−3.2585	−4.3021
P-value	0	0.0015	0.0014	0
R squared	0.22384763			
Degrees of freedom	145			

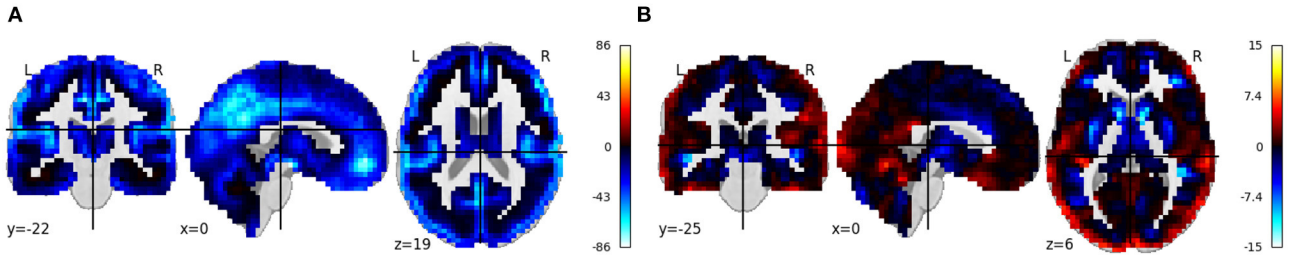


FIGURE 8
(A, B) Rendered images show voxel-wise β values corresponding to the *age* and *sex* covariates using CMI VBM Vault data in COINSTAC. For age, negative values show that the gray matter volume decreases with age. For sex, positive values indicate male's gray matter volume is greater than female's gray matter volume and vice versa.

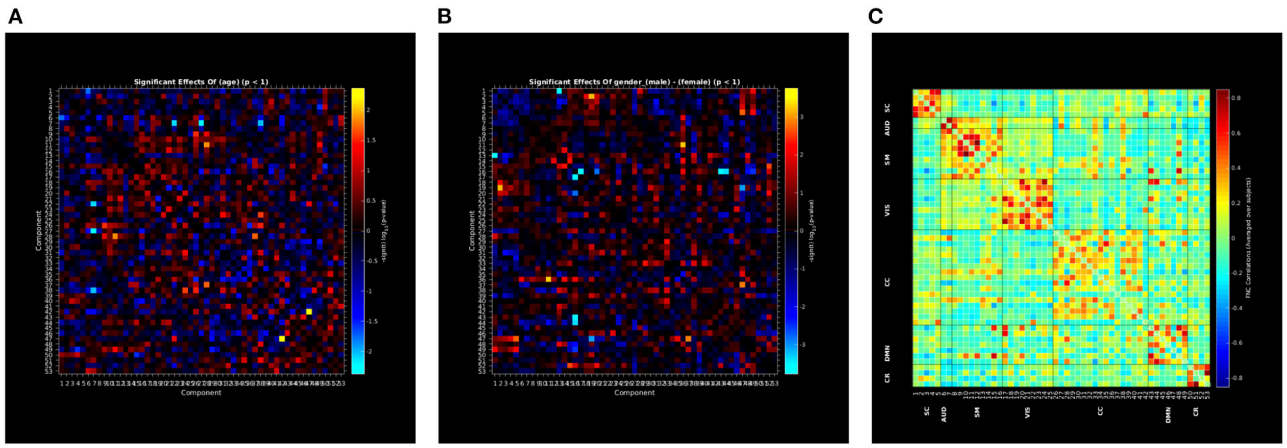


FIGURE 9
(A–C) Rendered images show univariate regression results demonstrating the effects of age and sex on correlation between the 53 independent components and FNC correlation map using vault data in COINSTAC.

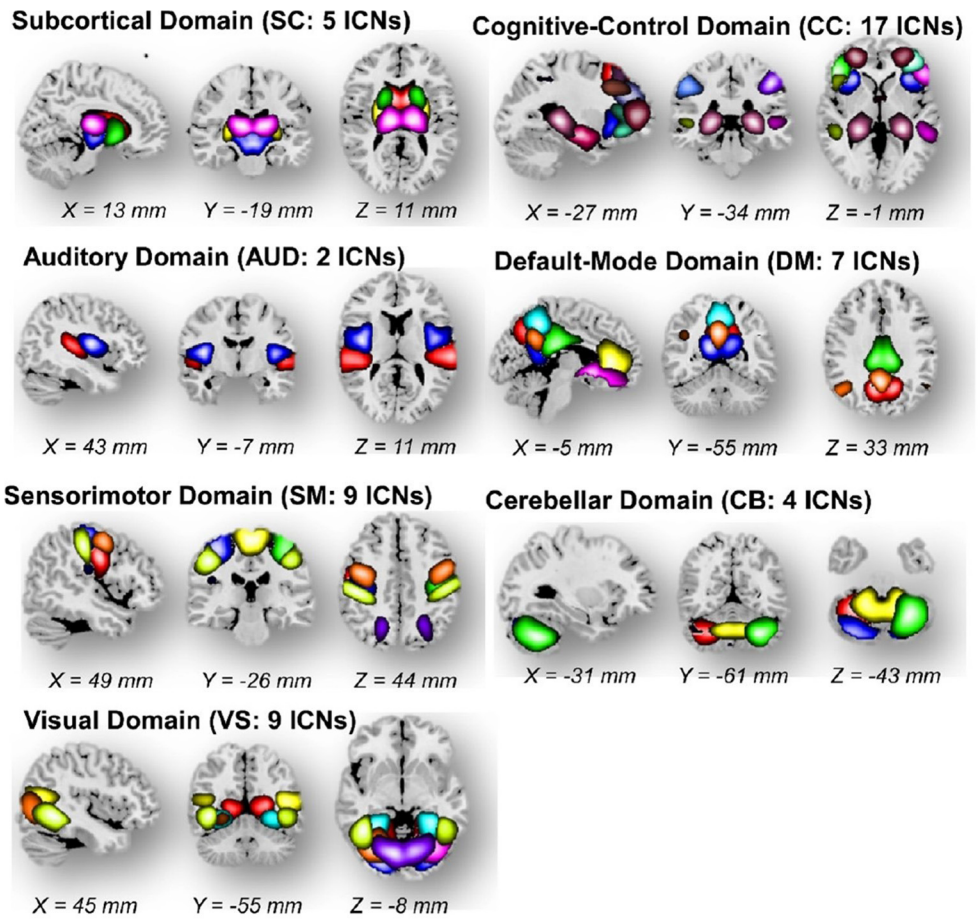


FIGURE 10
The Neuromark fMRI 1.0 template with 53 intrinsic networks (components) from 7 major networks.

limitations, including technological, policy, administrative, and methodological barriers that can hinder progress. COINSTAC Vaults and the federated computing framework within COINSTAC uniquely address these challenges by enabling data analysis while maintaining privacy protection, specifically in the context of neuroimaging research. The "always-on" status of Vaults streamlines collaboration between institutions by eliminating the need for synchronized efforts across users. The accessibility and user-friendly interface of COINSTAC Vaults serve as powerful tools for reproducible research, an area that has faced significant criticism in recent years. By bolstering the collaborative capabilities of federated learning and addressing the limitations of traditional data sharing, COINSTAC Vaults provide a cutting-edge solution for the neuroimaging community, pushing the boundaries of data analysis and open science.

COINSTAC offers a user-friendly GUI for the neuroimaging field, enabling federated learning on neuroimaging data with ease. Its extensive library includes numerous algorithms and pipelines, facilitating efficient processing of large datasets. Currently, over twenty computations are available in open-source repositories, allowing users to create versatile analytic pipelines. The integration of Vaults further enhances the user experience by providing access to diverse datasets, enabling efficient analysis with robust data, and fostering collaboration across institutions asynchronously.

Compared to OpenNeuro,[9] and OpenfMRI (Poldrack and Gorgolewski, 2017) like projects, where users can access data, download them and perform analysis on their own, Vaults allow users to perform neuroimaging analysis in federated learning platform immediately, without the need to download data and toolboxes onto a centralized computing environment. Vaults can help researchers to run an initial test on a data or their algorithm quickly to help setup their hypotheses or validate it to save time before they commit to a big project.

In addition to being faster to execute by being immediately available with no downloading or manual coordination, curated Vaults that follow documented standards make studies easier to design, execute, and reproduce. For example: Neuroimaging datasets can contain a large number of variables that apply to each subject: demographic information, cognitive measures, etc. The number of these variables can range from tens to hundreds. Using standard naming conventions makes it easier for researchers to understand what each variable tracks so that they can select the relevant variables for their study. Standard and predictable ways for handling missing data in Vaults makes it easier for researchers to design their analyses.

COINSTAC is unique in its commitment to open science, with its open-source platform promoting seamless integration of modular computations and streamlining federated analyses. The addition of COINSTAC Vaults reinforces this commitment by simplifying dataset inclusion in federated analyses, encouraging community contributions, and preserving privacy for private datasets. By offering easy access to public datasets and enabling secure contributions from private dataset owners, COINSTAC Vaults foster collaboration and dedication to open science.

9 https://openneuro.org/

4.1. Limitations and challenges

COINSTAC Vaults offer numerous benefits, but there are also limitations and challenges to consider, particularly in the areas of data privacy and security, and resource usage.

One concern is that allowing arbitrary summary queries on a dataset might enable an attacker to reconstruct the data. To mitigate such risks, the system must be privacy-preserving from "end-to-end," incorporating techniques like secure multiparty computation or differential privacy. Implementing these methods can be difficult due to floating point implementation issues (Mironov, 2012; Ilvento, 2020a,b) and the introduction of noise, which may increase error or variance in the analysis results.

While differentially private algorithms can provide stronger privacy guarantees, sharing data derivatives without differential privacy might be adequate in some situations, depending on the trust model and privacy concerns of data holders. These issues should be addressed on a case-by-case basis.

Vault owners can currently restrict computations on their data to a pre-approved list. To enhance privacy protection, further improvements are recommended. Potential solutions include allowing Vault owners to:

- Approve or deny individual analysis runs.
- Specify users and consortia that are allowed to run analyses.
- Limit the overall number of computation runs for a vault.
- Set expiration dates for specific approval permissions.

Another challenge is handling slowdowns or crashes during resource-intensive analyses due to high compute usage. To address this issue, Vault owners can be given more control over resource usage and compute capacity. They could limit the number of concurrent computations and overall CPU usage. Improving compute capacity could involve strategies like deploying multiple instances behind a load balancer or dynamically scaling resources.

Additional challenges include data distribution, network bandwidth, and communication speed. Federated learning and open-source solutions can help address some of these problems, but further research and development are needed to optimize COINSTAC Vaults' performance in various research settings. Our "Decentralized Sparse Deep Artificial Neural Networks in COINSTAC (CPU and GPU enabled)" algorithm allows users to save network bandwidth when transferring thousands of derived data/machine learning parameters across nodes.

In summary, COINSTAC Vaults mark a significant advancement in federated neuroimaging research, data privacy preservation, and open science promotion. By tackling the existing limitations and challenges, COINSTAC Vaults can further improve collaboration and innovation within the field.

5. Conclusion

The neuroimaging field is experiencing rapid growth, generating substantial data volumes. However, access to this data is challenged by technological, privacy, administrative, and methodological constraints. In this study, we present COINSTAC Vaults as a solution that streamlines data access and analysis,

specifically in the context of neuroimaging research. COINSTAC Vaults ensure continuous availability of high-quality data, promoting the advancement of open science and fostering efficient collaboration between researchers.

We invite researchers to use COINSTAC Vaults in their studies and to host their own datasets using COINSTAC Vaults. By adopting COINSTAC Vaults, the neuroimaging community can overcome the barriers associated with traditional data sharing and analysis methods, paving the way for groundbreaking discoveries.

5.1. Future work

The long-term vision for COINSTAC and COINSTAC Vaults includes:

- Introducing new user interface features, such as the ability to search Vaults and filter by covariates, to improve user experience and efficiency.
- Making new datasets available as Vaults, including those from OpenNeuro, the Autism Brain Imaging Data Exchange (ABIDE), the National Institute of Mental Health Data Archive (NDA), the Open Access Series of Imaging Studies (OASIS), and the Image and Data Archive (IDA), to enhance the diversity of Vaults.
- Increase BIDS (Brain Imaging Data Structure) support to all major neuroimaging modalities and Vault datasets, to ensure interoperability and ease of use.
- Increase compliance to programs such as the FAIR (Findability, Accessibility, Interoperability, and Reuse) Guiding Principles for scientific data management and stewardship, to enhance the overall data sharing ecosystem.
- Exploring the integration of differential privacy techniques to further safeguard data privacy, while preserving the utility of data analysis.

Data availability statement

Publicly available datasets were analyzed in this study. This data can be found here: http://fcon_1000.projects.nitrc.org/indi/cmi_healthy_brain_network/sharing_neuro.html, http://fcon_1000.projects.nitrc.org/indi/retro/cobre.html.

Author contributions

DM, RK, and VC: conceptualization. DM, SB, SPa, and PP: methodology. DM, SB, SPa, KR-M, PP, BB, and JR: writing—original draft preparation. SB and SPa: data analysis. SPl and VC: supervision. All authors: writing—review and editing, read, and agreed to the published version of the manuscript.

Funding

This work was funded by the National Institutes of Health (Grants: R01DA040487, R01DA049238, and R01MH121246).

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

References

Aine, C. J., Bockholt, H. J., Bustillo, J. R., Cañive, J. M., Caprihan, A., Gasparovic, C., et al. (2017). Multimodal neuroimaging in schizophrenia: description and dissemination. *Neuroinformatics* 15, 343–364. doi: 10.1007/s12021-017-9338-9

Alexander, L. M., Escalera, J., Ai, L., Andreotti, C., Febre, K., Mangone, A., et al. (2017). An open resource for transdiagnostic research in pediatric mental health and learning disorders. *Sci. Data* 4, 1–26. doi: 10.1038/sdata.2017.181

Andrade, C. (2020). Sample size and its importance in research. *Indian J. Psychol. Med.* 42, 102–103. doi: 10.4103/IJPSYM.IJPSYM_504_19

Babayan, A., Baczkowski, B., Cozatland, R., Dreyer, M., Engen, H., Erbey, M., et al. (2022). *MPI-Leipzig Mind-Brain-Body Dataset*.

Biswal, B. B., Mennes, M., Zuo, X.-N., and Milham, M. P. (2010). Toward discovery science of human brain function. *Proc. Natl. Acad. Sci. U.S.A.* 107, 4734–4739. doi: 10.1073/pnas.0911855107

Bonawitz, K., Ivanov, V., Kreuter, B., Marcedone, A., McMahan, H. B., Patel, S., et al. (2016). *Practical Secure Aggregation for Federated Learning on User-Held Data*. Technical Report. doi: 10.48550/arXiv.1611.04482

Bonawitz, K., Ivanov, V., Kreuter, B., Marcedone, A., McMahan, H. B., Patel, S., et al. (2017). "Practical secure aggregation for privacy-preserving machine learning," in *Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security, CCS '17* (New York, NY: ACM), 1175–1191.

Calhoun, V. D., Adali, T., Pearlson, G., and Pekar, J. (2001). "Group ICA of functional MRI data: separability, stationarity, and inference," in *Proceeedings of the International Conference on ICA and BSS* (San Diego, CA), 155.

Du, Y., Fu, Z., Sui, J., Gao, S., Xing, Y., Lin, D., et al. (2020). NeuroMark: an automated and adaptive ICA based pipeline to identify reproducible fMRI markers of brain disorders. *Neuroimage Clin.* 28:102375. doi: 10.1016/j.nicl.2020.102375

Dwork, C., and Roth, A. (2013). The algorithmic foundations of differential privacy. *Found. Trends Theoret. Comput. Sci.* 9, 211–407. doi: 10.1561/0400000042

Esteban, O., Markiewicz, C. J., Blair, R. W., Moodie, C. A., Isik, A. I., Erramuzpe, A., et al. (2019). fMRIPrep: a robust preprocessing pipeline for functional MRI. *Nat. Methods* 16, 111–116. doi: 10.1038/s41592-018-0235-4

Gazula, H., Kelly, R., Romero, J., Verner, E., Baker, B. T., Silva, R. F., et al. (2020). COINSTAC: Collaborative informatics and neuroimaging suite toolkit for anonymous computation. *J. Open Source Softw.* 5, 2166. doi: 10.21105/joss.02166

Gazula, H., Rootes-Murdy, K., Holla, B., Basodi, S., Zhang, Z., Verner, E., et al. (2023). Federated analysis in COINSTAC reveals functional network connectivity and spectral links to smoking and alcohol consumption in nearly 2,000 adolescent brains. *Neuroinformatics* 21, 287–301. doi: 10.1007/s12021-022-09604-4

Gollub, R. L. (2013). The MCIC collection: a shared repository of multi-modal, multi-site brain image data from a clinical investigation of schizophrenia. *Neuroinformatics* 11, 367–388. doi: 10.1007/s12021-013-9184-3

Gorgolewski, K. J., Auer, T., Calhoun, V. D., Craddock, R. C., Das, S., Duff, E. P., et al. (2016). The brain imaging data structure, a format for organizing and describing outputs of neuroimaging experiments. *Sci. Data* 3, 1–9. doi: 10.1038/sdata.2016.44

Gupta, C. N., Calhoun, V. D., Rachakonda, S., Chen, J., Patel, V., Liu, J., et al. (2015). Patterns of gray matter abnormalities in schizophrenia based on an international mega-analysis. *Schizophr. Bull.* 41, 1133–1142. doi: 10.1093/schbul/sbu177

Heikkilä, M. A., Koskela, A., Shimizu, K., Kaski, S., and Honkela, A. (2020). Differentially private cross-silo federated learning. *arXiv preprint arXiv: 2007.05553*. doi: 10.48550/arXiv.2007.05553

Homer, N., Szelinger, S., Redman, M., Duggan, D., Tembe, W., Muehling, J., et al. (2008). Resolving individuals contributing trace amounts of dna to highly complex mixtures using high-density SNP genotyping microarrays. *PLoS Genet.* 4, e1000167. doi: 10.1371/journal.pgen.1000167

Ilvento, C. (2020a). "Implementing differentially private integer partitions," in *Presented at the 2020 Workshop on the Theory and Practice of Differential Privacy*.

Ilvento, C. (2020b). "Implementing sparse vector," in *Presented at the 2020 Workshop on the Theory and Practice of Differential Privacy*.

Imtiaz, H., Mohammadi, J., Silva, R., Baker, B., Plis, S. M., Sarwate, A. D., et al. (2021). A correlated noise-assisted decentralized differentially private estimation protocol, and its application to fMRI source separation. *IEEE Trans. Signal Process.* 69, 6355–6370. doi: 10.1109/TSP.2021.3126546

Jwa, A. S., and Poldrack, R. A. (2022). The spectrum of data sharing policies in neuroimaging data repositories. *Hum. Brain Mapp.* 43, 2707–2721. doi: 10.1002/hbm.25803

Kairouz, P., McMahan, H. B., Avent, B., Bellet, A., Bennis, M., Bhagoji, A. N., et al. (2021). Advances and open problems in federated learning. *Found. Trends Mach. Learn.* 14, 1–210. doi: 10.1561/2200000083

Markiewicz, C. J., Gorgolewski, K. J., Feingold, F., Blair, R., Halchenko, Y. O., Miller, E., et al. (2021). The OpenNeuro resource for sharing of neuroscience data. *eLife* 10, e71774. doi: 10.7554/eLife.71774.sa2

McGuire, A. L., Basford, M., Dressler, L. G., Fullerton, S. M., Koenig, B. A., Li, R., et al. (2011). Ethical and practical challenges of sharing data from genome-wide association studies: the emerge consortium experience. *Genome Res.* 21, 1001–1007. doi: 10.1101/gr.120329.111

Ming, J., Verner, E., Sarwate, A., Kelly, R., Reed, C., Kahleck, T., et al. (2017). COINSTAC: decentralizing the future of brain imaging analysis. *F1000Research* 6, 1512. doi: 10.12688/f1000research.12353.1

Mironov, I. (2012). "On significance of the least significant bits for differential privacy," in *Proceedings of the 2012 ACM Conference on Computer and Communications Security (CCS)* (Raleigh, NC), 650–661.

Plis, S. M., Sarwate, A. D., Wood, D., Dieringer, C., Landis, D., Reed, C., et al. (2016). COINSTAC: a privacy enabled model and prototype for leveraging and processing decentralized brain imaging data. *Front. Neurosci.* 10, 365. doi: 10.3389/fnins.2016.00365

Poldrack, A. R. and Gorgolewski, K. J. (2017). OpenfMRI: open sharing of task fMRI data. *Neuroimage* 144(Pt B), 259–261. doi: 10.1016/j.neuroimage.2015.05.073

Rootes-Murdy, K., Gazula, H., Verner, E., Kelly, R., DeRamus, T., Plis, S., et al. (2022). Federated analysis of neuroimaging data: a review of the field. *Neuroinformatics* 20, 377–390. doi: 10.1007/s12021-021-09550-7

Senanayake, N., Podschwadt, R., Takabi, D., Calhoun, V., and Plis, S. (2022). NeuroCrypt: machine learning over encrypted distributed neuroimaging data. *Neuroinformatics* 20, 91–108. doi: 10.1007/s12021-021-09525-8

Thompson, P. M., Andreassen, O. A., Arias-Vasquez, A., Bearden, C. E., Boedhoe, P. S., Brouwer, R. M., et al. (2017). ENIGMA and the individual: predicting factors that affect the brain in 35 countries worldwide. *Neuroimage* 145, 389–408. doi: 10.1016/j.neuroimage.2015.11.057

Thompson, P. M., Stein, J. L., Medland, S. E., Hibar, D. P., Vasquez, A. A., Renteria, M. E., et al. (2014). The ENIGMA consortium: large-scale collaborative analyses of neuroimaging and genetic data. *Brain Imaging Behav.* 8, 153–182. doi: 10.1007/s11682-013-9269-5

Turner, J. A., Calhoun, V. D., Thompson, P. M., Jahanshad, N., Ching, C. R., Thomopoulos, S. I., et al. (2022). ENIGMA + COINSTAC: improving findability, accessibility, interoperability, and re-usability. *Neuroinformatics* 20, 261–275. doi: 10.1007/s12021-021-09559-y

Vogt, N. (2023). Reproducibility in MRI. *Nat. Methods* 20, 34. doi: 10.1038/s41592-022-01737-3

frontiers | Frontiers in Neuroinformatics

TYPE Methods
PUBLISHED 18 July 2023
DOI 10.3389/fninf.2023.1174156

OPEN ACCESS

EDITED BY
Maaike M. H. Van Swieten,
Integral Cancer Center Netherlands (IKNL), Netherlands

REVIEWED BY
B. Nolan Nolan Nichols,
Maze Therapeutics, United States
Alexandre Rosa Franco,
Nathan Kline Institute for Psychiatric Research, United States

*CORRESPONDENCE
Nazek Queder
✉ nqueder@uci.edu

RECEIVED 25 February 2023
ACCEPTED 27 June 2023
PUBLISHED 18 July 2023

CITATION
Queder N, Tien VB, Abraham SA, Urchs SGW, Helmer KG, Chaplin D, van Erp TGM, Kennedy DN, Poline J-B, Grethe JS, Ghosh SS and Keator DB (2023) NIDM-Terms: community-based terminology management for improved neuroimaging dataset descriptions and query.
Front. Neuroinform. 17:1174156.
doi: 10.3389/fninf.2023.1174156

NIDM-Terms: community-based terminology management for improved neuroimaging dataset descriptions and query

Nazek Queder[1,2]*, Vivian B. Tien[3], Sanu Ann Abraham[4], Sebastian Georg Wenzel Urchs[5], Karl G. Helmer[6,7], Derek Chaplin[6], Theo G. M. van Erp[8,9], David N. Kennedy[10], Jean-Baptiste Poline[5], Jeffrey S. Grethe[11], Satrajit S. Ghosh[12] and David B. Keator[1]

[1]Department of Psychiatry and Human Behavior, School of Medicine, University of California, Irvine, Irvine, CA, United States, [2]Department of Neurobiology and Behavior and Center for the Neurobiology of Learning and Memory, University of California, Irvine, Irvine, CA, United States, [3]Fairmont Preparatory Academy, Anaheim, CA, United States, [4]McGovern Institute for Brain Research, Massachusetts Institute of Technology, Cambridge, MA, United States, [5]NeuroDataScience–ORIGAMI Laboratory, McConnell Brain Imaging Centre, The Neuro (Montreal Neurological Institute-Hospital), Faculty of Medicine, McGill University, Montreal, QC, Canada, [6]Massachusetts General Hospital, Boston, MA, United States, [7]Harvard Medical School, Boston, MA, United States, [8]Clinical Translational Neuroscience Laboratory, Department of Psychiatry and Human Behavior, School of Medicine, University of California, Irvine, Irvine, CA, United States, [9]Center for the Neurobiology of Learning and Memory, University of California, Irvine, Irvine, CA, United States, [10]Departments of Psychiatry and Radiology, University of Massachusetts Chan Medical School, Worcester, MA, United States, [11]Department of Neurosciences, School of Medicine, University of California, San Diego, San Diego, CA, United States, [12]McGovern Institute for Brain Research, Massachusetts Institute of Technology, Cambridge, MA, United States

The biomedical research community is motivated to share and reuse data from studies and projects by funding agencies and publishers. Effectively combining and reusing neuroimaging data from publicly available datasets, requires the capability to query across datasets in order to identify cohorts that match both neuroimaging and clinical/behavioral data criteria. Critical barriers to operationalizing such queries include, in part, the broad use of undefined study variables with limited or no annotations that make it difficult to understand the data available without significant interaction with the original authors. Using the Brain Imaging Data Structure (BIDS) to organize neuroimaging data has made querying across studies for specific image types possible at scale. However, in BIDS, beyond file naming and tightly controlled imaging directory structures, there are very few constraints on ancillary variable naming/meaning or experiment-specific metadata. In this work, we present NIDM-Terms, a set of user-friendly terminology management tools and associated software to better manage individual lab terminologies and help with annotating BIDS datasets. Using these tools to annotate BIDS data with a Neuroimaging Data Model (NIDM) semantic web representation, enables queries across datasets to identify cohorts with specific neuroimaging and clinical/behavioral measurements. This manuscript describes the overall informatics structures and demonstrates the use of tools to annotate BIDS datasets to perform integrated cross-cohort queries.

KEYWORDS

neuroimaging, dataset, query, annotation, NIDM

1. Introduction

There is a "crisis of replication" in neuroscience (Button et al., 2013; Szucs and Ioannidis, 2017). Interpreting, reproducing, and validating results of experiments depends critically on our ability to understand the conditions under which the data were acquired and processed. Efficient discovery and reuse of existing data relies on the data and metadata adhering to the FAIR: Findable, Accessible, Interoperable and Reusable principles (Wilkinson et al., 2016; Schulz, 2018). The biomedical research community is motivated to share and reuse data from studies and projects by an increasing number of requirements from funding agencies (e.g., NIH-wide Policy for Data Management and Sharing[1]) and publishers (PMID: 34914921). There are a growing number of data repositories (Das et al., 2012; Book et al., 2013; Poldrack et al., 2013; Ambite et al., 2015; Crawford et al., 2016; Kennedy et al., 2016), each with their own data structures and data dictionaries (Eickhoff et al., 2016). With dozens of neuroimaging data sharing sources now available, we need better methods to annotate datasets and to search across those datasets without a significant investment in time to develop database mediation services (Keator et al., 2008; Turner et al., 2015; Wang et al., 2016; Niso et al., 2022) or creating "crosswalks" mapping variables across datasets.

Critical barriers to finding and reusing data include the use of undefined variables and/or an insufficient degree of variable annotations that make it difficult to understand the data available without significant interaction with the original authors. Further, determining whether cohorts from different studies can be combined, based on phenotypes or acquisition parameters is currently difficult, requiring a significant investment in effort from the researcher. The ability to conduct searches across diverse datasets is difficult and typically requires sufficient annotation of the study variables to understand what was collected and how to query each dataset to find meaningful results. For example, a query such as: "identify datasets that contain a measure of depression, age, IQ, and a T1-weighted MRI scan" is not easy to implement. Historically, this type of query would have to be posed to multiple data repositories separately, through each repository's interface, and the results manually combined by the investigator. Often, the returned results would depend upon the annotations used in each repository and the level of granularity to which each data object was annotated which may require the investigator to download complete datasets in order to manually extract the data of interest. In some cases, this query can not be satisfied without an expert user because often the same annotation term collection is not used across repositories, as terms used to annotate collected study variables are inconsistent. Each lab can freely name study variables such that they are not guaranteed to be meaningful or sufficient, either for understanding or for querying each interface and each dataset.

Building off the example query above, it has proven difficult to query arbitrary datasets to find out whether they contain images with contrast types relevant to the research question. The Brain Imaging Data Structure (BIDS) (Gorgolewski et al., 2015; incf-nidash, 2016) was designed to provide software developers and the neuroimaging community with file- and directory-naming conventions for organizing imaging data. Because of its simplicity, BIDS has been quickly supported by a number of analysis tools and database platforms (COINS,[2] XNAT,[3] Scientific Transparency,[4] OpenfMRI,[5] LORIS[6]). In BIDS, the organization of the data is required to conform to strict naming and directory-structure conventions. The adoption of BIDS has addressed the imaging-related parts in our example query above because with BIDS and the associated PyBIDS[7] Python library, one can use the location of data within the directory structure to determine the type of images included in that dataset. Beyond file naming and tightly controlled imaging directory structures, there are very few constraints on ancillary variable naming/meaning or experiment-specific metadata in BIDS. As such, we still have difficulty satisfying the query above because: (1) we cannot guarantee that variable names will be meaningful; (2) data dictionaries are optional in BIDS and there is no validation that data dictionaries, if supplied, contain important or sufficient information (e.g., units, frames of reference, etc.). Therefore, searching and combining information across independent BIDS datasets is often difficult for data beyond image types and metrics describing those images. Finally, there is no query engine that natively supports BIDS datasets.

To address these concerns regarding the ability to query across BIDS datasets, as well as the desire to create a web of linked human neuroimaging data, an international team of cognitive scientists, computer scientists, and statisticians are developing a (meta)data representation model and tools to support its use. The goal is to provide the foundational infrastructure in a well-defined and easily expandable model, to link datasets using unambiguous annotations. This effort, built upon the resource description framework (RDF) and the PROV standard[8] (Moreau et al., 2008; PROV-Overview, 2016), is called the Neuroimaging Data Model (NIDM)[9] (Keator et al., 2013; Maumet et al., 2016; NIDM, 2016). By using RDF as the foundation for NIDM, it benefits from a variety of sophisticated query languages (e.g., SPARQL, RQL, TRIPLE, Xcerpt), an open world assumption allowing users to add as many statements about the data as they like without constraints on header sizes as is the case with typical image formats, and direct use of web-accessible terminologies and ontologies to provide multiple layers to link and infer relationships among data and metadata. A full description of NIDM is beyond the scope of this manuscript, but NIDM was designed to facilitate queries across neuroscientific datasets. A Python library was built (PyNIDM[10]) to create NIDM annotation documents and a tool was also created to represent a BIDS dataset, along with all the associated behavioral and/or clinical data, as a NIDM document. Using PyNIDM and NIDM documents, one could use RDF query languages to satisfy the example query above across BIDS or other datasets.

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

2 https://coins.trendscenter.org/

3 https://www.xnat.org/

4 https://scitran.github.io/

5 https://openfmri.org/

6 http://mcin-cnim.ca/neuroimagingtechnologies/loris/

7 https://github.com/bids-standard/pybids

8 https://www.w3.org/TR/prov-overview/

9 http://nidm.nidash.org/

10 https://github.com/incf-nidash/PyNIDM

To annotate datasets for future discovery or integration, researchers need to be able to rely on a set of common properties for precisely defining study variables, beyond what is already offered by BIDS for imaging data. In other domains beyond neuroimaging, tools have been developed to aid in dataset annotation such as the open source ISA framework (Sansone et al., 2012) for life sciences research, the Clinical Data Interchange Standards Consortium (CDISC) RDF framework (Facile et al., 2022) focused on the medical and healthcare domains, and Frictionless Data[11] developed to support climate scientists, to humanities researchers, to government data centers, and others. In this manuscript, we focus on the research neuroimaging community yet many of the methods presented are general and could be applied to other domains in synergy with related efforts. Here we describe NIDM-Terms, a toolkit that employs both the NIDM data model and associated terminologies to aid in querying across datasets. We provide tools to more fully annotate BIDS datasets and provide user-friendly community-based annotation and terminology management tools to assure proper definitions and metadata are provided with the annotations. Further, we show how these annotations, along with the NIDM data model, can be used to search across publicly available neuroimaging datasets.

2. Materials and methods

In the following sections, we begin by formalizing our definitions of different data element types. We then define a small set of properties we consider critical to include when annotating study-specific data elements to be able to both understand, at a high level, what was collected and assure such annotations have the necessary information for researchers to understand how to reuse and/or combine these with other studies. Finally, we describe some tools, both command-line and graphical, for creating these data element annotations.

2.1. Data element types

Data elements can be simply defined as annotations on data, where data can be variable names or content, file names or content of files. In this work, we introduce two distinct types of data elements and a conceptual element: (1) data elements that are often locally defined and represent study variables (personal data elements; PDEs), (2) data elements that are defined by a community or a standards body (common data elements; CDEs), and (3) terms that capture an abstract idea or a general notion (concepts). Within the NIDM terminology work, each of these distinct types of elements play an important role in the detailed description of datasets.

Personal Data Elements (PDEs) refer to the typical study variables and require strict definitions, ranges, value types, and, if categorical, complete definitions of the categories and their potential mapping from numerical categories to text-based strings (e.g., 0 = right handed, 1 = left handed, 2 = ambidextrous) to be easily reused across studies. PDEs may define common terms in non-standard ways or use non-standard terms for commonly acquired variables or processing steps. PDEs may also combine separate annotation terms into a single term, e.g., "age_months" that combines the duration "age" with the units of "months." In general, since PDEs are used locally, there is no requirement to adhere to a standard convention and users are typically free to name and annotate such elements as they wish.

Common data elements (CDEs) are those that have been adopted for use by a group, often either a consortium operating in a specific domain or standards body. Ideally, a rigorous adoption process is implemented that entails the proposal of a term, identification of whether similar terms already exist in other terminologies, determination of how the term will fit into the logical structure of the existing terminology and whether it adheres to standards already established by the group. Often though, CDE collections may be simply that, a collection of terms that a group has decided to use, without the establishment of any standards or logical framework.

Concepts are distinctly different from CDEs and PDEs. Concepts (also known as "classes" in RDF) are those terms that represent "higher order" ideas, e.g., the concept of "age" is the notion of a duration of time from some predetermined starting point to the current moment. Concepts are used to aid in querying across datasets and provide a mechanism for researchers to annotate their study-specific PDEs (or CDEs if they are used within a study) with abstract ideas or general notions about a PDE which helps us to query across datasets. For example, two studies collect a data element meant to measure the participant's dominant hand. Dataset one names the variable simply "handedness" and is stored as a categorical variable with values indicating whether the participant is predominantly right-handed, left-handed, or ambidextrous. Dataset two instead collects the Edinburgh handedness inventory, names their study variable "ehi" and whose values are integers ranging from −40 (left handed) to 40 (right handed). Therefore, *no query for a single variable name would return data from both datasets*. However, if each dataset annotated their handedness assessment data with a concept describing the general notion of "handedness assessment," for example term *ILX:0104886*[12] in the InterLex repository, querying across datasets would then return handedness data from both datasets. One could then investigate each returned dataset and understand, through the data element properties (see section "2.2. Properties"), the distinctions between how each was measured.

The use of properly defined CDEs, concepts, and properties provides the foundation for NIDM documents to: (1) abstract the concepts inherent in PDEs to allow for meaningful searches across data collections, (2) provide an extensible collection of general and domain-specific terms used to describe data, and (3) allow for an inherently flexible annotation of data to an arbitrary level of detail. As an example of the above points, reconceptualizing the PDEs "age_months" and "YEARSOLD" from different datasets with the properties "isAbout" the concept "age" and "hasUnits" of "months" and "years," respectively, allows an automated system to discover both PDEs when "age" is searched for, as well as not having to define a separate variable each time a different duration unit is required.

11 https://frictionlessdata.io

12 http://uri.interlex.org/ilx_0104886

TABLE 1 Data element properties.

Property	Definition
Description	An explanation of the nature, scope, or meaning of the data element.
Label	Short text string for referring to the data element.
ValueType	A value representation such as integer, float, string, date/time (e.g., xsd: int, xsd: float, xsd: string).
UnitCode	Unit of measurement (e.g., years, millimeters, etc.).
MaxValue	The upper value of the data element (in case of ordered data).
MinValue	The lower value of the data element (in case ordered data).
Choices	Choices is a concept that corresponds to the BIDS (https://bids.neuroimaging.io/) "levels" standard for categorical variables where you're mapping the value (often an integer) to some text string. Using the handedness example from above, the choices would be {1 = Right, 5 = Left, 10 = Ambidextrous}.
IsAbout	Used to record the relationship between a data element and a broader concept. Annotating using is About can be used to search across datasets. The is About annotations consist of a url to identify the concept and an optional label for the concept.
Source_variable	Variable name from dataset. This applies to personal data elements which are data elements defined within a specific study, typically referred to as "study variables."
MeasureOf	Describes what the data element measures (e.g., volume, area, distance, intensity, health status, duration/period, intelligence).
datumType	What type of datum it is (e.g., range, count, scalar etc.).
IsPartOf	Used to link data elements to assessments (e.g., WAIS_Vocab_Raw linked to WAIS scale (https://www.cognitiveatlas.org/task/id/tsk_4a57abb949f12/#). Typically this is not added by the user and is often done as an additional annotation to link data elements with other classes of information.
SubtypeCDEs	This property is typically added during curation. It links the term to lower-level (child) terms to provide some limited ontological relationships.
SupertypeCDEs	This property is typically added during term curation. It links the term to higher-level (parent) terms to provide some limited ontological relationships.
AssociatedWith	List of strings used to associate data elements with communities (e.g., BIDS, NIDM, etc.) for grouping data elements or searching within communities for specific data elements.

2.2. Properties

Properties play an important role in disambiguating and simplifying the annotation of data, as well as the mapping of data elements between data sources, especially those that use PDEs. In reviewing available terminologies and ontologies for use in human neuroimaging studies, we found that data elements in these terminologies often lacked important properties such as units, value types, ranges, etc. When researchers try to reuse data collected by other laboratories they often request data dictionaries which describe the study variables collected (PDEs) and hopefully provide precise definitions and properties for those variables. If important properties are missing, such as "units," the data is either not usable or users must contact the dataset providers, if they are reachable, to correctly and confidently reuse the data. The NIDM team found this to be a significant problem when trying to reuse retrospective data and query across studies to build cohorts matching various search criteria. We therefore started out by defining a minimal set of properties (**Table 1**) that we felt were important to properly define data elements of various types (see section "2.1. Data element types"). A larger set of properties are available in the terminology used in the experimental description component of NIDM (i.e., NIDM-Experiment).

Many of the terms that are used for annotation are part of NIDM-Experiment (NIDM-E), an ontology that can be used to describe neuroscience experiments. NIDM-E was originally constructed with an emphasis on terms describing imaging-based studies, in particular those employing MRI, but has since been expanded to encompass other modalities. NIDM-E was built through the annotation of several real-world multi-modality neuroscience-based data sets. The goal of NIDM-E is to provide semantically-aware tools, a collection of defined terms organized in a structure that can be used to annotate data to an arbitrary level of detail. The structure of NIDM-E allows a user both to annotate complicated data collections and accommodate terms for new modalities and acquisition methods. NIDM-E also comprises tools to discover terms, webpages for term URL resolution, and a framework for community conversations regarding the terms.

As per good ontological practice (Arp et al., 2015), NIDM-E reuses terms from other ontologies before creating new terms. Terms are reused from such active ontologies such as the Semanticscience Integrated Ontology (SIO) (Dumontier et al., 2014), Information Artifact Ontology (IAO) (Ceusters, 2012), and Prov-O.[13] These general ontologies provide the framework to which domain-specific terms were added to create NIDM-E. Terms created for NIDM-E have formal Aristotelian definitions in the "X is a Y that Z" format (Seppälä et al., 2017). NIDM-E also includes many imported data type, object, and annotation properties.

Because NIDM-E began with neuroimaging data, it has particularly strong coverage in that domain. It contains two unique properties: "hadImageContrastType" and "hadImageUsageType" that can be used to distinguish between the physics-based mechanism for the contrast in an image (e.g., "T2-weighted") and the eventual use of that image (e.g., "Anatomical"). These are important for the discovery of imaging data in and across

13 https://www.w3.org/TR/prov-o

10.3389/fninf.2023.1174156

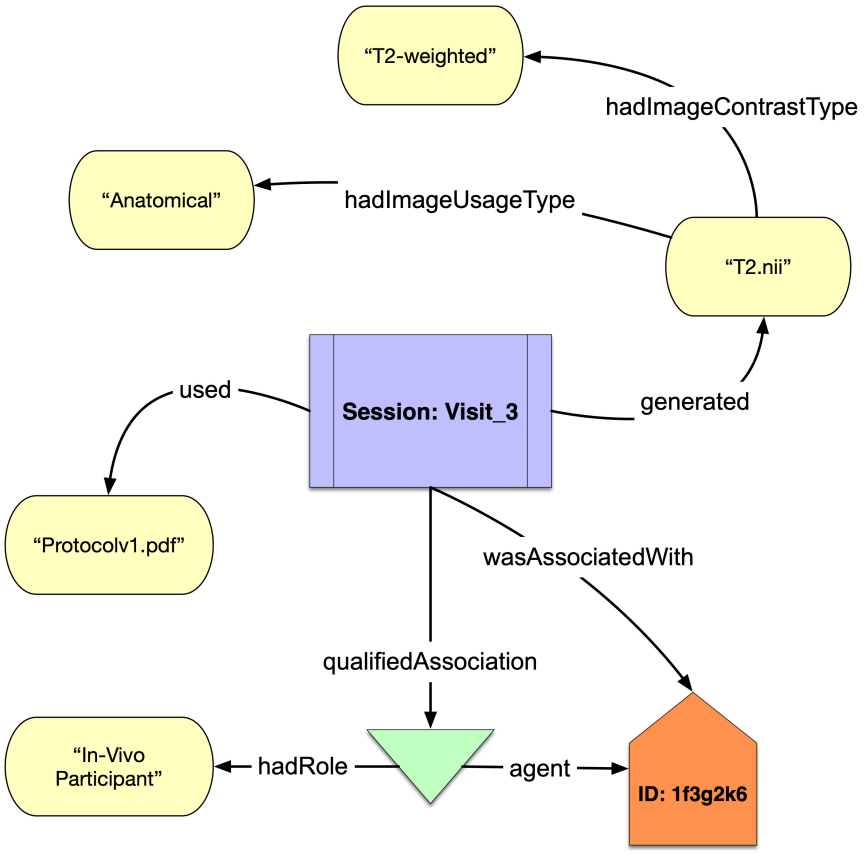


FIGURE 1
A schematic of a acquisition object "T2.nii" that was generated from the Session "Visit_3" of participant "1f3g2k6" annotated with the protocol used and "hadImageUsageType" and "hadImageContrastType."

repositories, where datasets with different image contrasts may be annotated by the same term. For example, T2*-weighted, and T2-weighted images may both be stored as "Functional" data. NIDM-E also includes terms from two widely used standards: DICOM[14] and the BIDS standards, the former of which is ubiquitous in the neuroimaging domain for the formatting of raw image data and is used in multiple imaging modalities. We have created a set of DICOM tag data type properties that can be used to associate acquisition parameters with an acquisition object. We have also included BIDS terms so that BIDS-organized datasets can be annotated using terms directly from the official BIDS schema.

We show in Figure 1 a simple example of how NIDM-E can annotate an acquisition object, "T2.nii," with an image contrast type of "T2-weighted" and an image usage type of "Anatomical," and showing the scan session activity it was acquired at ("Session:Visit_3"), the protocol that was used ("Protocolv1.pdf"), and the study participant ("ID:1f3g2k6") from which it was acquired and who had the role of "*In Vivo* Participant."

The NIDM-E term-resolution and schema pages are available in GitHub[15] which includes a web-accessible infrastructure built so that (1) the neuroscientific community can suggest or edit terms to the NIDM-E vocabulary using GitHub issue templates, (2) terms have resolvable URI's, and (3) the ontology can be browsed to facilitate term discovery. GitHub issue templates allow us to have a public record of the discussion surrounding each term. To discover NIDM-E terms, we provide a "Schema Browser"[16] webpage that allows users to view the NIDM-E term graph, including all of the imported terms. For semantic-web applications, we also provide a "Terms Resolution"[17] page in which each term has a unique URL so that terms used by applications have a unique reference location.

2.3. SHACL validation

Beyond just defining useful properties for annotating data elements, it is critically important that researchers include such properties in their data annotations (i.e., data dictionaries). To ensure that data elements annotated by the community and contributed to the NIDM-Terms ecosystem contain the appropriate properties according to their type, we have built a validation schema using the Shapes Constraint Language (SHACL) (Pareti et al., 2019). SHACL is a W3C-supported language for validating RDF graphs according to a schema (i.e., a SHACL shape). Each data element type (e.g., PDE, CDE, and concept)

14 https://www.dicomstandard.org/
15 https://github.com/incf-nidash/nidm-experiment
16 https://incf-nidash.github.io/nidm-experiment/schema_menu.html
17 https://incf-nidash.github.io/nidm-experiment/

has a separate SHACL shape used for validation. These shapes specify the required properties, the value type of each property's values, and the number of such properties in each data element definition. Validation is done when new data elements are added to the NIDM-Terms GitHub repository through pull requests, either using the NIDM-Terms UI (see section "3.1. NIDM-Terms user interface") or through Github Actions and Github Pull Requests. The git action uses the Python validation framework provided by ReproSchema.[18] In brief, Reproschema offers a way to standardize the underlying representation of assessment tools. It comes with an open and accessible library of questionnaires with appropriate conversion [e.g., from/to RedCap (Patridge and Bardyn, 2018)] and data collection tools [e.g., MindLogger (Klein et al., 2021), RedCap, etc.] to enable a more consistent acquisition across projects, with data being harmonized by design. The techniques described here have been aligned with ReproSchemas to both support automated annotation of data collected and shared using assessments from ReproSchemas and to align our data element descriptions so there is consistency across representations. Such consistency will help the user who wants to share their study data when collected using ReproSchemas.

2.4. Terminology management resources used in NIDM terms

In order to provide users with the ability to easily annotate their data and link selected PDEs to broader concepts, a simple means to query across existing terminologies is needed. These query services are provided by Interlex[19] (Surles-Zeigler et al., 2021), a dynamic lexicon, initially built on the foundation of NeuroLex (PMID: 24009581), of biomedical terms and common data elements designed to help improve the way that biomedical scientists communicate about their data, so that information systems can find data more easily and provide more powerful means of integrating data across distributed resources. One of the challenges for data integration and FAIR data is the inconsistent use of terminology and data elements. InterLex allows for the association of data fields and data values to common data elements and terminologies, enabling the crowdsourcing of data-terminology mappings within and across communities. InterLex also provides a stable layer on top of the many other existing terminologies, lexicons, ontologies (i.e., provides a way to federate ontologies for data applications), and common data element collections to enable more efficient search for users. To support annotation using CDEs, InterLex has been expanded to include the full NIMH Data Archive (NDA) CDE library. Through available RESTful web-services, InterLex is supporting alignment of data elements and terminologies through PyNIDM developed to simplify creation, editing, and querying of NIDM documents. To further expand our available terminologies, PyNIDM supports querying the Cognitive Atlas[20] as an additional information source for dataset annotation. Similar to Interlex, Cognitive Atlas provides a systematic approach to representing cognitive neuroscience entities and biomedical terminologies.

3. Results

In previous sections, we have described the foundational principles used in this work to annotate study variables. Research laboratories often reuse PDEs across research projects or, alternatively, define new PDEs for studies that have previously been used in other projects. In an effort to help labs maintain an internal list of PDEs and share them with others in the community, we have developed both terminology management and dataset annotation tools. In the following sections, we describe three such annotation tools and a terminology management interface. We then show how proper dataset annotations can be useful in querying across publicly available MRI-related neuroimaging data.

3.1. NIDM-Terms user interface

To facilitate the community's interaction in managing the neuroimaging terminology, we developed a JavaScript (using Visual Code Studio: version 1.67.1) NIDM-Terms User Interface[21] (UI), hosted on GitHub Pages, that allows community curators to define and interact with their lab-specific terminologies as well as reuse terms from other neuroimaging communities. The UI is designed around the Git version control system and uses the NIDM-Terms/terms[22] GitHub repository as a backend, providing JavaScript Object Notation - Linked Data (JSON-LD)[23] formatted files for each PDE, CDE, and concept contributed by the community.

The NIDM-Terms UI provides the following supportive functions: browse, search, edit, and export available terms and their properties. The "Browse Terms" function (**Figure 2**, Panel A) fetches the NIDM-Terms GitHub repository and displays the JSON-LD formatted files in a treeview format, including a tag for the term's data type (e.g., concepts and data elements). Users are then able to filter through the available communities and terms based on the label of the term they're interested in. We have developed additional functionality that allow users to suggest edits to the available terms and their properties, across the neuroimaging communities hosted on the UI. The UI will create a JSON-LD formatted dictionary with the user's suggested edits to a specific term and using the edits as a query parameter string. Upon submission, a new browser tab will open a new Github pull request, with the edited term and its properties, to the NIDM-Terms repository allowing the user to use their login information to complete the pull request. The "Suggest new terms" function (**Figure 2**, Panel B) works in a similar manner to edit terms. Suggested terms will be formatted as a JSON-LD file. The JSON-LD file is then stringified and sent as a query parameter to the pull request to the NIDM-Terms repository; in case of any technical difficulties, the UI will submit a github issue to the NIDM-Terms repository with the suggested term describing the problem specifics while submitting the pull request. Upon the term's approval by a community's curator, a JSON-LD representation of the new

18 https://github.com/ReproNim/reproschema-py
19 https://scicrunch.org/scicrunch/interlex/dashboard
20 http://www.cognitiveatlas.org
21 https://nidm-terms.github.io/
22 https://github.com/NIDM-Terms/terms
23 https://json-ld.org/

10.3389/fninf.2023.1174156

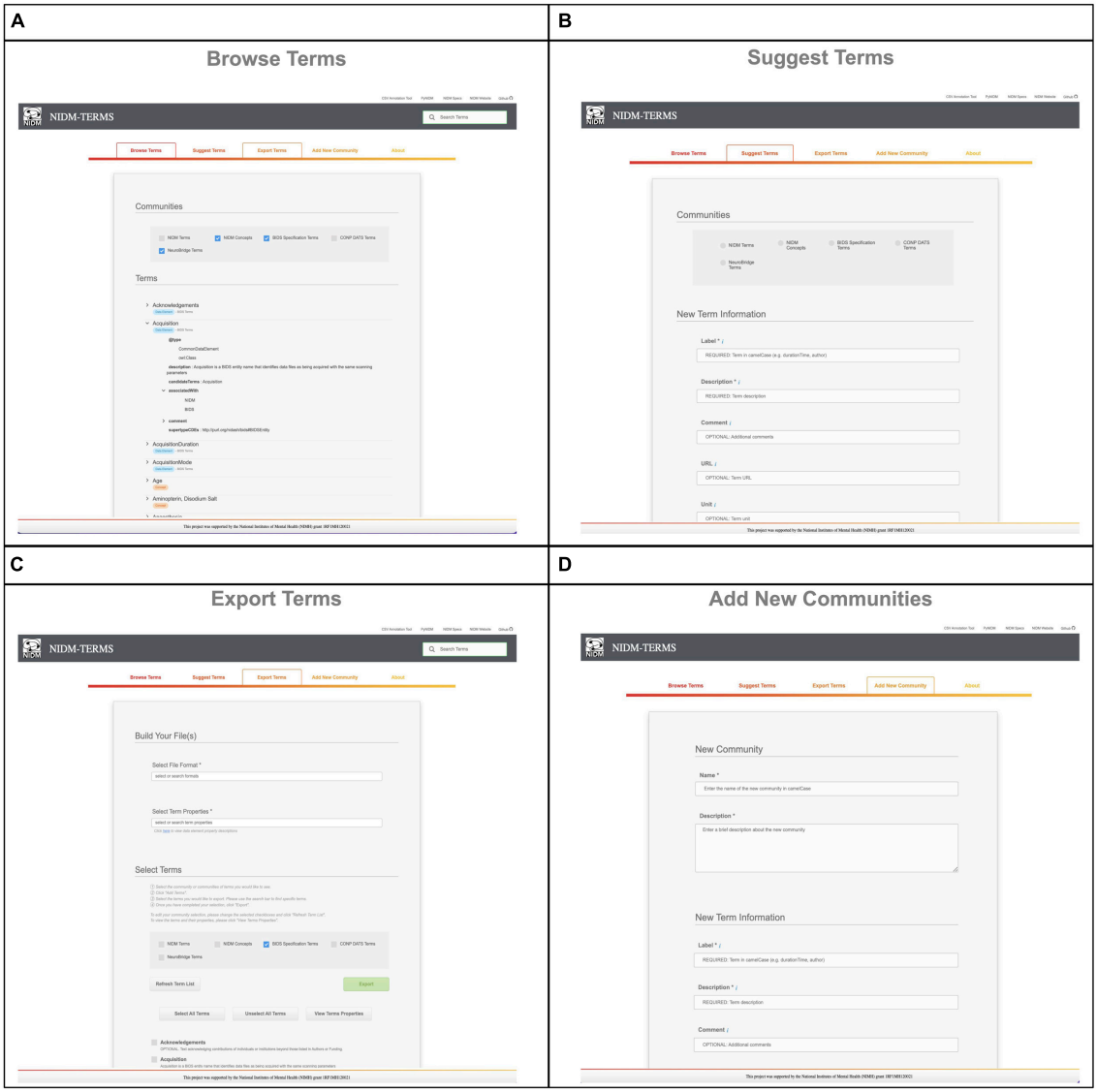


FIGURE 2
This figure illustrates the various functionalities the NIDM-Terms User Interface (UI) supports including: browse terms **(A)**, suggest terms **(B)**, export terms **(C)**, and add new communities **(D)**.

terms will be added to the repository and a tree-view display of the new term will appear under the "Browse Terms" section of the UI. Note, each community has its own curators responsible for approving/interacting with users suggesting new terms and/or editing terms. In this way, each community has responsibility for their own terms. The "Export selected terms" function allows users to export terms, across communities, along with their properties, in several file formats: (1) A Markdown table for possible inclusion in community's documentation (e.g., BIDS reference manual); (2) JSON; (3) JSON-LD; (4) CSV; (5) N-Quads (Figure 2, Panel C). Finally, the "Add a new community" function which allows for the addition of a new community to NIDM-Terms. Similar to "Suggest Terms," the "Add a new community" functionality submits a pull request with the new community as a query parameter. Upon submitting new terms and communities to the NIDM-Terms GitHub repository, all new terms are validated using our SHACL Schema Validator (see section "2.3. SHACL validation"), consistent with the term properties described in (section "2.2. Properties").

To further enhance our list of neuroimaging communities and support communities who may want complete control over their repository, we have provided instructions for cloning the NIDM-Terms UI and associated GitHub repository in order to host their own community in their name-space prior to merging them with the NIDM-Terms repository for broader community use. Together, these tools form a user-friendly interface allowing the neuroimaging community curators to interact with and reuse terminologies across communities, backed by a version control system.

3.2. Dataset annotation tools

In this project, we have created several tools to assist the neuroimaging community in annotating datasets using the terminology management tools we developed (see section "3.1. NIDM-Terms user interface"), consistent with our data element types and properties. This includes defining study-specific variables and their properties, as well as linking the variables to higher level concepts using the properties in Table 1. A rich set of annotations increases a dataset's Findability and Reusability and can make publicly available datasets more FAIR by enabling scientists to efficiently discover datasets using concept-based queries.

To achieve this goal, we have built several annotation tools that allow scientists to efficiently and effectively annotate their study variables. We have built both command-line and graphical annotation tools. First, the "bidsmri2nidm" tool enables scientists to annotate BIDS structured datasets by iterating over the dataset and its variables contained in the "participants.tsv" file or other phenotypic files stored in the "phenotypes" directory through a command-line interface. A series of questions about each study variable will then be displayed on the screen allowing users to input specific properties describing those variables such as description, unit, minimum value, maximum value, etc. Additionally, the tool queries concepts from information sources such as InterLex and Cognitive Atlas for users to select the best matching concept to their study variable. The tool suggests concepts that are fuzzy-matched to the study variable name and provides a mechanism for users to refine such queries. Often when searching large information sources such as InterLex for concepts, users might find multiple concepts that could be applicable to the variable to be annotated. In our annotation tools, we initially present to the user the term deemed closest to the study variable amongst the list of concepts used for prior data annotations. For example, if annotating a study variable that stores the age of the participant, each data set provider should annotate this variable with the same "age" concept to increase consistent term usage. Our tools attempt to restrict the space of concepts by re-using concepts already used in data annotations from other users. In this way, we reduce the space to a single concept for "age." The user can always broaden their search for concepts but this initial reduction in the search space helps to steer the user in selecting a concept that increases the potential for finding these data across studies. This reduction in search space is accomplished by the tool searching the NIDM-Terms github repository which maintains a list of concepts selected for annotations by users of the tool. To prevent duplicate choices for common study variables often used in queries (e.g., age, sex, handedness) the list of prior concepts is currently being manually curated. This is a place ripe for development using AI natural language processing techniques to keep the list of concepts relatively small and consistent.

After the annotation process is completed, the tool will export a JSON dictionary with the variables and their properties in addition to a NIDM-Experiment RDF document. This tool is a great addition to the neuroimaging community because it allows scientists to more easily add detailed and standard annotations to their BIDS structured datasets. In addition to "bidsmri2nidm," we have also developed "csv2nidm," which also allows for annotation of study variables however it uses tabular data [e.g., comma-separated values files (CSV) or tab–separated values files (TSV)] instead of requiring a complete BIDS datasets. Both of our command line interface tools, "bidsmri2nidm" and "csv2nidm" are open source and available with the PyNIDM (see text footnote 10) tools. To expand the use of our tools, we have additionally built a user-friendly web-based Graphical user interface version of csv2nidm[24].

A second web-based annotation tool that has been developed by the ReproNim[25] community is the Neurobagel[26] annotation tool. This graphical annotation tool loads a tabular phenotypic file - for example a BIDS participants.tsv file–and then guides the user through two annotation stages. In the first stage (Figure 3), the user is presented with a number of pre-configured categories (i.e., CDEs), which have been previously agreed on across a number of dataset providers, and is asked to identify the columns of the loaded phenotypic file that contain information about each category (e.g., sex, or clinical diagnosis). To accomplish this step in the user interface (UI), the user first selects a category by clicking on the corresponding colored button, and then clicks on each column from the phenotypic file that she wants to associate with the category. An existing association between a column and a category is represented in the UI by highlighting the column name with the respective category color. In the second stage of the annotation process (Figure 4), the user is asked to annotate the values in each column that has been associated with a category (continuous values can be transformed into a standardized format). Each category has a predefined list of terms from a controlled vocabulary that a user has to choose from to annotate the values in their phenotypic file (from a list of common data elements). Constraining the annotation terms is a design choice to make the annotation process easier and to facilitate consistency across annotations at the expense of flexibility. However, the predefined categories will be configurable in the next version of the annotation tool to help communities choose the most appropriate set of terminologies. After completing the annotation, the neurobagel annotator creates a BIDS compatible data dictionary (JSON) file, that contains the additional semantic annotations as additional properties and can be converted to a NIDM file.

These annotation tools are beneficial for the neuroimaging community because they allow users to quickly and accurately annotate their study variables in a standardized way. This helps ensure that their data is consistently structured and more easily understood by other researchers working on similar projects and to facilitate cross-dataset queries. By integrating Interlex and Cognitive Atlas, it also allows scientists to quickly and easily match their variables to existing concepts, making it easier to formulate sophisticated scientific queries and to interpret their results.

3.3. Use case

To evaluate the developed tools and overall terminology management, annotation, and query workflows presented in this manuscript section, we focus on a specific use-case, that of querying across publicly available MRI-related neuroimaging data

24 https://incf-nidash.github.io/nidmterms-ui/?#/annotate
25 https://www.repronim.org/
26 https://annotate.neurobagel.org

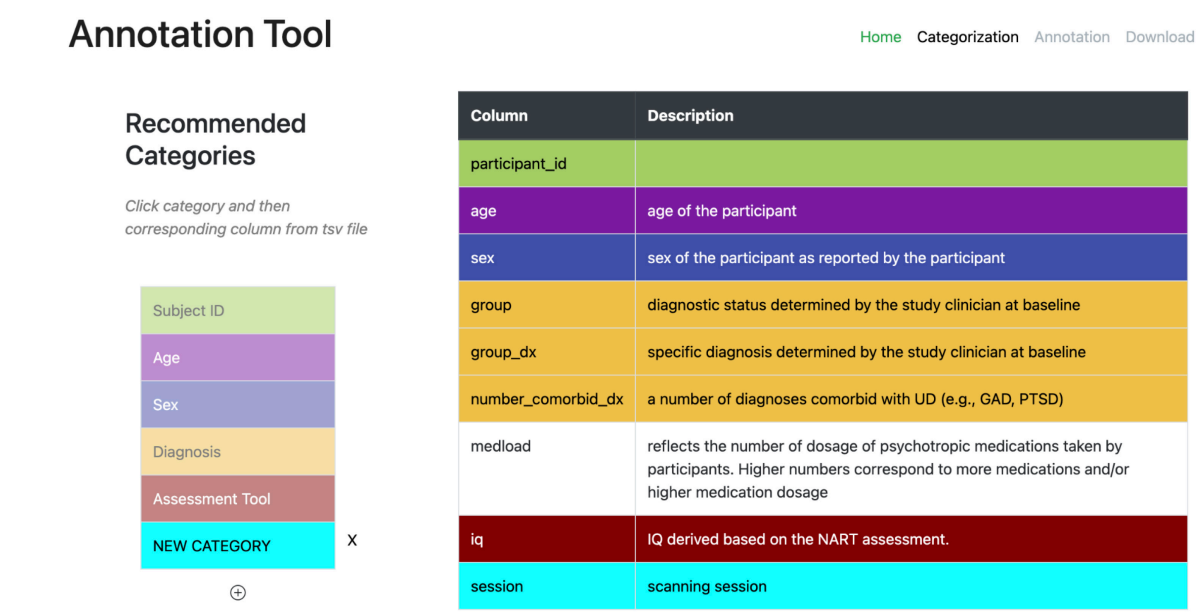


FIGURE 3
Stage 1 of the neurobagel annotator workflow. **Left:** a list of pre-defined categories (common data elements) is shown, each associated with a specific color. **Right:** the column names of a loaded demographic.tsv file with optional descriptions are displayed. The user now selects each category by clicking on it (e.g., "Diagnosis" in yellow), and then associates the category with each column that contains information about this category by clicking on it (e.g., "group," "group_dx," and "number_comorbid_dx"). An existing association is reflected by the column being highlighted in the color of the category.

to identify potential cohorts of interest. For these tests we use publicly available projects contained in the OpenNeuro archive at the time our work began, the ABIDE[27] dataset, and the ADHD200[28] dataset, all of which are available from each dataset provider and were accessed using DataLad[29] (Halchenko et al., 2021; **Figure 5**). Each of these datasets and the projects within the OpenNeuro archive are available in the BIDS format and generally contain MRI imaging data along with selected demographics and additional cognitive and/or behavioral assessments at varying levels of complexity. In addition, the selected datasets contain differing amounts of annotations. For ABIDE and ADHD200 datasets, full data dictionaries are available from the dataset providers; although, not in a readily parsable format (e.g., PDF format). For OpenNeuro projects, approximately 25% had annotations in the form of BIDS "sidecar" JSON files and the rest did not.

To prepare these reference datasets for query, given their varying levels of existing annotations and organizational form (e.g., BIDS containing phenotype data, BIDS for imaging data and phenotype data stored as separate tabular data files outside of BIDS), we used various NIDM-related tools. For the ABIDE study, each study site created their own BIDS dataset containing the imaging data. When we started our work, the phenotype data was stored separately for all sites as a CSV file. In later versions of the BIDS datasets, phenotype data was stored in the BIDS "participants.tsv" files. Although each study site collected the same phenotypic variables, it was often the case that the variable names were slightly different across sites in terms of the spelling, capitalization, and word-connection indicators such as spaces, dashes, or underscores. This inconsistency, even within a single study, demonstrates the difficulties users may have in trying to query across datasets simply using variable names. Further, there were no BIDS "sidecar" files included with any of the site's BIDS datasets. To convert the ABIDE BIDS datasets into a NIDM document for query we used the following procedure:

- Download each ABIDE site's BIDS dataset via Datalad.
- Manually convert the PDF-formatted data dictionary into a NIDM JSON-formatted data dictionary.
  - Add entries to JSON-formatted data dictionaries to accommodate all heterogeneity in variable naming across ABIDE sites.
  - Add high-level concept associations to the JSON-formatted data dictionary for selected variables using the isAbout property.
- For each ABIDE site.
  - Run PyNIDM tool "bidsmri2nidm" with a local path to the BIDS dataset.
  - Run PyNIDM tool "csv2nidm" with a local path to the phenotype CSV file, the JSON-formatted data dictionary, and the NIDM file created by the "bidsmri2nidm" step above.

The procedure above results in one NIDM document per site, containing both the imaging and phenotype metadata, along with the data dictionaries and concept annotations. In this procedure we created the JSON-formatted data dictionary and did concept

27 http://fcon_1000.projects.nitrc.org/indi/abide/abide_I.html
28 http://fcon_1000.projects.nitrc.org/indi/adhd200/
29 https://www.datalad.org/

10.3389/fninf.2023.1174156

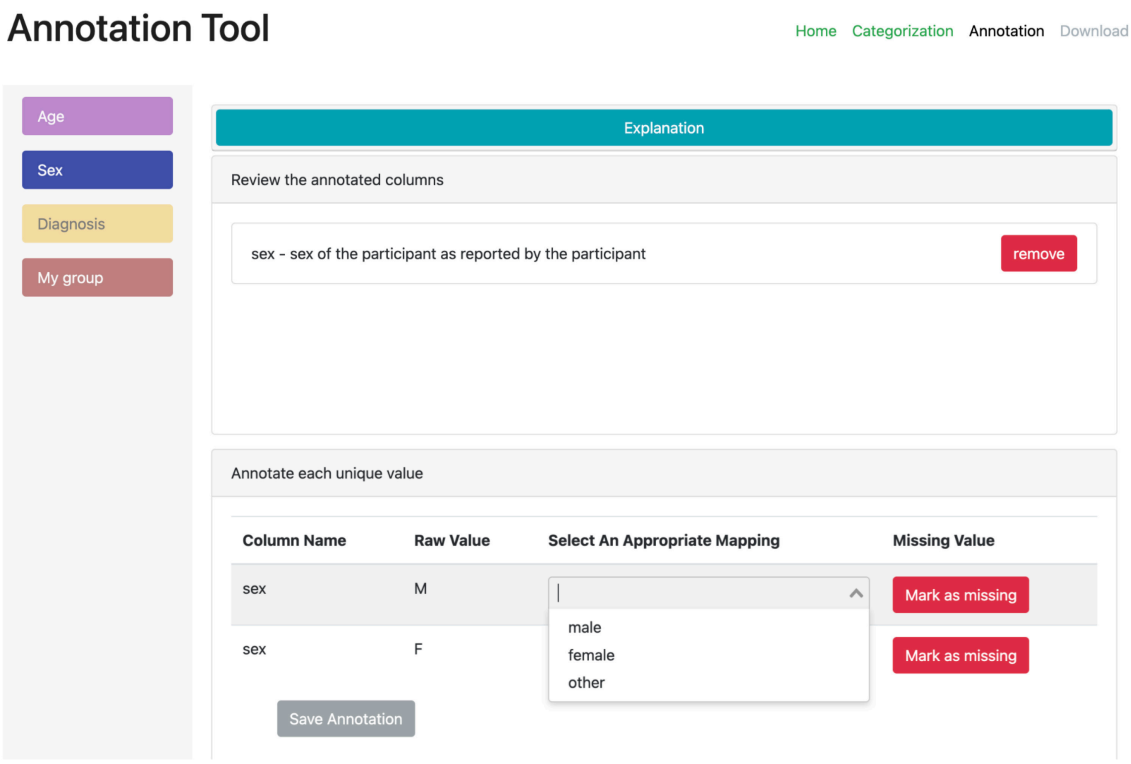


FIGURE 4
Stage 2 of the neurobagel annotator workflow. The user is now asked to annotate the values inside each column that has been linked to one of the predefined categories. **Left:** each category associated with at least one column is represented with a colored button. By clicking on each button, the user can annotate the values in the associated columns. **Right:** the annotation view for each category (here "Sex" in blue) contains specific elements such as an explanation (collapsed here), an overview of the associated column names (here "sex"), and an overview of the unique values in the associated column **(bottom)**. The user can map each unique value to a pre-defined list of controlled terms (here with a drop-down menu) or indicate that the value reflects a "missing value" (e.g., a data entry error or a truly missing response).

associations manually since we already had a reference PDF document with variable definitions. Alternatively, one could use any of the annotation tools discussed in (section "3.2. Dataset annotation tools").

To prepare the ADHD200 dataset for query we follow a similar procedure as for the ABIDE dataset except that the BIDS formatted dataset contained the imaging and phenotypic data. Similar to ABIDE, each study site's BIDS data was stored separately and there was a variety of variable name heterogeneity. Here again, to improve the efficiency and account for the heterogeneity of variable names, we manually created the data dictionaries by transcribing the information from PDF-formatted documents. Similar to the ABIDE dataset, one could have used our annotation tools (from section "3.2. Dataset annotation tools") as an alternative approach, which we think is far easier and less prone to transcription errors. Yet to be able to capture the heterogeneity of variable names across all the sites, we would have had to run 'bidsmri2nidm' many times, once for each site, answering all the annotation questions about variables that have already been annotated but have slightly different names (e.g., one as a space whereas another site used an underscore). To complete this task at the scale we were working at, it was simpler for us to create a single data dictionary with all the variable name variations and provide this to the "bidsmri2nidm" tool. At the end of this procedure, we have a NIDM document per site containing both the imaging and phenotype data along with the data dictionaries and concept annotations.

Finally, we created NIDM documents for each dataset available in the OpenNeuro archive via DataLad at the time we performed these experiments. For the datasets in the OpenNeuro archive we used the following procedure:

- Download each OpenNeuro BIDS dataset via Datalad.
- Evaluate whether a data dictionary (JSON sidecar file) is available and export all variable names and properties to a Google spreadsheet along with project name and contact emails.
    - If data dictionary is present.
        - Add concept annotations to the spreadsheet manually.
    - If data dictionary is not present.
        - Evaluate variables for consistency with BIDS schema recommended data type, units, etc. (e.g., age variable suggested to be years, etc.).
        - Ask dataset providers for clarity when needed.
        - Add concept annotations to spreadsheet manually.
- Convert Google spreadsheet entries to BIDS JSON sidecar files for each project using our additions.
- Run PyNIDM tool 'bidsmri2nidm' with a local path to the BIDS dataset and using our BIDS JSON "sidecar" files.

10.3389/fninf.2023.1174156

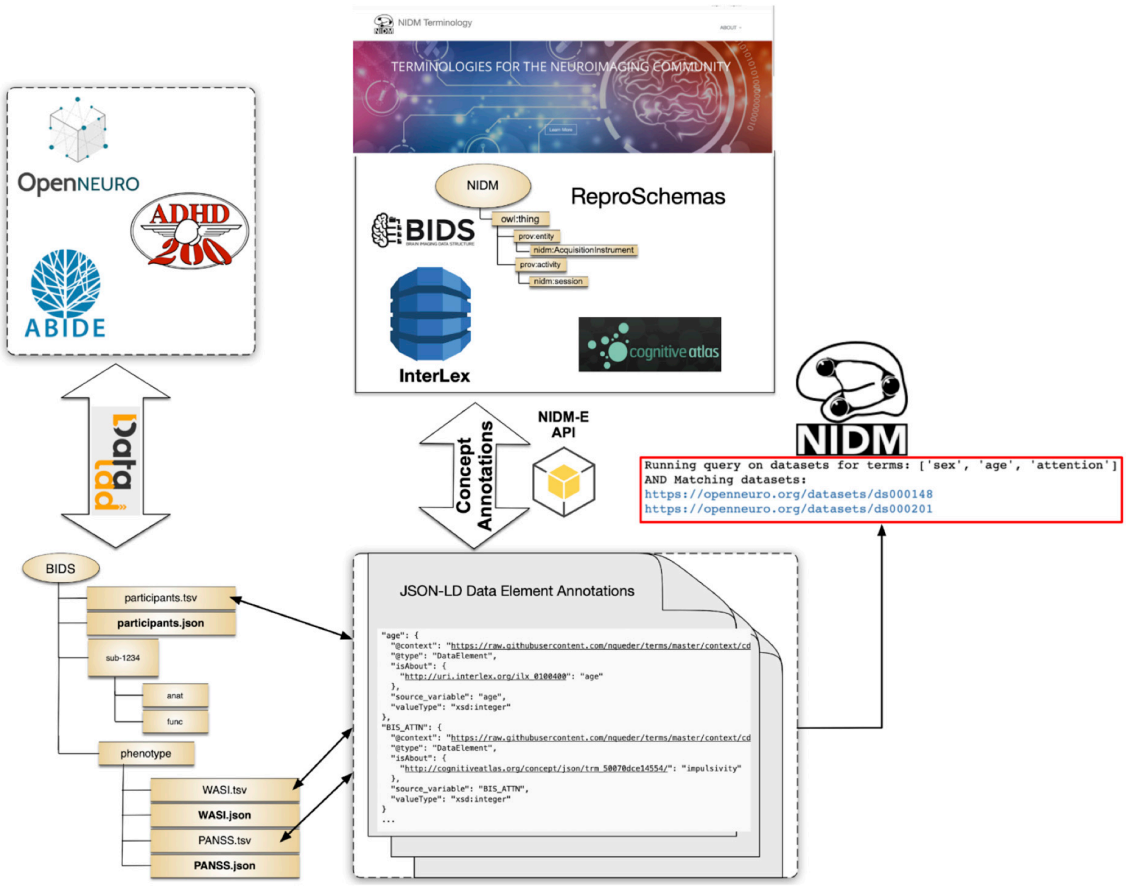


FIGURE 5
An illustration of the NIDM-Terms workflow.

The procedure above resulted in the creation of a NIDM document for each OpenNeuro dataset available via Datalad at the time of initial query. For these datasets we used a different procedure from the ABIDE and ADHD200 NIDM conversions. Here we had to do the annotations in bulk for approximately 300 datasets while we developed, in parallel, the robust concept annotation capabilities of the 'bidsmri2nidm' tool. To save time we decided to crowd-source the annotation activities amongst our NIDM-Terms team by using an export to a Google spreadsheet. As described previously, in practice for a smaller number of datasets, one could (and should) use any of the annotation tools provided with our work.

3.3. Concept-based queries

Now that each of our example datasets (i.e., OpenNeuro, ABIDE, and ADHD200) has been annotated using the methodologies presented here and a NIDM file representation created, we began testing concept-based integration queries. We created two Jupyter notebook query demonstrations available directly in the NIDM-Terms GitHub repository via Binder (see README - Demos[30]): (1) Using the JSON-LD version of our BIDS-compliant JSON "sidecar" files to query across OpenNeuro datasets; (2) Using the NIDM files across all three datasets to search by concept and neuroimaging type. We feel these demonstrations serve to show how a user can query across BIDS datasets using concepts without any backend database (example 1) and using NIDM files across three datasets facilitated through the ReproLake metadata database (example 2) supported by ReproNim (see text footnote 25). In example 2, we add the additional capability of querying for image type alongside concepts. Note, there are many additional pieces of metadata in the NIDM files that could be included, along with the ones shown here, in a production query interface.

With respect to example 1, the Jupyter notebook starts by pulling all the JSON-LD "sidecar" files for the OpenNeuro datasets from the NIDM-Terms GitHub repo. It then creates a dictionary of the concepts used in annotating those data by accessing the "isAbout" property in those JSON-LD files. Next, it uses ipywidgets[31] to create a simple drop-down interface within the Jupyter notebook listing all the concepts available across all annotated OpenNeuro datasets. The user can then add concepts to a query list and perform AND-based or OR-based queries on the list. The notebook then returns a list of datasets in OpenNeuro that

30 https://github.com/NIDM-Terms/terms/blob/master/README.md

31 https://ipywidgets.readthedocs.io/en/stable/

satisfy the query with links out to the OpenNeuro interface for the datasets. These queries are fairly efficient and require no additional database backend.

With respect to example 2, first the NIDM files, created here, for all studies (i.e., ABIDE, ADHD200, and OpenNeuro) were uploaded to ReproLake. ReproLake is a publicly available metadata archive developed on the StarDog[32] platform. Although it is still in development, ReproLake will, in the near future, provide a metadata archive containing NIDM files describing many publicly available neuroimaging datasets. Because querying large RDF graphs across thousands of datasets is quite resource intensive, using a database to support these queries makes them more efficient. One could instead use a local metadata database to store and query these NIDM files by cloning the "Simple2_NIDM_Examples"[33] repository and looking in the folder. Different from example 1, there are no JSON-LD or JSON files used in this demonstration. Here we use the NIDM files directly, served by ReproLake. The Jupyter notebook begins by performing a SPARQL query, sent to the ReproLake server, on the NIDM documents to retrieve the concepts via the "isAbout" predicate. It then queries the neuroimaging scan types from the NIDM documents by looking for data acquisition activities in the NIDM graphs that contain the "nidm:hadImageContrastType" predicate, a term that is part of the NIDM terminology (see section "2.2. Properties"). Next, similar to example 1, these concepts and contrast types get added to ipywidgets and the user can select criteria to query on. The tool then formulates a SPARQL query, presenting this query to the user for educational purposes, and sends the query to the ReproLake StarDog instance. Depending on the complexity of the query, the results can take a few seconds to many minutes (or longer) to complete. Because the ReproLake utility is still in development, no server-side optimizations have been done and limited server resources are available. As ReproNim continues to develop this resource, query response will improve.

4. Discussion

The dataset annotations and terminology management tools presented here have shown to be a useful and pragmatic approach to querying across datasets and linking datasets through mappings from dataset-specific variables and terms to broader concepts. Most of the tools and techniques presented here have been pragmatically-focused and developed, in part, to support building the ReproLake metadata database. We've tried to create models that are sufficiently expressive to capture important information needed to enable data reuse, while minimizing the burden on researchers. Thus far, through efforts connected to ReproNim and the overall NIDM work, we have found the minimal set of properties we've selected to be sufficient to find and reuse data, amongst the datasets we chose.

Through our query demonstrations and additional work with the ReproLake, concept annotations have been successful in helping us search across datasets. During our initial experiments, using the datasets described here and annotated by several individuals in our research team, we found that there was some ambiguity surrounding several similar concept choices. Even for simple variables such as age, sex, and handedness, there were multiple concepts that could be selected from the many available in large terminology management resources such as InterLex. To address this complexity, we've taken two main approaches, enabled by our choice to use RDF and JSON-LD: (1) constrain the search space for often-used concepts; (2) use RDF and linked-data capabilities to start connecting similar/equivalent terms in InterLex. Constraining the search space was accomplished using our NIDM-Terms GitHub repository to maintain a list of concepts selected for annotating previous datasets and to initially present those concepts to users of our tools, effectively giving them a single choice for age, sex, and handedness concepts. This procedure works well if the annotations are performed using our tools and curated term lists but does not address the problem when users are manually annotating data using term resources without guidance. The second approach, that of connecting similar/equivalent terms together within InterLex, has been an on-going project for many years and that project continues to make progress on that front. By connecting terms within InterLex using the RDF framework, one could perform equivalence mapping at query time via the SPARQL query language. Then, one could theoretically select concepts from InterLex without much concern for whether other dataset providers selected the same concept because the similarity and/or equivalency has already been modeled by the InterLex team and is used directly within the ReproLake query engine. This approach would satisfy those doing manual annotations but only when using InterLex. To make this approach scale, the research community should move toward using linked data methods across all metadata included with publicly available datasets. By creating a rich web of linked neuroimaging information, the overhead involved in database-dependent mediation services could be reduced and this linked terminology information would be available to any web resource. This is the promise of linked-data and we are seeing signs of this goal coming to fruition in the broader web, outside of neuroimaging-based scientific data.

Data-sharing requirements from funding agencies and journals, have done much to increase the amount of data available for reuse in the neuroimaging and other related communities over the last 10 years. The work presented here has been successful at providing a framework for annotating study variables in ways to make them more reusable by providing a formal (and minimal) list of properties and tools to support them in the context of the popular BIDS data structure. Further, the process of linking concepts to selected study variables has been successful at showing the promise of an integrated metadata search utility (i.e., ReproLake). Despite these advances, there is still much work to be done to realize a web of linked neuroimaging (neuroscience) data that is fully reusable and findable at scale and across studies. Through continued support from funding bodies and international informatics organizations such as the International Neuroinformatics Coordinating Facility (INCF),[34] we expect

32 https://www.stardog.com/

33 https://github.com/dbkeator/simple2_NIDM_examples/tree/master/datasets.datalad.org

34 https://www.incf.org/

the remaining barriers to slowly crumble such that data shared by any laboratory, globally, could be reused for the advancement of science.

Data availability statement

Publicly available datasets were analyzed in this study. The annotated terms from those datasets and their properties can be found here: https://github.com/nidm-terms.

Author contributions

All authors listed have made a substantial, direct, and intellectual contribution to the work, and approved it for publication.

Funding

This work has been supported by the National Institutes of Health (NIH) grants 1RF1MH120021-01 from the National Institute of Mental Health (NIMH), P41 EB019936 from the National Institute of Biomedical Imaging and Bioengineering (NIBIB), and the International Neuroinformatics Coordinating Facility (INCF). J-BP and SU were partly funded by the Michael J. Fox Foundation (LivingPark), the National Institutes of Health (NIH) NIH-NIBIB P41 EB019936 (ReproNim), the National Institute of Mental Health of the NIH under Award Number R01MH096906 (Neurosynth), as well as the Canada First Research Excellence Fund, awarded to McGill University for the Healthy Brains for Healthy Lives initiative and the Brain Canada Foundation with support from Health Canada. This work has been in part made possible by the Brain Canada Foundation, through the Canada Brain Research Fund, with the financial support of Health Canada and the McConnell Brain Imaging Centre.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

References

Ambite, J., Tallis, M., Alpert, K., Keator, D., King, M., Landis, D., et al. (2015). SchizConnect: Virtual Data Integration in Neuroimaging. *Data Integr. Life Sci.* 9162, 37–51. doi: 10.1007/978-3-319-21843-4_4

Arp, R., Smith, B., and Spear, A. (2015). *Building Ontologies with Basic Formal Ontology*. Cambridge, MA: MIT Press. doi: 10.7551/mitpress/9780262527811.001.0001

Book, G., Anderson, B., Stevens, M., Glahn, D., Assaf, M., and Pearlson, G. (2013). Neuroinformatics Database (n.d.) – A modular, portable database for the storage, analysis, and sharing of neuroimaging data. *Neuroinformatics* 11, 495–505. doi: 10.1007/s12021-013-9194-1

Button, K., Ioannidis, J., Mokrysz, C., Nosek, B., Flint, J., Robinson, E., et al. (2013). Power failure: why small sample size undermines the reliability of neuroscience. *Nat. Rev. Neurosci.* 14:365. doi: 10.1038/nrn3475

Ceusters, W. (2012). An information artifact ontology perspective on data collections and associated representational artifacts. *Stud. Health Technol. Inform.* 180, 68–72.

Crawford, K., Neu, S., and Toga, A. (2016). The image and data archive at the laboratory of neuro imaging. *Neuroimage* 124, 1080–1083. doi: 10.1016/j.neuroimage.2015.04.067

Das, S., Zijdenbos, A., Harlap, J., Vins, D., and Evans, A. (2012). LORIS: a web-based data management system for multi-center studies. *Front. Neuroinform.* 5:37. doi: 10.3389/fninf.2011.00037

Dumontier, M., Baker, C., Baran, J., Callahan, A., Chepelev, L., Cruz-Toledo, J., et al. (2014). The Semanticscience Integrated Ontology (SIO) for biomedical research and knowledge discovery. *J. Biomed. Semantics* 5:14. doi: 10.1186/2041-1480-5-14

Eickhoff, S., Simon, E., Nichols, T., Van Horn, J., and Turner, J. (2016). Sharing the wealth: Neuroimaging data repositories. *Neuroimage* 124, 1065–1068. doi: 10.1016/j.neuroimage.2015.10.079

Facile, R., Muhlbradt, E., Gong, M., Li, Q., Popat, V., Pétavy, F., et al. (2022). Use of Clinical Data Interchange Standards Consortium (CDISC) standards for real-world data: expert perspectives from a qualitative delphi survey. *JMIR Med. Inform.* 10:e30363. doi: 10.2196/30363

Gorgolewski, K., Tibor, A., Calhoun, V., Cameron Craddock, R., Samir, D., Duff, E., et al. (2015). The brain imaging data structure: a standard for organizing and describing outputs of neuroimaging experiments. *bioRxiv* [Preprint]. doi: 10.1101/034561

Halchenko, Y., Meyer, K., Poldrack, B., and Solanky, D. (2021). DataLad: distributed system for joint management of code, data, and their relationship. *J. Open Source* 6:63. doi: 10.21105/joss.03262

incf-nidash (2016). *incf-nidash/nidm*. San Francisco, CA: GitHub.

Keator, D., Grethe, J., Marcus, D., Ozyurt, B., Gadde, S., Murphy, S., et al. (2008). A national human neuroimaging collaboratory enabled by the Biomedical Informatics Research Network (BIRN). *IEEE Trans. Inf. Technol. Biomed.* 12, 162–172. doi: 10.1109/TITB.2008.917893

Keator, D., Helmer, K., Steffener, J., Turner, J., Van Erp, T., Gadde, S., et al. (2013). Towards structured sharing of raw and derived neuroimaging data across existing resources. *Neuroimage* 82, 647–661. doi: 10.1016/j.neuroimage.2013.05.094

Kennedy, D., Haselgrove, C., Riehl, J., Preuss, N., and Buccigrossi, R. (2016). The NITRC image repository. *Neuroimage* 124, 1069–1073. doi: 10.1016/j.neuroimage.2015.05.074

Klein, A., Clucas, J., Krishnakumar, A., Ghosh, S., Van Auken, W., Thonet, B., et al. (2021). Remote Digital Psychiatry for Mobile Mental Health Assessment and Therapy: MindLogger Platform Development Study. *J. Med. Internet Res.* 23:e22369. doi: 10.2196/22369

Maumet, C., Auer, T., Bowring, A., Chen, G., Das, S., Flandin, G., et al. (2016). Sharing brain mapping statistical results with the neuroimaging data model. *Sci. Data* 3:160102. doi: 10.1038/sdata.2016.102

Moreau, L., Luc, M., Bertram, L., Ilkay, A., Barga, R., Shawn, B., et al. (2008). Special Issue: The First Provenance Challenge. *Concurr. Comput.* 20, 409–418. doi: 10.1002/cpe.1233

NIDM (2016). *Neuroimaging Data Model*. New Delhi: NIDM.

Niso, G., Botvinik-Nezer, R., Appelhoff, S., De La Vega, A., Esteban, O., Etzel, J., et al. (2022). Open and reproducible neuroimaging: From study inception to publication. *Neuroimage* 263:119623. doi: 10.1016/j.neuroimage.2022.119623

Pareti, P., Konstantinidis, G., Norman, T., and Şensoy, M. (2019). *SHACL Constraints with Inference Rules. In: The Semantic Web – ISWC 2019*. Berlin: Springer International Publishing, 539–557. doi: 10.1007/978-3-030-30793-6_31

Patridge, E., and Bardyn, T. (2018). Research Electronic Data Capture (REDCap). *J. Med. Libr. Assoc.* 106:142. doi: 10.5195/jmla.2018.319

Poldrack, R., Barch, D., Mitchell, J., Wager, T., Wagner, A., Devlin, J., et al. (2013). Toward open sharing of task-based fMRI data: the OpenfMRI project. *Front. Neuroinform.* 7:12. doi: 10.3389/fninf.2013.00012

PROV-Overview (2016). *PROV-Overview. An Overview of the PROV Family of Documents*. Cambridge, MA: World Wide Web Consortium.

Sansone, S., Rocca-Serra, P., Field, D., Maguire, E., Taylor, C., Hofmann, O., et al. (2012). Toward interoperable bioscience data. *Nat. Genet.* 44, 121–126. doi: 10.1038/ng.1054

Schulz, S. (2018). Faculty of 1000 evaluation for The FAIR Guiding Principles for scientific data management and stewardship. *F1000* [Epub ahead of print]. doi: 10.3410/f.726216348.793543848

Seppälä, S., Ruttenberg, A., and Smith, B. (2017). Guidelines for writing definitions in ontologies. *Ciência Inform.* 46, 73–88.

Surles-Zeigler, M., Sincomb, T., Gillespie, T., de Bono, B., Bresnahan, J., Mawe, G., et al. (2021). Extending and using anatomical vocabularies in the Stimulating Peripheral Activity to Relieve Conditions (SPARC) program. *bioRxiv* [Preprint]. doi: 10.1101/2021.11.15.467961

Szucs, D., and Ioannidis, J. (2017). Empirical assessment of published effect sizes and power in the recent cognitive neuroscience and psychology literature. *PLoS Biol.* 15:e2000797. doi: 10.1371/journal.pbio.2000797

Turner, J., Pasquerello, D., Turner, M., Keator, D., Alpert, K., King, M., et al. (2015). Terminology development towards harmonizing multiple clinical neuroimaging research repositories. *Data Integr. Life Sci.* 9162, 104–117. doi: 10.1007/978-3-319-21843-4_8

Wang, L., Alpert, K., Calhoun, V., Cobia, D., Keator, D., King, M., et al. (2016). SchizConnect: Mediating neuroimaging databases on schizophrenia and related disorders for large-scale integration. *NeuroImage* 124, 1155–1167. doi: 10.1016/j.neuroimage.2015.06.065

Wilkinson, M., Dumontier, M., Aalbersberg, I., Appleton, G., Axton, M., Baak, A., et al. (2016). The FAIR Guiding Principles for scientific data management and stewardship. *Sci. Data* 3:160018.

frontiers | Frontiers in Neuroinformatics

TYPE Original Research
PUBLISHED 24 July 2023
DOI 10.3389/fninf.2023.1216443

OPEN ACCESS

EDITED BY
Maaike M. H. Van Swieten,
Integral Cancer Center Netherlands (IKNL), Netherlands

REVIEWED BY
Rafael Martínez Tomás,
National University of Distance Education (UNED), Spain
Anita Bandrowski,
University of California San Diego, United States

*CORRESPONDENCE
Jessica A. Turner
✉ jessica.turner@osumc.edu

RECEIVED 03 May 2023
ACCEPTED 05 July 2023
PUBLISHED 24 July 2023

CITATION
Sahoo SS, Turner MD, Wang L, Ambite JL, Appaji A, Rajasekar A, Lander HM, Wang Y and Turner JA (2023) NeuroBridge ontology: computable provenance metadata to give the long tail of neuroimaging data a FAIR chance for secondary use.
Front. Neuroinform. 17:1216443.
doi: 10.3389/fninf.2023.1216443

NeuroBridge ontology: computable provenance metadata to give the long tail of neuroimaging data a FAIR chance for secondary use

Satya S. Sahoo[1], Matthew D. Turner[2], Lei Wang[2], Jose Luis Ambite[3], Abhishek Appaji[4], Arcot Rajasekar[5], Howard M. Lander[5], Yue Wang[5] and Jessica A. Turner[2]*

[1]Case Western Reserve University, Cleveland, OH, United States, [2]Department of Psychiatry and Behavioral Health, The Ohio State University Wexner Medical Center, Columbus, OH, United States, [3]University of Southern California, Los Angeles, CA, United States, [4]BMS College of Engineering, Bengaluru, India, [5]University of North Carolina at Chapel Hill, Chapel Hill, NC, United States

Background: Despite the efforts of the neuroscience community, there are many published neuroimaging studies with data that are still not *findable* or *accessible*. Users face significant challenges in *reusing* neuroimaging data due to the lack of provenance metadata, such as experimental protocols, study instruments, and details about the study participants, which is also required for *interoperability*. To implement the FAIR guidelines for neuroimaging data, we have developed an iterative ontology engineering process and used it to create the NeuroBridge ontology. The NeuroBridge ontology is a computable model of provenance terms to implement FAIR principles and together with an international effort to annotate full text articles with ontology terms, the ontology enables users to locate relevant neuroimaging datasets.

Methods: Building on our previous work in metadata modeling, and in concert with an initial annotation of a representative corpus, we modeled diagnosis terms (e.g., schizophrenia, alcohol usage disorder), magnetic resonance imaging (MRI) scan types (T1-weighted, task-based, etc.), clinical symptom assessments (PANSS, AUDIT), and a variety of other assessments. We used the feedback of the annotation team to identify missing metadata terms, which were added to the NeuroBridge ontology, and we restructured the ontology to support both the final annotation of the corpus of neuroimaging articles by a second, independent set of annotators, as well as the functionalities of the NeuroBridge search portal for neuroimaging datasets.

Results: The NeuroBridge ontology consists of 660 classes with 49 properties with 3,200 axioms. The ontology includes mappings to existing ontologies, enabling the NeuroBridge ontology to be interoperable with other domain specific terminological systems. Using the ontology, we annotated 186 neuroimaging full-text articles describing the participant types, scanning, clinical and cognitive assessments.

Conclusion: The NeuroBridge ontology is the first computable metadata model that represents the types of data available in recent neuroimaging

studies in schizophrenia and substance use disorders research; it can be extended to include more granular terms as needed. This metadata ontology is expected to form the computational foundation to help both investigators to make their data FAIR compliant and support users to conduct reproducible neuroimaging research.

KEYWORDS

FAIR neuroimaging data, computable provenance metadata, NeuroBridge ontology, ontology text annotation, W3C PROV ontology

1. Introduction

Reproducible science involving replication and reproducibility using meta-analysis as well as mega-analyses are critical to the advancement of neuroimaging research (Dinov et al., 2010; Poldrack et al., 2017; Kennedy et al., 2019). Reanalysis of a study, either with alternate analyses of the original experiment or with novel analyses that conform to the data is relatively easy if the original data and the associated provenance metadata are available to other researchers (Sahoo et al., 2019; Huber et al., 2020). Well-designed mega- and meta-analyses require the identification of studies that use experimental methods and subjects that are similar or equivalent to the original study; therefore, provenance metadata that describes this contextual information is critical for the identification and harnessing of data from existing studies for rigorous replication. The Findable, Accessible, Interoperable, and Reusable (FAIR) guiding principles adopted in 2014 aim to facilitate the discoverability and accessibility of the useful datasets (Wilkinson et al., 2016). However, concrete implementation of the FAIR guiding principles has been a key challenge (Musen et al., 2022), especially for neuroimaging databases and repositories [The National Institute of Mental Health Data Archive (NDA), 2023], which are often stored in silos with limited support for FAIR principles.

For example, the neuroimaging data repositories supported by different divisions within the US National Institutes of Health (NIH) lack common terminology and representation format for metadata information describing the datasets [The National Institute of Mental Health Data Archive (NDA), 2023]. Similarly, the large volume of neuroimaging datasets that are collected in hundreds of laboratories around the world each year are only described in journal publications without being made accessible through organized data management systems (Sejnowski et al., 2014). These underutilized data form the "long tail of science" (Ferguson et al., 2014; Frégnac, 2017), and finding these datasets requires tedious search of published literature for relevant neuroimaging studies through manual review of papers to extract the provenance metadata of the studies. The metadata terms describe the structure and methods used in the study, such as the profile of the participants recruited for the study (e.g., patients with schizophrenia, cocaine users and their family members), the type of neuroimaging data collected [e.g., T1-weighted imaging, task-based functional magnetic resonance imaging (fMRI)], and the clinical and cognitive assessment instruments used in the study (e.g., SAPS/SANS, RAVLT, AUDIT).

PubMed and Google Scholar search features allow users to find papers related to a neuroimaging question of interest; however, the results do not analyze the study metadata such as the experimental design, the modality of data collected, and the status of data sharing. These existing search engines are powerful tools for exhaustive search using sophisticated artificial intelligence methods to find relevant results; however, the lack of support for FAIR principles makes it difficult for users to find relevant papers with accessible study data. To address this limitation, we are developing the NeuroBridge data discovery platform as part of the NIH-funded Collaborative Research in Computational Neuroscience (CRCNS) program to be a bridge between neuroimaging researchers and the relevant data published in literature. The NeuroBridge platform aims to identify, index, and analyze provenance metadata information from neuroimaging articles available in the PubMed Central repository and map specific studies to user queries related to research hypotheses. The NeuroBridge platform with its multiple components and sources is described in more detail in a companion paper in this Research Topic (Wang et al., Under Review). To enable the modeling of computable metadata that underpins the data search platform, we developed the NeuroBridge ontology based on FAIR guidelines for the neuroimaging domain.

1.1. Standardized provenance for implementing FAIR principles

The FAIR principles have been widely endorsed by funding agencies, including the NIH, individual researchers, and data curators to facilitate open science and maximize the reusability of existing resources. However, the lack of standardized metadata models that can be used by users in a specific domain to implement the FAIR principles and make their datasets FAIR compatible has been noted by recent studies (Musen et al., 2022). It is difficult for investigators to: (1) enumerate the relevant metadata terms that are necessary for understanding the experiment details that generated a dataset, which will ensure that the dataset can be reused either as part of a meta-analysis or new study; and (2) encode the relevant metadata terms in a machine interpretable standard format. In our earlier work in the field of data sharing in neurological disorders such as epilepsy and sleep disorder, we developed a metadata framework that classified provenance metadata related to research studies into the three categories of *study instrument*, *study data*,

and *study method* (called the S3 model) as part of the Provenance for Clinical and Health Research (ProvCaRe) project (Sahoo et al., 2019). The S3 model is built on many existing reproducibility focused metadata guidelines such as the Consolidated Standards of Reporting Trials (CONSORT) guidelines (Schulz et al., 2010), the Animals in Research: Reporting *In Vivo* Experiments (ARRIVE) guidelines (Kilkenny et al., 2010), and the Problem/Population, Intervention, Comparison, Outcome and Time (PICOT) model (Richardson et al., 1995), among other guidelines.

The S3 model was formalized into a computable, machine interpretable format called the ProvCaRe ontology, which extended the World Wide Web Consortium (W3C) PROV specification to represent provenance metadata for biomedical domain. The PROV specification was developed as a standard provenance model for cross-domain interoperability and has been widely used to support FAIR guidelines in a variety of applications (Richardson et al., 1995; Poldrack and Gorgolewski, 2014). The PROV ontology formalized the PROV terms in an ontology using the description logic-based Web Ontology Language (OWL) with built-in extensibility features, which was used to create the ProvCaRe ontology for the broad biomedical domain. The NeuroBridge ontology is built on the same PROV specifications, and it is focused on the neuroimaging domain to support sharing and secondary use of experiment data.

1.2. Related work and the NeuroBridge project

There has long been a recognition of the importance of data sharing in neuroimaging studies and there has been multiple efforts to standardize terminologies describing neuroimagin datasets (Poldrack and Gorgolewski, 2014). We have contributed to or developed multiple projects to formalize aspects of these terminologies, for example neuroanatomical concepts in the Neuroscience Information Framework (NIF) project (Imam et al., 2012), the cognitive processes and measures (CogAtlas) project (Turner et al., 2011), details of the behavioral experiments used in functional neuroimaging (Cognitive Paradigm Ontology) (Turner and Laird, 2012), and the neuroimaging data analysis (NIDM ontology) (Maumet et al., 2016). Although these previous projects include model terms related to various aspects involved in neuroimaging studies, they lack provenance metadata terms at the appropriate level of granularity to describe the clinical or cognitive instruments used, the types of neuroimaging data collected, and information about the groups of study participants. The NeuroBridge ontology addresses this gap for neuroimaging studies.

The NeuroBridge platform overall builds closely on our work done in the SchizConnect project, which was developed to access multiple institutional neuroimaging databases (Wang et al., 2016). The SchizConnect project allowed a researcher to query for datasets that were relevant to their study hypothesis regarding schizophrenia, for example, a query for datasets including individuals with a diagnosis of schizophrenia, male, over 35 years old, and with a resting state fMRI scan on a 3T scanner. In response to this user query, the SchizConnect platform returned the data matching the query criteria from the different studies indexed by the platform to the user for download and analysis. The development of the SchizConnect platform involved the creation of a terminology, a usable subset of terms to describe neuroimaging datasets, which were informed in part by users and in part by the Organization for Human Brain Mapping (OHBM) Committee on Best Practice in Data Analysis and Sharing (COBIDAS) for reporting fMRI studies databases (Turner et al., 2015). The SchizConnect terminology consisted of terms to describe the different types of schizophrenia groups included in the available studies, the imaging types, the scanner information and the other attendant clinical, cognitive, or behavioral data that were part of the SchizConnect database.

The NeuroBridge project aims to generalize and expand the SchizConnect platform to develop a data discovery system that can be a bridge between the needs of neuroimaging researchers and the relevant data from scientific literature. Published articles describing neuroimaging studies and datasets generated in these studies are an important resource for investigators. The NeuroBridge platform aims to automatically extract provenance metadata terms from these articles and use the terms to identify datasets that are highly relevant to a user's research question. In this paper, we describe a novel iterative ontology engineering process that was developed and implemented to create the NeuroBridge ontology that supports: (1) Fine granularity annotation of full text articles describing neuroimaging studies; (2) Automated parsing and indexing of terms describing experimental design details of neuroimaging studies; and (3) Interactive user queries to locate experimental studies that match research terms (**Figure 1**). The automated parsing and indexing of research papers as well as the interactive user queries require the development of machine learning algorithms and web application resources together with the NeuroBridge ontology, therefore, they are outside the scope of this paper and are described in the companion paper (Wang et al., Under Review). The rest of the paper is structured as follows: In the Section "2. Materials and methods," we describe the core components of the NeuroBridge ontology development process for text annotation; In the Section "3. Results," we describe the resulting neuroimaging metadata ontology and its use in annotation of published literature; and in Section "4. Discussion and conclusion," we discuss the broader impact of the NeuroBridge ontology engineering process, the terms of the ontology, and its application in making neuroimaging studies FAIR.

2. Materials and methods

The first phase of the ontology engineering process involved defining the scope of the ontology to support FAIR guidelines in the neuroimaging domain. Given the lack of existing community standards for modeling neuroimaging metadata, we built on our experience in dataset sharing efforts in the SchizConnect project (e.g., subject groups, neuroimaging modalities, cognitive and clinical assessments), and extended them to current literature describing substance abuse disorders studies using neuroimaging studies.

In the second phase, the metadata terms were classified into the three ProvCaRe S3 model categories of study data, instruments, and method. These metadata terms were collaboratively modeled

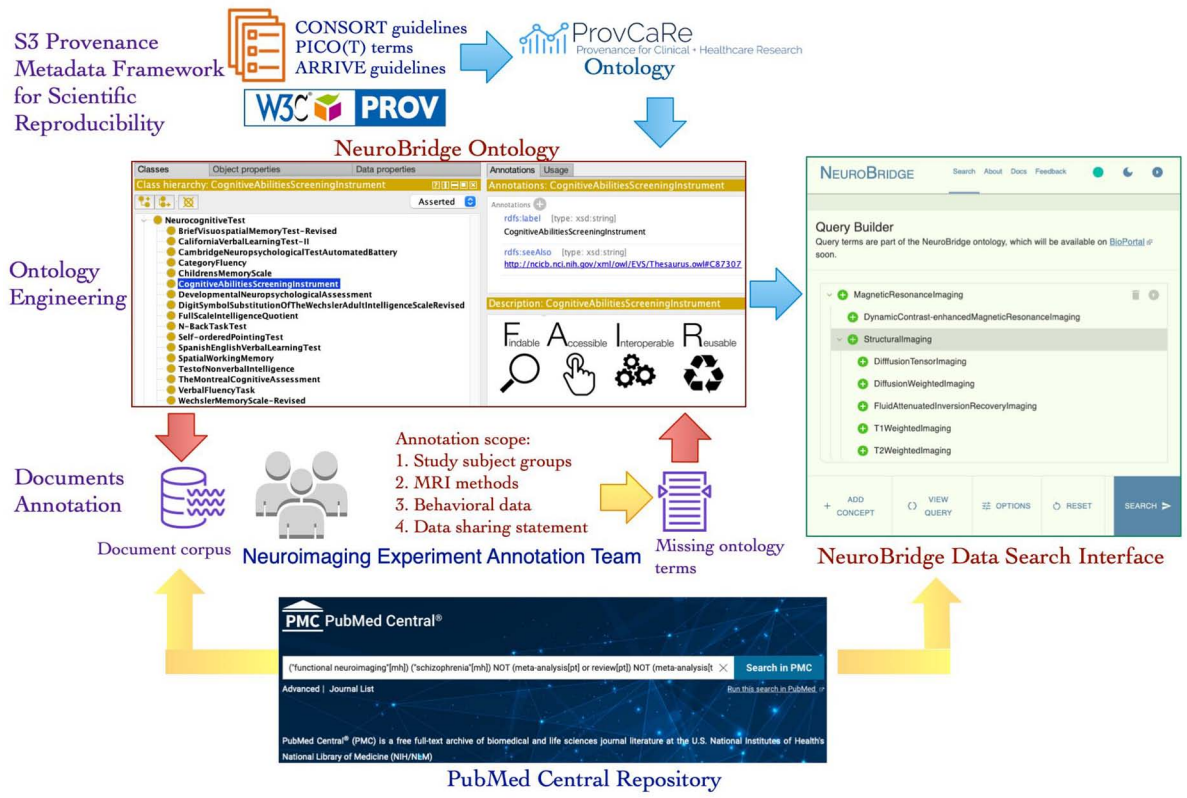


FIGURE 1
An overview of the new ontology engineering workflow developed in the NeuroBridge project to create a FAIR provenance ontology for neuroimaging experiments. The NeuroBridge ontology is based on the ProvCaRe S3 framework for metadata to support FAIR principles and it is used in annotation of full text articles retrieved from PubMed Central repository.

in the NeuroBridge ontology and subsequently used to annotate full text articles describing neuroimaging experiments as part of the third phase of the ontology engineering process. In the final phase, the feedback from the metadata annotation phase was used to evaluate the NeuroBridge ontology followed by extensive restructuring and expansion to meet FAIR guidelines for neuroimaging datasets. Figure 1 is an overview of the new ontology engineering process developed in this project to model computable provenance metadata for neuroimaging experiments.

2.1. Document corpus describing neuroimaging experiments

We created a document corpus consisting of articles describing potential fMRI datasets generated from schizophrenia related studies by querying the PubMed Central repository for papers published between 2017 and 2020 using the following phrases:

```
Query 1: ("functional neuroimaging"[mh]) ("schizophrenia"[mh]) NOT (meta-analysis[pt] or review[pt]) NOT (meta-analysis[ti] or review[ti])
```

Similarly, the following query expanded on the above query with a focus on substance abuse aspect:

```
Query 2: ("functional neuroimaging"[mh]) ("substance-related disorders"[mh]) NOT (meta-analysis[pt] or review[pt]) NOT (meta-analysis[ti] or review[ti])
```

The first query expression generated a corpus consisting of 255 articles, while the second query expression generated 200 articles. We selected 100 articles from each query result to manually process and annotate them using provenance metadata terms. During the annotation phase, we removed articles that were reviews, or meta-analyses, or position papers related to the neuroimaging domain, which resulted in a final count of 186 articles in the document corpus. This corpus included a few papers published on the psychosis datasets available through SchizConnect, but the entirety of the substance abuse papers, and majority of the schizophrenia papers were not part of the SchizConnect project.

2.2. Modeling neuroimaging metadata terms in the NeuroBridge ontology

The W3C PROV specifications support the modeling of provenance metadata for multiple applications, including the description of how datasets were generated to enable their meaningful use (secondary use), reproducibility, and ensuring data quality (Lebo et al., 2013). To achieve these objectives, the PROV model consists of *prov:Entity*, which may be physical or digital (e.g., fMRI images), *prov:Activity* to model the process of creation or modification of entities (e.g., imaging protocol), and

prov:Agent, which takes responsibility for an activity (e.g., study participant). In addition to these terms, the PROV specifications also includes relationships that can be used to represent detailed provenance metadata, for example an experimental study *prov:used* a neurocognitive test of language function [we refer to the PROV specification for further details (Moreau and Missier, 2013)]. The PROV ontology standardized these provenance metadata terms and relationships using OWL expressions. The PROV ontology was extended in the ProvCaRe ontology to standardize the S3 model (Sahoo et al., 2019).

Although the ProvCaRe ontology models the core provenance metadata terms associated with biomedical health domain, the ontology does not model terms at the required level of granularity for neuroimaging experiments. Therefore, the NeuroBridge ontology restructured and expanded the ProvCaRe ontology with a focus on neuroimaging experiments and broadly neuroscience research studies. Our approach is based on ontology engineering best practices to re-use and expand existing ontologies for specific domain applications (Bodenreider and Stevens, 2006). In the initial phase of the ontology engineering process, we reviewed many existing neuroimaging terminologies to identify suitable terms for inclusion in the NeuroBridge ontology. First, we reviewed the SchizConnect terminology list that describes: (1) demography (e.g., socioeconomic status, and handedness scales questions in Edinburgh inventory rating scale among others); (2) psychopathology symptoms (e.g., Calgary depression scale, and Young mania rating scale); (3) extrapyramidal symptoms (e.g., Abnormal involuntary movement scale); (4) functional capacity (e.g., history of motor skills); and (5) medical condition (e.g., Structured clinical interview for the diagnostic statistical manual of mental disorder, SCID) (Spitzer et al., 1990). These five categories of SchizConnect terms were modeled as subtypes of different rating scales in the NeuroBridge ontology (**Figure 2** shows a screenshot of the ontology class hierarchy representing these terms).

In the next step, we reviewed the Neuroimaging Data Model-Experiment (NIDM-E) ontology that was designed to describe different modalities of neuroscience datasets, including terms from the Digital Imaging and Communications in Medicine (DICOM) and Brain Imaging Data Structure (BIDS) specifications (Gorgolewski et al., 2016). We focused on mapping NIDM-E ontology terms describing the method used to generate neuroimaging data and its application in the NeuroBridge ontology. This process was facilitated by collaborative meetings with members of the NIDM-E team members to coordinate the reuse and mapping of terms between the two ontologies. In addition to NIDM-E ontology, we also used the National Center for Biomedical Ontologies (NCBO) BioPortal resource to create mappings between the NeuroBridge ontology and existing ontologies, such as the Radiology Lexicon (RadLex), the Systematized Nomenclature of Medicine Clinical Terms (SNOMED CT), and the National Institute on Drug Abuse (NIDA) common data elements (CDE). The NIDA Clinical Trials Network (CTN) recommended CDEs are part of the National Cancer Institute Data Standards Repository (caDSR), which were created using the metadata registry standard (ISO/IEC 11179) (National Institutes of Health, 2023).

At the end of this first phase of the ontology engineering process, the NeuroBridge ontology had a broad representation of provenance metadata terms describing neuroimaging studies. To evaluate the coverage of the NeuroBridge ontology, we used it for manual annotation of full text articles in our document corpus.

2.3. A two-pass process for text annotation using provenance metadata terms

In the next phase, we implemented a two-pass text annotation process that was designed to be *repeatable*, which could be used to annotate new metadata features in papers as they are identified, and *extensible*, which could be customized for annotation of experimental studies described in broader neuroscience articles. The first "draft expansion" pass was marked by extensive collaboration between the members of the text annotation and the ontology engineering team. The goal of this pass was to identify metadata terms that were needed for annotation of the papers, but they were missing in the first version of the ontology.

This phase was implemented using a variety of online tools, including spreadsheets and shared copies of published articles from the document corpus that were distributed using Google Drive. There were two main workspaces: the first workspace, implemented as a spreadsheet, listed the assignment of annotation team members to specific documents (two annotators per document), which recorded the citation, links to the documents, basic bibliographic data, and notations related to the annotation process, such as the agreement between annotators regarding the metadata annotations. The annotation team members were trained remotely via teleconference due to the coronavirus pandemic. The original team of annotators were trained by co-author JAT to find the relevant parts of papers for annotation.

After this training phase was completed, the annotation teams (with at least two members) were assigned the articles for annotation. A second workspace contained the annotations made by the annotation team members, with each row of this spreadsheet corresponding to a reviewed article and the metadata annotations listed in the columns. Both the workspaces were live documents that were modified by all the members of the annotation team.

The annotation team members focused on the title, abstract, and methods sections of the papers. Their goal was to identify the available or needed labels for each article with four categories of provenance metadata:

1. **Subject groups:** This included disorder types (e.g., schizophrenia, substance abuse) as well as control groups (identified as "no known disorder") if present in the article.
2. **Imaging methods used in the study:** For example, resting state or task-based functional imaging, and T1 weighted imaging.
3. **Behavioral data collected in the study**. For example, standardized scales for symptom severity, cognitive batteries, personality assessments. In addition, unique non-standardized scales, and measures such as medication status or specific cognitive experiment data were also identified and annotated.
4. **Data and resource sharing**. Mark the presence or absence of a formal data sharing statement for the project.

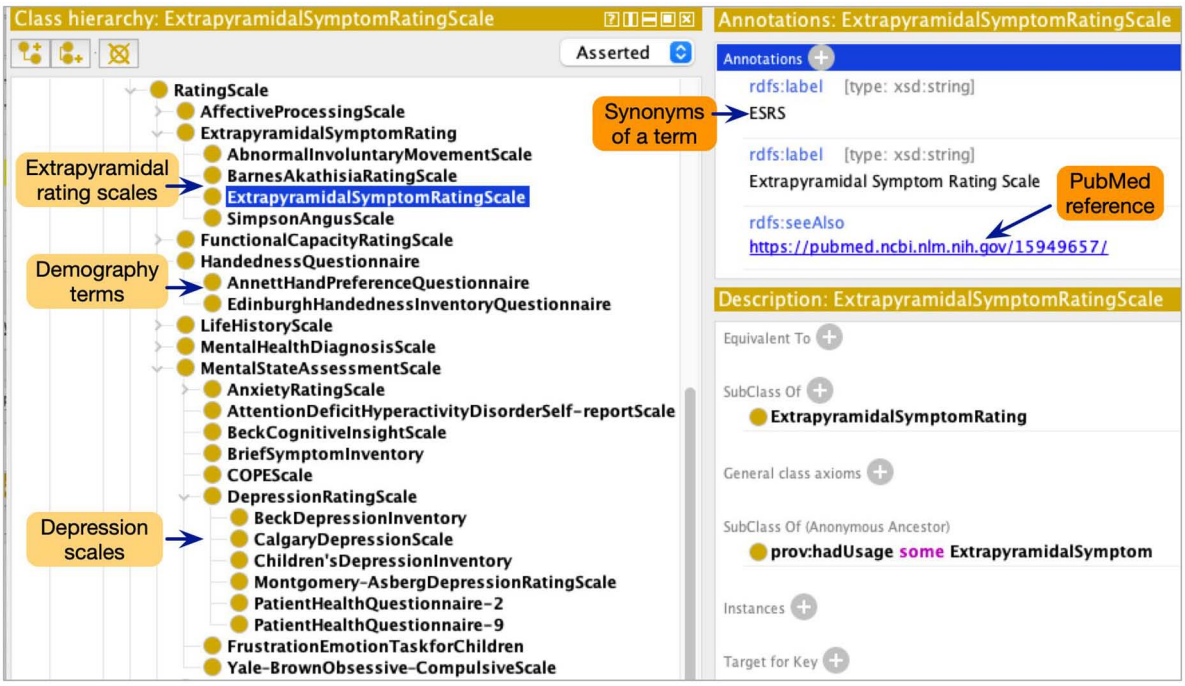


FIGURE 2
The NeuroBridge ontology models a variety of terminology collected as part of the previous SchizConnect project to describe schizophrenia related studies.

The second workspace was used to record the above four categories of provenance metadata annotations associated with specific sections of text in the article. Additional columns in this workspace were used to record the agreement between members of the annotation team regarding the category of metadata terms. The annotators also used this workspace to record metadata terms that could not be mapped to an appropriate ontology term. These missing terms in the ontology together with feedback related to class structure of the ontology were used as feedback by the ontology engineering team to revise the NeuroBridge ontology.

2.4. Revision of the NeuroBridge ontology using text annotation feedback

As part of the tightly coupled cycle of ontology engineering and text annotation, the ontology engineering team agreed that no existing ontology terms were to be removed to preserve backward compatibility with metadata terms already used to annotate the articles. However, the annotations could be modified after the expanded version of the ontology was finalized. The feedback from the annotation phase identified missing terms across all the four categories of provenance metadata, that is, *subject groups*, *imaging methods*, *behavioral assessments*, and *data sharing* policy description.

Within the subject groups category, the ontology engineering team (co-authors, SSS, JAT, and LW), reviewed the modeling approach for representing the distinction between samples of unaffected family members of a study subject with a particular disorder, and "healthy controls." We had already identified that "healthy controls" in any given study may or may not be defined in a consistent manner across studies; therefore, these terms were annotated as a group with "no known disorder" (the corresponding ontology term *NeuroBridge:NoKnownDisorder* was modeled as subclass of *NeuroBridge:ClinicalFinding*). This modeling approach allowed us to represent the information that these participants did not have the given disorder that characterized the other samples in the same study, but there was no guarantee they did not have some other disorder. It is important to note that in disorders with genetic risk, the relatives of affected individuals are of special interest. However, we deferred modeling this provenance information to the next version of the ontology as it required the annotation of a new set of articles describing whether family members of subjects are included in the "no known disorder" group, and the complexity of a family tree (sibling, parents, and multiple generations, among other terms). If the family members were not reported to have been diagnosed with any disorders, the annotation noted that the study collected the "no known disorder" subject group.

A particular challenge in annotating the articles with subject group metadata terms was the need to model modifying attributes of the descriptors in the NeuroBridge ontology. For example, the diagnosis label was not sufficient as a growing number of papers explicitly included subjects with the first episode of psychosis versus subjects with chronic schizophrenia, and unmedicated or medicated status of the study subject. Further, an important distinction in substance use research was not only the type of substance being used but also the "current status of use"; for example, it is important to distinguish between "currently abstinent users," "currently dependent users," and "children of people with addiction." In the NeuroBridge ontology, we represented these through conjunctions of labels, "unmedicated and schizophrenia," "currently abstinent," and "currently using" (Figure 3).

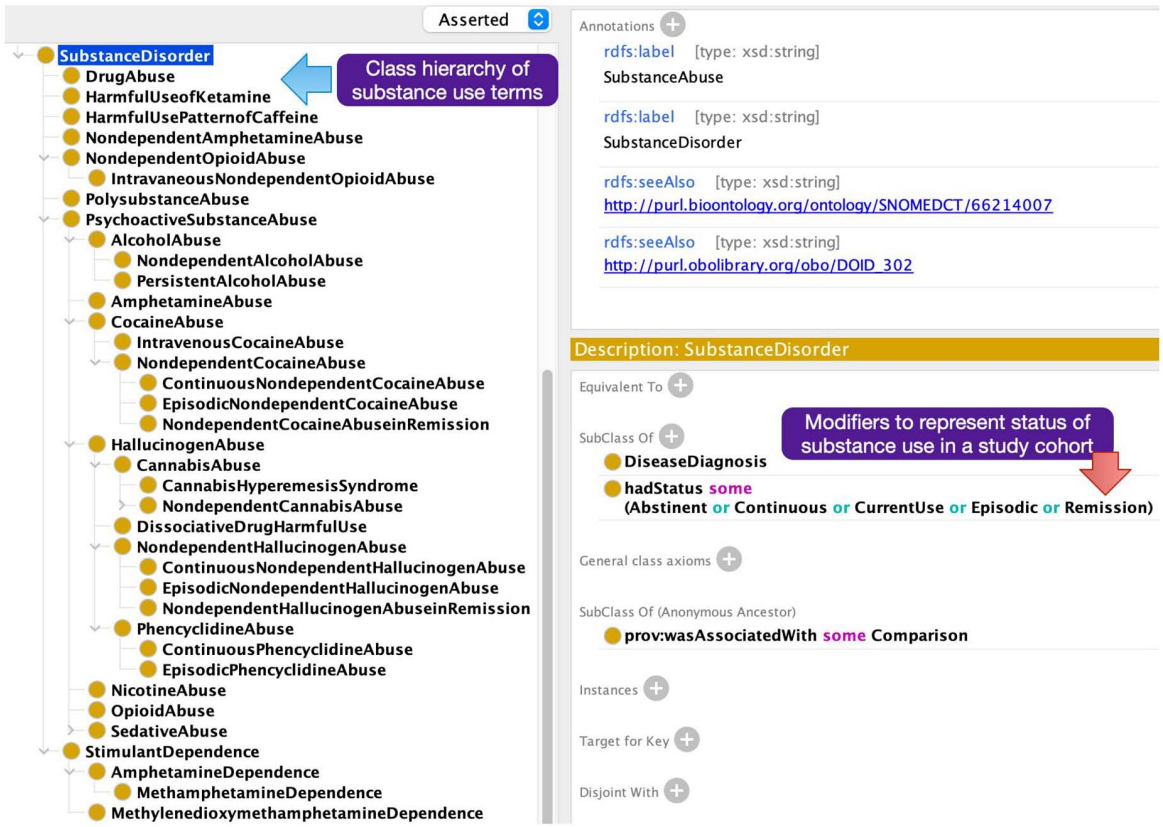


FIGURE 3
Additional neuroimaging metadata terms added to the expanded NeuroBridge ontology based on feedback from the annotation phase.

Similarly, we added new classes to the NeuroBridge ontology to distinguish between imaging protocol, and the imaging modality of the data collected by the imaging protocol. Within both the modality and the protocol branches of the ontology classes, there are common terms describing the types of structural and functional imaging, including task-based and resting state fMRI. The annotation team identified 30 unique terms to describe fMRI tasks in the article corpus. Nine of these terms had been modeled in the CogPO, which had been included in the NeuroBridge ontology. Figure 4A shows the ontology classes describing imaging protocols, which were modeled separately from the imaging modalities. In addition, the NeuroBridge ontology was expanded to model terms describing clinical symptom assessments, diagnostic interviews, and neuropsychological (cognitive) tests. Within the substance use disorder literature, however, there is a research effort focused on impulsivity's relationships with addictive behavior, as well as measures of emotion regulation or openness or other personality traits. We created an initial branch in the ontology for personality assessments as well, to capture those measures. A subset of the Rating Scales is shown in Figure 4B showing (starting in the upper left) the AUDIT scale as an example of the Alcohol Use Scale, which is a type of Substance Use Scale; Substance Craving scales are a separate branch. Neurocognitive scales are not expanded in this view but include various cognitive batteries. The Barratt Impulsivity Scale (not shown in Figure 4) would be an Impulsivity scale class as a subclass of Personality Assessments (top). Clinical ratings of Depression severity (far right) are examples of Mental State Assessments, which are distinct from scales primarily used for diagnosis (modeled as subclasses of the Mental Health Diagnosis Scale class).

2.5. Final annotation phase of the document corpus

Following the first annotation pass through the corpus, and the extensions to the ontology that it entailed as discussed above, the second pass of annotations had a twofold goal of: (a) generating high quality, manually annotated text describing neuroimaging experiments, which were subsequently used to train a Bidirectional Encoder Representations from Transformers (BERT) deep learning model (Devlin et al., 2019; Wang et al., 2022); and (b) validate the metadata term coverage of the NeuroBridge ontology.

To achieve these two goals, an independent set of annotators used the *Inception* text annotation tool to confirm that the annotations originally marked in the spreadsheets could be used in annotating the text (Klie et al., 2018). The Inception tool allows users to select text spans (individual words or phrases) and then connect these spans to terms in the ontology. We used the revised version of the NeuroBridge ontology in this annotation pass (we note that the structure of the ontology remained unchanged during this pass). The annotation team members consisted of trained annotators and a curator. The curator had a supervisory role during the annotation process, specifically with the authority to make

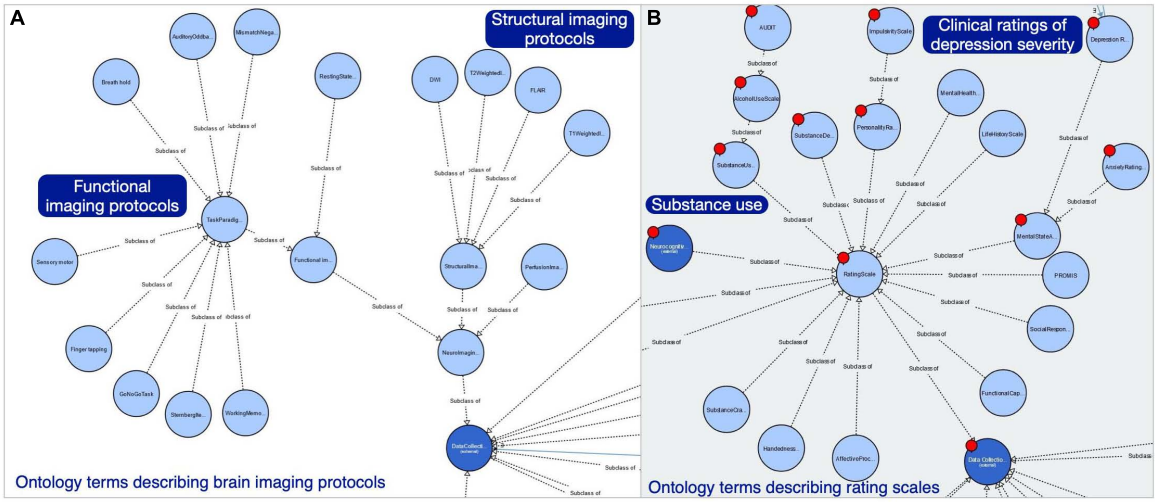


FIGURE 4
Expanded NeuroBridge ontology class structure for **(A)** imaging protocols, and **(B)** rating scales metadata terms used to describe neuroimaging experimental studies.

unilateral decisions in the annotation process. Curators reviewed the work of the annotators and resolved differences in their joint annotations, as well as reviewed all the annotations, and played an important role in ensuring consistency in term usage and application across the corpus.

The roles of annotator and curator were separated, with one of the annotators from the first annotation pass now serving as a curator, and new annotation team members were recruited for this annotation pass. The annotation process was implemented in the following steps: during step 1, a pair of annotators were assigned to an article. The annotators had access to the annotations in the spreadsheets from the first pass, which alerted them to the presence of expected metadata labels in each article. In step 2, each annotator individually reviewed the assigned article and using the annotation from first phase as a guide applied the final metadata annotations. Any Issues identified during this phase were reviewed by the supervisor. The annotators selected the spans of text representing metadata terms describing neuroimaging experiments and marked these with appropriate links to the ontology terms (Figure 5).

After each pair of annotators marked their work as complete, the papers were reviewed by a curator. The Inception software has a curator view of each document that allows direct comparison of the work of each assigned annotator. When the annotators agree completely, the curator can simply mark the annotations as correct or incorrect. When annotators disagree, the curator can decide how to resolve any differences in the final document. Initially the curator was the annotation supervisor. However, at this point some of the more senior annotators from the previous pass had developed sufficient skill; therefore, they were designated as curators for this phase. This allowed volunteer annotators, who had gained significant experience and knowledge about provenance metadata, to move onto a different category of annotation task.

The inter-annotator agreement was computed for the annotations done in Inception; the initial work that used online spreadsheets required the annotators to work in pairs to identify the terms needed for the ontology expansion, so agreement would not be meaningful. Inception calculated Cohen's kappa as measures of pair-wise agreement between annotators, which ranged from 0.75 to 1.0 (mean 0.92). We exported the annotated text corpus from the Inception tool as *WebAnno TSV 3.x* files (this NeuroBridge resource[1]).

3. Results

The new iterative ontology engineering process implemented in this paper resulted in the first release version of the NeuroBridge ontology, consisting of more than 660 classes and 3,200 axioms representing a variety of neuroimaging experiment related provenance metadata. The ontology class expressions leverage more than 40 OWL object properties together with class level restrictions to represent the four categories of metadata information used during the annotation phase of this study. The ontology was evaluated using the Protégé built-in FaCT++ reasoner, which performed classification of concepts using subsumption reasoning followed by satisfiability to identify incorrect subsumptions (Tsarkov and Horrocks, 2006). The standard inference results computed by the reasoner across class, object property, and data property hierarchies as well as class, and object property assertions did not identify any errors in the ontology. The NeuroBridge ontology is made available at the National Center for Biomedical Ontologies (NCBO) Bioportal, as https://bioportal.bioontology.org/ontologies/NEUROBRG.

3.1. Provenance metadata terms used to annotate the document corpus

The 186 articles in the document corpus included annotations with 153 unique metadata terms. The annotation team used the

1 https://github.com/NeuroBridge/Annotation-Project/releases

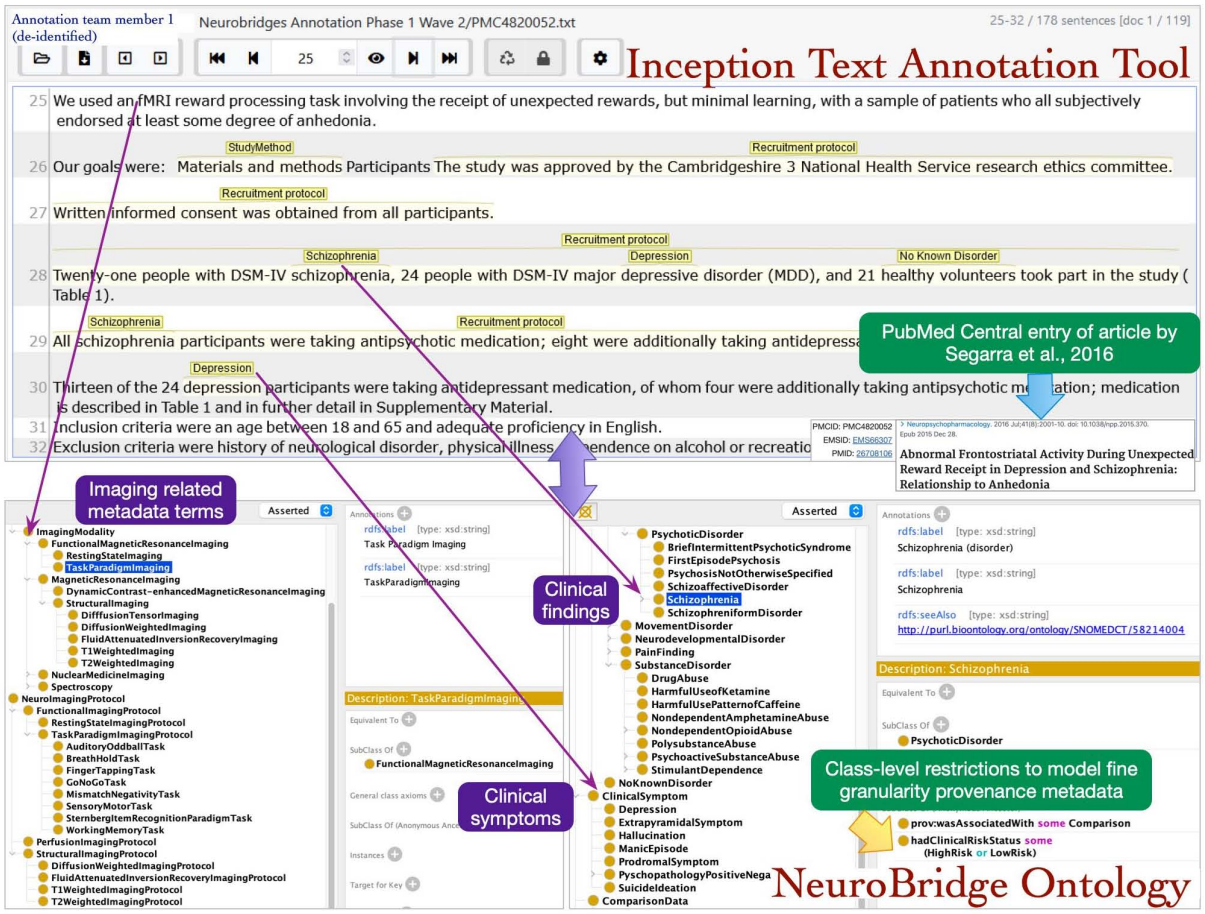


FIGURE 5
The annotation team members identified the text spans in articles during review and mapped the terms to NeuroBridge ontology classes.

metadata terms to label the study method text, the subject groups, the imaging techniques, and additional data. The annotation process ensured that there would be a minimum of one provenance metadata term for each of the first three of those categories, which resulted in a minimum number of concepts per article to be three, with the assumption that there was no other cognitive or behavioral data or information about the recruitment, which may occur with the use of legacy data. Table 1 shows the metadata annotations per paper (including repeats of the same annotation on different blocks of text) and the number of distinct metadata terms per paper. Given multiple imaging data types, multiple possible subject groups, and a wide range of assessments, the number of concepts annotated within the description of the study could range notably, as shown in Table 1.

Conversely, the number of papers referencing each concept ranged between 1 and the entire corpus; the median and average number of *papers per concept* were 3 and 12.54, respectively, representing a skewed distribution of papers referring to concepts. The most common concepts in this corpus are shown in Table 2. As expected, after Study Method and Recruitment Protocol, which is included in almost every study except some papers which used legacy data and gave no details, is the most common subject group (*NeuroBridge:NoKnownDisorder*), and the most common imaging techniques (*NeuroBridge:FunctionalMagneticResonanceImaging* and *NeuroBridge:T1WeightedImaging*). Disorders represented in these papers were chosen to include *NeuroBridge:schizophrenia*, which account for its common use; but substance use disorder was more diverse, with *NeuroBridge:AlcoholAbuse* and *NeuroBridge:CocaineAbuse* being the most common metadata terms.

Surprisingly, only 22% of the articles in our corpus of 186 recently published papers had an explicit data sharing and access statement, despite the increasing focus on data sharing within different domains of biomedical research. This statistic clearly highlights the challenges in making neuroimaging data findable and accessible.

TABLE 1 The descriptive statistics on the paper annotations.

	Annotations per paper	Concepts per paper
Minimum	5	3
Median	33	10
Mean	35	10
Maximum	84	21

TABLE 2 Concepts referred to in at least 10 papers, as well as their general superclass and the number of papers which referred to them.

Concept	Relevant superclass	Number of papers
StudyMethod	Activity	184
RecruitmentProtocol	StudyMethod	183
NoKnownDisorder	ClinicalFinding	154
FunctionalMagneticResonanceImaging	ImagingModality	128
T1WeightedImaging	StructuralImaging	99
MagneticResonanceImaging	ImagingModality	89
Schizophrenia	MentalDisorder/DiseaseDiagnosis	85
RestingStateImaging	FunctionalMagneticResonanceImaging	72
TaskParadigmImaging	FunctionalMagneticResonanceImaging	69
StructuredClinicalInterviewforDSMDisorders	RatingScale	61
MagneticResonanceImagingInstrument	ImagingInstrument	44
PositiveandNegativeSyndromeScale	RatingScale	44
StructuralImaging	ImagingModality	41
AlcoholAbuse	SubstanceDisorder	36
FunctionalImagingProtocol	BrainImaging	32
T2WeightedImaging	StructuralImaging	28
NeurocognitiveTest	RatingScale	26
SubstanceDisorder	DiseaseDiagnosis	21
AlcoholUseDisordersIdentificationTest	RatingScale	18
SchizoaffectiveDisorder	MentalDisorder/DiseaseDiagnosis	17
PsychoticDisorder	MentalDisorder/DiseaseDiagnosis	16
Questionnaire	DataCollectionInstrument	15
CocaineAbuse	SubstanceDisorder	15
FagerstromTestforNicotineDependence	RatingScale	14
ScaleforAssessmentofNegativeSymptoms	RatingScale	12
Electroencephalogram	DiagnosticProcedureOnBrain	12
MedicationStatus	ObservableMeasurement	12
BeckDepressionInventory	RatingScale	11
MentalHealthDiagnosisScale	RatingScale	11
NicotineAbuse	SubstanceDisorder	11
DrugDependence	DrugRelatedDisorder (SNOMED)	10
SubstanceUseScale	RatingScale	10
BipolarDisorder	MentalDisorder/DiseaseDiagnosis	10

3.2. Use of the ontology in the NeuroBridge user portal

In addition to its use in annotation of full-text articles, the NeuroBridge ontology also is incorporated into the NeuroBridge platform for use. The NeuroBridge platform allows users to compose a search query using ontology terms together with logical connectives such as AND, OR. The query expression is automatically expanded using OWL reasoning to include relevant subclasses of a selected ontology term, and this expanded query expression is used to search for neuroimaging experimental studies that match the query constraints (Hitzler et al., 2009). Please see our companion NeuroBridge paper in this Research Topic issue for more details of the platform (Wang et al., Under Review).

4. Discussion and conclusion

The NeuroBridge ontology combines the experience gained from neuroimaging data sharing projects, such as SchizConnect, NI-DM, CogPO and CogAtlas, with the S3 framework of the ProvCaRe project. This combination expands both ProvCaRe

and the previous terminologies to capture important features of multiple domains of biomedical research. This positions NeuroBridge as a backbone for interoperability in annotating the neuroimaging literature. By incorporating substance use disorder papers in the corpus of this study, we confirmed that the S3 framework and the basic SchizConnect terminologies were sufficient to capture metadata information about neuroimaging studies in a different subfield of mental health. However, each of the different categories of metadata terms modeled in the ontology can be further extended to model additional study metadata describing its subject recruitment and data collection methods.

The metadata terms describing MRI techniques are similar across mental health studies and within our 186 functional imaging papers, 84 used task-based imaging, and the remaining 102 (55%) used resting state approaches. Within the task-based neuroimaging, there were a surprisingly limited number of tasks in these papers. The task name was not always specified in the text: For example, the Balloon Analogue Risk Task (BART) or the Monetary Incentive Delay Task were used in 15 of the papers, variations on a cue-reactivity paradigm were used in another nine papers, and the Stop Signal task in another seven papers for example. Naturalistic viewing was used in two papers, and another two dozen tasks such as reality monitoring, paced serial addition, or visual perspective taking were used once each in the corpus. A dozen papers did not include an explicitly named or recognizable imaging paradigm in the text. Future extensions of this corpus are planned to extend the representation of the task paradigms, and to annotate the descriptions of the task in the text. This would allow automated methods for text mining to group papers based on similar task descriptions, and to identify potential task labels that will be added to the ontology.

The rating scales and questionnaires used in the studies in this corpus cover a wide range of topics. We did not create labels for every scale identified in the annotation process. In the ontology, we classified the scales based on higher-level use, such as symptom severity ratings, personality scales, social function scales, and craving scales among others. This is not a challenging issue unique to neuroimaging study metadata, as every domain has its own clinical and cognitive tools, and new assessments and scales are developed continually. The NIH CDEs represent an effort to make data more interoperable, by representing common variables with standard terms. The NIMH National Data Archive (NDA) contains data from highly varied NIMH-funded studies across multiple experimental study designs and subject groups, all tagged with CDE terms. We explored using the NDA's CDEs and matching those against the terms identified in the papers and incorporating them into the ontology, as the data archive is representative of recent research techniques. But there are several notable challenges to that approach, for example the CDEs do not have standardized structure which can be modeled as computable metadata terms. The CDE terms describe specific questions based on the studies that submit them. This can lead to idiosyncratic effects, for example, the term for the Scale for the Assessment of Positive Symptoms (SAPS) is defined as only the formal thought disorder symptom severity part of the SAPS, linked to psychiatric outcomes in Parkinson's Disease, rather than being defined as the Scale itself. These limitations in terms of standardization and lack of structure excluded the NIH CDE for these scales from being modeled in the ontology.

We note that this study successfully demonstrated the implementation of an internationally coordinated metadata annotation process, and online annotation efforts of neuroimaging papers across multiple naive teams. Teams were recruited from several undergraduate programs in the US and in India, and students worked for research credit or in some cases for a summer stipend. The use of current distributed-access tools allowed interactions across teams, levels of expertise, and time zones.

4.1. Ontology-based data access and the application of NeuroBridge ontology

Searching for relevant information over a large corpus is challenging and this task is more difficult if the objective is to query information described in the article's text. In this scenario, exact matches between query term and terms in text are difficult; therefore, deeper domain knowledge in the form of an ontology has been acknowledged to be an effective approach for processing unstructured text in the knowledge representation and Semantic Web communities. The main contribution in our approach is the use of the NeuroBridge ontology in the NeuroBridge search over published articles. The NeuroBridge search feature is designed to use machine learning techniques together with ontologies to support queries beyond simple syntactic and grammar-based term matching; it is designed to use multi-faceted ontology structure to perform domain-specific search. This captures the nuances of data references without being tied down to any specific syntactic structure.

The NeuroBridge ontology and the NeuroBridge platform are distinct from traditional systems such as the Ontology Based Data Access (OBDA) (also called Ontology Mediated or Ontology Based Query Answering) (OMQA/OBQA) (Kock-Schoppenhauer et al., 2017; Xiao et al., 2018; Corcho et al., 2020; Franco et al., 2020; Pankowski, 2021), which are mostly based on relational databases, either across a single database or federated databases with related schemas. The NeuroBridge model can be viewed as a reverse of the mapping advocated by OBDA systems. In our approach we look at ontologies as providing the entities in a database schema and map these ontological structures to sentences or groups of sentences in published articles. This reverse mapping allows us to find references to datasets of interest in an article. This reverse mapping from articles to ontologies is facilitated through the human-annotated stage where identification of relevant sentence structure is performed. These manually annotated examples are used to train machine learning (ML) model to identify similar mappings [described in our companion paper (Wang et al., Under Review)].

4.2. Limitations

A key limitation of this study is the use of a time-intensive ontology engineering process, which makes it challenging to scale the NeuroBridge ontology to include other domains such as cardiac or spinal imaging studies, or even brain tumor scanning. This would require novel expansion methods to be implemented to add new terms in the ontology. As noted above, we do not explicitly

model all published assessments, and the model of subject groups, as currently implemented, does not capture all possibilities. We also have not modeled all the possible details of a neuroimaging study, for example, imaging protocol parameters, quality assurance steps, data processing and analysis phases together with their many parameters, and the statistical results or their interpretation. This version of the ontology would not support, for example, searching for datasets of a certain sample size, which used a particular MRI platform machine, or user queries based on their conclusions (e.g., searching for datasets which were used to support a certain hypothesis).

The representation of neuroimaging behavioral tasks in the ontology does not include the Cognitive Paradigm Ontology (CogPO) approach, which focused on describing the choice of stimuli, the instructions given to the subject, and the responses that the subjects were expected to make till now (Turner and Laird, 2012). The CogPO approach would be more detailed, and would allow disambiguation between, for example, studies which claimed the same type of task but used different stimuli, or between studies which used different names for the same task. This level of detail was considered to be outside of the scope for the first version of the ontology; therefore, it will be part of future expansion of the ontology.

5. Conclusion

The goal of this project is to apply metadata annotations which address FAIR guidelines to the literature of published human neuroimaging studies, even though the studies themselves may not be sharing their datasets through FAIR-compliant methods. The objective of the study is to meet an important requirement of making neuroimaging metadata computable through the NeuroBridge ontology, which will enable neuroimaging data to be compliant with the FAIR guidelines. The use of the ontology for text annotation and supporting user queries in the NeuroBridge portal will allow us to identify and present the relevant neuroimaging papers to the user and to request access from the study authors as necessary for re-use of experimental data. For the purposes of finding neuroimaging datasets that use similar methods and that could be aggregated for a novel analysis, the NeuroBridge ontology is addressing what current ontologies in the fields today are lacking, i.e., describing the methods that neuroimaging studies employed to collect the data. The NeuroBridge ontology is available at https://neurobridges.org/, and in BioPortal (Musen et al., 2012). The corpus will be available on the NeuroBridge website as well for re-use by the community (see text footnote 1). The NeuroBridge platform has been submitted to rrids.org for consideration for a research resource identifier.

Data availability statement

The original contributions presented in this study are publicly available. The NeuroBridge ontology data is available through the NCBO Bioportal as https://bioportal.bioontology.org/ontologies/NEUROBRG and the annotated text will be available through https://github.com/NeuroBridge/Annotation-Project/releases.

Author contributions

SS, LW, MT, and JT conceptualized and designed the study. JT, MT, AA, and YW developed the annotation process involving the identification of metadata terms and their use in annotation of neuroimaging articles, with input from HL. SS, LW, and JT revised and developed the ontology with input from JA, AA, HL, AR, MT, and YW. JT, MT, and SS co-wrote the first draft of the manuscript. All authors contributed to the manuscript revision and read and approved the submitted version of the manuscript.

Funding

The efforts described in this manuscript are funded by NIDA grant R01 DA053028 "CRCNS:NeuroBridge: Connecting big data for reproducible clinical neuroscience," the NSF Office of Cyberinfrastructure OCI-1247652, OCI-1247602, and OCI-1247663 grants, "BIGDATA: Mid-Scale: ESCE: DCM: Collaborative Research: DataBridge–A Sociometric System for Long Tail Science Data Collections," and by the NSF IIS Division of Information and Intelligent Systems grant number #1649397 "EAGER: DBfN: DataBridge for Neuroscience: A novel way of discovery for Neuroscience Data," NIMH grant U01 MH097435," SchizConnect: Large-Scale Schizophrenia Neuroimaging Data Mediation and Federation," NSF grant 1636893 SP0037646, "BD Spokes: SPOKE: MIDWEST: Collaborative: Advanced Computational Neuroscience Network (ACNN)."

Acknowledgments

We would like to acknowledge the following students from Georgia State University for their work on the initial annotations, and from BMS College of Engineering, India for their contribution to the final annotation process: From GSU, Jordan Guffie, Caroline Soares Iizuka, Inara Jawed, Shrusti Joshi, Akhila Madichetty, Grace Marcus, Sumeet Singh, and Vaishnavi Veeranki; from BMS College of Engineering, Ananya Markande, Ojasvi Bhasin, Shriraksha M, Swarna Kedia, and Vaishnavi Haritwal.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

References

Bodenreider, O., and Stevens, R. (2006). Bio-ontologies: Current trends and future directions. *Brief. Bioinform.* 7, 256–274. doi: 10.1093/bib/bbl027

Corcho, O., Priyatna, F., and Chaves-Fraga, D. (2020). Towards a new generation of ontology based data access. *Semant. Web* 11, 153–160. doi: 10.3233/SW-190384

Devlin, J., Chang, M. W., Lee, K., and Toutanova, K. (2019). Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv* [preprint] arXiv:181004805.

Dinov, I., Lozev, K., Petrosyan, P., Liu, Z., Eggert, P., Pierce, J., et al. (2010). Neuroimaging study designs, computational analyses and data provenance using the LONI pipeline. *PLoS One* 5:e13070. doi: 10.1371/journal.pone.0013070

Ferguson, A. R., Nielson, J. L., Cragin, M. H., Bandrowski, A. E., and Martone, M. E. (2014). Big data from small data: Data-sharing in the'long tail'of neuroscience. *Nat. Neurosci.* 17:1442. doi: 10.1038/nn.3838

Franco, W., Avila, C. V. A. O. S., Maia, G., Brayner, A., Vidal, V. M. P., Carvalho, F., et al. (2020). "Ontology-based question answering systems over knowledge bases: A survey," in *Proceedings of the 22nd international conference on enterprise information systems (ICEIS)*, (Setúbal: SCITEPRESS Digital Library). doi: 10.5220/0009392205320539

Frégnac, Y. (2017). Big data and the industrialization of neuroscience: A safe roadmap for understanding the brain? *Science* 358, 470–477. doi: 10.1126/science.aan8866

Gorgolewski, K. J., Auer, T., Calhoun, V. D., Craddock, R. C., Das, S., Duff, E. P., et al. (2016). The brain imaging data structure, a format for organizing and describing outputs of neuroimaging experiments. *Sci. Data* 3, 1–9. doi: 10.1038/sdata.2016.44

Hitzler, P., Krötzsch, M., Parsia, B., Patel-Schneider, P. F., and Rudolph, S. (2009). *OWL 2 web ontology language primer*. Cambridge, MA: World Wide Web Consortium W3C.

Huber, S. P., Zoupanos, S., Uhrin, M., Talirz, L., Kahle, L., Häuselmann, R., et al. (2020). AiiDA 1.0, a scalable computational infrastructure for automated reproducible workflows and data provenance. *Sci. Data* 7:300. doi: 10.1038/s41597-020-00638-4

Imam, F. T., Larson, S. D., Bandrowski, A., Grethe, J. S., Gupta, A., and Martone, M. E. (2012). Development and use of ontologies inside the neuroscience information framework: A practical approach. *Front. Genet.* 3:111. doi: 10.3389/fgene.2012.00111

Kennedy, D. N., Abraham, S. A., Bates, J. F., Crowley, A., Ghosh, S., Gillespie, T., et al. (2019). Everything matters: The ReproNim perspective on reproducible neuroimaging. *Front. Neuroinform.* 13:1. doi: 10.3389/fninf.2019.00001

Kilkenny, C., Browne, W. J., Cuthill, I. C., Emerson, M., and Altman, D. G. (2010). Improving bioscience research reporting: The ARRIVE guidelines for reporting animal research. *PLoS Biol.* 8:e1000412. doi: 10.1371/journal.pbio.1000412

Klie, J. C., Bugert, M., Boullosa, B., de Castilho, R. E., and Gurevych, I. (2018). "The inception platform: Machine-assisted and knowledge-oriented interactive annotation," in *Proceedings of the 27th international conference on computational linguistics: System demonstrations*, (Santa Fe, NM: Association for Computational Linguistics).

Kock-Schoppenhauer, A. K., Kamann, C., Ulrich, H., Duhm-Harbeck, P., and Ingenerf, J. (2017). Linked data applications through ontology based data access in clinical research. *Stud. Health Technol. Inform.* 235, 131–135.

Lebo, T., Sahoo, S. S., and McGuinness, D. (2013). *PROV-O: The PROV ontology*. Cambridge, MA: World Wide Web Consortium.

Maumet, C., Auer, T., Bowring, A., Chen, G., Das, S., Flandin, G., et al. (2016). Sharing brain mapping statistical results with the neuroimaging data model. *Sci. Data* 3:160102. doi: 10.1038/sdata.2016.102

Moreau, L., and Missier, P. (2013). *PROV data model (PROV-DM)*. Cambridge, MA: World Wide Web Consortium W3C.

Musen, M. A., Noy, N. F., Shah, N. H., Whetzel, P. L., Chute, C. G., Story, M. A., et al. (2012). The national center for biomedical ontology. *J. Am. Med. Inform. Assoc.* 19, 190–195. doi: 10.1136/amiajnl-2011-000523

Musen, M. A., O'Connor, M. J., Schultes, E., Martínez-Romero, M., Hardi, J., and Graybeal, J. (2022). Modeling community standards for metadata as templates makes data FAIR. *Sci. Data* 9:696. doi: 10.1038/s41597-022-01815-3

National Institutes of Health (2023). *NIH CDE repository*. Available online at: https://cde.nlm.nih.gov/home (accessed June 30, 2023).

Pankowski, T. (2021). Modeling and querying data in an ontology-based data access system. *Procedia Comput. Sci.* 192, 497–506. doi: 10.1016/j.procs.2021.08.051

Poldrack, R. A., and Gorgolewski, K. J. (2014). Making big data open: Data sharing in neuroimaging. *Nat. Neurosci.* 17, 1510–1517. doi: 10.1038/nn.3818

Poldrack, R. A., Baker, C. I., Durnez, J., Gorgolewski, K. J., Matthews, P. M., Munafò, M. R., et al. (2017). Scanning the horizon: Towards transparent and reproducible neuroimaging research. *Nat Rev. Neurosci.* 18, 115–126. doi: 10.1038/nrn.2016.167

Richardson, W. S., Wilson, M. C., Nishikawa, J., and Hayward, R. S. (1995). The well-built clinical question: A key to evidence-based decisions. *ACP J. Club* 123, A12–A13. doi: 10.7326/ACPJC-1995-123-3-A12

Sahoo, S. S., Valdez, J., Kim, M., Rueschman, M., and Redline, S. (2019). ProvCaRe: Characterizing scientific reproducibility of biomedical research studies using semantic provenance metadata. *Int. J. Med. Inform.* 121, 10–18. doi: 10.1016/j.ijmedinf.2018.10.009

Schulz, K. F., Altman, D. G., Moher, D., and Consort Group (2010). CONSORT 2010 statement: Updated guidelines for reporting parallel group randomised trials. *J. Clin. Epidemiol.* 63, 834–840. doi: 10.1016/j.jclinepi.2010.02.005

Sejnowski, T. J., Churchland, P. S., and Movshon, J. A. (2014). Putting big data to good use in neuroscience. *Nat. Neurosci.* 17, 1440–1441. doi: 10.1038/nn.3839

Spitzer, R. L., Williams, J. B., Gibbon, M., and First, M. B. (1990). *User's guide for the structured clinical interview for DSM-III-R: SCID*. Washington, DC: American Psychiatric Association.

The National Institute of Mental Health Data Archive (NDA) (2023). Available online at: https://data-archive.nimh.nih.gov/ (accessed June 30, 2023).

Tsarkov, D., and Horrocks, I. (2006). *FaCT++ description logic reasoner: System description. automated reasoning*. Berlin: Springer Berlin Heidelberg.

Turner, J. A., and Laird, A. R. (2012). The cognitive paradigm ontology: Design and application. *Neuroinformatics* 10, 57–66. doi: 10.1007/s12021-011-9126-x

Turner, J. A., Frishkoff, G., and Laird, A. R. (2011). "Ontology harmonization between fMRI and ERP: CogPO and NEMO," in *Proceedings of the 41th annual meeting of the society for neuroscience*, Washington, DC.

Turner, J. A., Pasquerello, D., Turner, M. D., Keator, D. B., Alpert, K., King, M., et al. (2015). Terminology development towards harmonizing multiple clinical neuroimaging research repositories. *Data Integr. Life Sci.* 9162, 104–117. doi: 10.1007/978-3-319-21843-4_8

Wang, L., Alpert, K. I., Calhoun, V. D., Cobia, D. J., Keator, D. B., King, M. D., et al. (2016). SchizConnect: Mediating neuroimaging databases on schizophrenia and related disorders for large-scale integration. *Neuroimage* 124(Pt B), 1155–1167. doi: 10.1016/j.neuroimage.2015.06.065

Wang, L., Ambite, J. L., Appaji, A. M., Bijsterbosch, J., Dockès, J., Herrick, R., et al. (Under Review). NeuroBridge: A prototype platform for discovery of the long-tail neuroimaging data. *Front. Neuroinform.*

Wang, X., Wang, Y., Ambite, J.-L., Appaji, A., Lander, H., Moore, S., et al. (2022). "Enabling scientific reproducibility through FAIR data management: An ontology-driven deep learning approach in the NeuroBridge Project," in *Proceedings of the AMIA annual symposium*, Washington, DC.

Wilkinson, M. D., Dumontier, M., Aalbersberg, I. J., Appleton, G., Axton, M., Baak, A., et al. (2016). The FAIR Guiding Principles for scientific data management and stewardship. *Sci. Data* 3:160018.

Xiao, G., Calvanese, D., Kontchakov, R., Lembo, D., Poggi, A., Rosati, R., et al. (2018). "Ontology-based data access: A survey," in *Proceedings of the 27 international joint conferences on artificial intelligence*, Stockholm, 5511–5519. doi: 10.24963/ijcai.2018/777

frontiers | Frontiers in Neuroinformatics

TYPE Original Research
PUBLISHED 29 August 2023
DOI 10.3389/fninf.2023.1233121

OPEN ACCESS

EDITED BY
Christian Haselgrove,
UMass Chan Medical School, United States

REVIEWED BY
Dov Greenbaum,
Yale University, United States
Orla Shortall,
The James Hutton Institute, United Kingdom

*CORRESPONDENCE
Damian Eke
✉ damian.eke@dmu.ac.uk

RECEIVED 01 June 2023
ACCEPTED 09 August 2023
PUBLISHED 29 August 2023

CITATION
Eke D, Ogoh G, Knight W and Stahl B (2023) Time to consider animal data governance: perspectives from neuroscience. *Front. Neuroinform.* 17:1233121. doi: 10.3389/fninf.2023.1233121

Time to consider animal data governance: perspectives from neuroscience

Damian Eke[1]*, George Ogoh[2], William Knight[1] and Bernd Stahl[2]

[1]Centre for Computing and Social Responsibility, De Montfort University, Leicester, United Kingdom, [2]School of Computer Science, University of Nottingham, Nottingham, United Kingdom

Introduction: Scientific research relies mainly on multimodal, multidimensional big data generated from both animal and human organisms as well as technical data. However, unlike human data that is increasingly regulated at national, regional and international levels, regulatory frameworks that can govern the sharing and reuse of non-human animal data are yet to be established. Whereas the legal and ethical principles that shape animal data generation in many countries and regions differ, the generated data are shared beyond boundaries without any governance mechanism. This paper, through perspectives from neuroscience, shows conceptually and empirically that there is a need for animal data governance that is informed by ethical concerns. There is a plurality of ethical views on the use of animals in scientific research that data governance mechanisms need to consider.

Methods: Semi-structured interviews were used for data collection. Overall, 13 interviews with 12 participants (10 males and 2 females) were conducted. The interviews were transcribed and stored in NviVo 12 where they were thematically analyzed.

Results: The participants shared the view that it is time to consider animal data governance due to factors such as differences in regulations, differences in ethical principles, values and beliefs and data quality concerns. They also provided insights on possible approaches to governance.

Discussion: We therefore conclude that a procedural approach to data governance is needed: an approach that does not prescribe a particular ethical position but allows for a quick understanding of ethical concerns and debate about how different positions differ to facilitate cross-cultural and international collaboration.

KEYWORDS

animal research, animal data, neuroscience, data governance, ethics dumping, regulations

1. Introduction

In the last decade, the need to ensure reproducibility of research results and to justify public investment in research has led to increased sharing of research data and the imperative for open sharing. Open data platforms supported by research projects have increasingly become the center piece for facilitating open sharing of research data. In neuroscience, a number of these open platforms exist and share big, multitype and multifunctional data

from diverse species of organisms for both research and innovation. As Poldrack and Gorgolewski (2014) pointed out, these platforms not only encourage data re-use and increase statistical power to stimulate translational knowledge but also expand the reach and impact of neuroscience research. A critical implication of this data re-use expansion is that data increasingly interacts with different jurisdictions with different regulatory requirements. While most of these datasets are from human participants, many are generated from animals.

Unlike human data that is nationally or regionally regulated (e.g., EU General Data Protection Regulation[1]), there are no established legal frameworks that govern the sharing and re-use of animal data nationally or internationally. One reason for this is that the sharing or use of animal data does not raise the traditional data use concerns associated with human data (Stahl et al., 2019) such as; privacy, fairness, human rights and security. The ethical and legal issues around animal data are usually raised during data creation. The scientific, ethical and legal validity of animal research data are mostly determined by the nature of the research procedure/experiment. The moral and legal questions of animal experiments do not always revolve around the implications of animal data usage but on the ethical and legal permissibility of its scientific generation. Crucially, the regulatory and ethical principles that shape animal data generation in many countries and regions are different while the generated data are shared beyond boundaries without any governance mechanism. This means that animal data generated in less restrictive places are openly shared in countries with very restrictive requirements mostly through open data platforms. Thus, this paper asks the question: is it time to consider animal data governance?

The paper shows conceptually and empirically that there is a need for animal data governance that is informed by ethical concerns. Animal data raises different ethical concerns from human data and thus needs to be treated differently. There is a plurality of ethical perspectives on the use of animals in research which any data governance regime will need to take into account. This article therefore arrives at the conclusion that a procedural approach to data governance is called for that does not prescribe a particular ethical position but allows for a quick understanding of ethical concerns and debate about how different positions differ to facilitate cross-cultural and international collaboration.

We organize this paper as follows. We explore the current international data governance ecosystem for responsible biomedical research and innovation; the continued use of animals for neuroscience research, especially non-human primates. We then provide a thematic analysis of interviews conducted with international neuroscientists who conduct animal experiments. On this basis we arrive at the conclusion that a procedural approach to international data governance is called for.

2. The continued use of animals in neuroscience research

Despite the increasing requirements to implement the 3Rs (replacement, refinement, and reduction) (Russell and Burch, 1960; Guhad, 2005), the use of animals in research continues in many parts of the world, especially in neuroscience research (perhaps more than in any other field of biomedical research) (Jones, 2021). Although public interest in the use of animals for research has significantly reduced the number of animal experiments in Europe and North America (Lankau et al., 2014; European Commission, 2019), animals continue to be central to scientific research in other parts of the world. The rationale for this is varied. Neuroscience research often involves invasive and non-invasive methods that cannot be conducted with humans because of associated risks and ethical concerns. Thus, neuroscientists turn to other animals such as; rodents, ferrets, dogs, pigs, zebra fish and monkeys whose usage in scientific research comes with considerably fewer ethical concerns. The argument for this, borders on the lack of safe and non-invasive approaches to studying the human brain (Preuss, 2010). Rodents are widely used because of their short gestation periods, their cost-effective production and have proved unquestionably effective (Neuhaus, 2018). Fruit flies (Zweier et al., 2009) and zebrafish (Haesemeyer and Schier, 2015) have also proved to be important research resources for neuroscience research.

The use of animals in invasive and intrusive neuroscience research is generally informed by the legal and ethical restrictions of such research with human brains. However, many neuroscientists suggest that there are anatomical and genetic limitations/differences that hamper the scope of their use (Garner, 2014; Windisch, 2014). These animal subjects or models present challenges Kaiser and Feng (2015) referred to as the lack of face validity and predictive validity. The delays in the development of new interventions for brain diseases are associated with the differences in pathophysiological mechanisms between rodents and humans (Ting and Feng, 2013). Differences in brain functions and cognitive behavior, difficulties in scalability of dosage regimens, differences in recovery times and differences in the ratio of white to gray matter in the brain are some of the issues that inform this lack of transnationality (Varki, 2000; Weatheall, 2015). For better predictive validity, some neuroscientists assert that non-human primates (NHPs) are particularly better subjects.

2.1. NHPs in neuroscience

Available draft genome sequences of primates have shown that there are important similarities between human and NHP genomes. NHPs have, indeed, been revealed as our closest relatives with regard to the DNA sequence of our genome (Li and Saunders, 2005). While the Chimpanzee genome is 98.77% similar to the human genome (Chimpanzee Sequencing and Analysis Consortium, 2005), the rhesus macaque has a 93.5% similarity (Disotell and Tosi, 2007). This phylogenetic proximity to humans and related similarities in anatomy, physiology and behavior form the basis for justification of the use of NHPs in neuroscience experiments (Tardif et al., 2013; Friedman et al., 2017). A further comparative study of primate genomics has also shown that the most significant evolutionary change between primates happened in the brain (Sikela, 2006). These similarities form the basis of justification and a remarkable curiosity to unlock the brain's complex structure and functions through experiments with NHPs. Today, NHPs are mainly used in basic/fundamental

1 https://gdpr-info.eu/

neurobiological research to explore how brain circuits contribute to human brain activities such as; perception, attention, memory and emotion (Bystron et al., 2006). In Europe, some neuroscientists are engaged in this type of research even though the number of invasive/intrusive experiments ongoing is not known (Scientific Committee on Health, Environmental and Emerging Risks [SCHEER], 2017). They are also used in translational and applied neuroscience research (Capitanio and Emborg, 2008) aimed at understanding the causes and development of potential treatments of brain disorders. Even though the scientific and translational validity of NHP experiments have been questioned (Knight, 2007; Bailey and Taylor, 2016), NHP research in neuroscience continues to develop new tools and approaches.

There is also a steady rise in neurological alterations of NHPs to model human brain diseases/functions and to study the genetic mechanisms that inform human specific neurological changes (Shi et al., 2019). This can be in the form of neural grafting (Bjugstad and Sladek, 2006), transgenesis (Chan, 2014) or other forms of human-NHP chimerism. These experiments significantly alter the neurobiological appearance, behavior or genetic makeup of the NHP causing phenotypic changes. Transgenesis refers to the artificial transfer of a foreign gene into the genome of another organism in order to introduce or delete characteristics of the phenotype (Mepham et al., 1998). In neuroscience, this can involve the use of HLS (human lineage specific) genes to create transgenic NHPs to demonstrate changes in brain structure (like brain size), function (such as high cognition) or to model diseases (like autism, Huntington diseases etc.). These genetic alterations of NHPs are developed with customized mutations and have shown to be able to model human brain disorders like Parkinson's (Yun et al., 2015), Schizophrenia (Qiu and Li, 2017), Alzheimer's (Yeo et al., 2015), autism (Cyranoski, 2016; Zhao et al., 2018) and Huntington's disease (Tomioka et al., 2017). Another invasive NHP experiment in neuroscience involves the transplantation of human-derived neural cells into an NHP to model "human-like behavior"- neural grafting. These and similar neuroscience research experiments with NHPs present unique concerns. The controversy is not confined to the scientific community but extends to the wider public. But the fact that significant neurological similarities between humans and NHPs raise ethical concerns that needs attention and some agreement by many stakeholders (Conlee and Rowan, 2012; Carvalho et al., 2018). In essence, the unique usefulness of NHPs neuroscience research also shapes the unique ethical and legal questions they raise. Overall, this increases the imperative for animal data governance in neuroscience.

3. Current international data governance ecosystem

Data governance is defined as "the principles, procedures, frameworks, and policies that ensure acceptable and responsible processing of data at each stage of the data life cycle, from collection, storage, processing, curation, sharing, and use to deletion" (Eke D. O. et al., 2022). The emphasis on the data life cycle demonstrates that data governance is not only required for a specific stage of the life-cycle. It is a robust framework that starts before data collection and continues to the deletion stage. Fothergill et al. (2019) described it as the overall management of the availability, usability, integrity, quality, and security of data in order to ensure that the potential of the data is maximized while regulatory and ethical compliance is achieved within a specific organizational context. This definition introduces ethical compliance as an important aspect of data governance. Data governance is therefore more than legal compliance (Eke D. et al., 2022). It includes adherence to available ethical principles.

Furthermore, whereas the interpretations of data governance in organizations, disciplines and projects are different (Stahl et al., 2018), its goals and objectives are rooted in available laws and ethics. The question of whose laws and ethical values is determined by the context. Available regulations and ethical values are still jurisdictionally constrained while data continues to cross borders and socio-cultural contexts. For instance, data protection laws are established for specific jurisdictions (e.g., EU GDPR, USA HIPAA[2], Canada's PIPEDA[3] etc.). Ethical values and principles that shape data governance also emerge from specific socio-cultural backgrounds. This means that the meaning or interpretations of data ethics principles such as trust, autonomy, privacy and consent are different in different cultures and societies. These inform relative interpretations of data governance in different cultures.

It is also important to note that established data related regulations and ethical narratives focus mainly on human data. The literature and practice of data ethics and data protection exists to address issues that affect humans in the data processing pipelines. To the best of our knowledge, there is no existing regulation established to address ethical, legal and societal issues related to animal data. This is true for both research and non-research settings. In their systematic literature review of ethical principles that shape data governance discourse in neuroscience, (Ochang et al., 2022) identified a number of ethical principles that often shape data governance discourse in brain research. None of the principles identified touched on animal data concerns. That means that data governance mechanisms often exclude animal data concerns as it relates to ethics and the law. Aspects of technical elements of animal data governance are, however, often included in discussions on findable, accessible, interoperable, reusable (FAIR) data principles. These include aspects of data standardization, integration and interoperability. This falls short of addressing ethical and legal concerns related to the different stages of animal data lifecycle while animal experiments continue to be critical parts of biomedical research.

4. Emerging interests in animal data governance in neuroscience

For a number of reasons, including socio-cultural, ethical and legal differences that inform what is considered permissible use of animals in research, there is a growing interest in the governance of animal data (Eke D. O. et al., 2022). Perceptions on ethical concerns about the use of animals, particularly NHPs, in biomedical research are fundamentally shaped by diverse socio-cultural norms, ethical principles and regulatory requirements. This is evident in the fact that while the use of NHPs in research has decreased

2 https://www.hhs.gov/hipaa/index.html

3 https://www.priv.gc.ca/en/privacy-topics/privacy-laws-in-canada/the-personal-information-protection-and-electronic-documents-act-pipeda/

significantly in Europe and America due to increased social animal welfare activism, NHPs are still the central focus of big neuroscience projects in Asia (Okano et al., 2016; Poo et al., 2016).

Advancements in genome-editing technologies such as the CRISPR/Cas system and artificial intelligence amplifies the huge possibilities in generating genetically modified NHPs. Although this emerging field (genetically modified NHP) has the "potential to transform the study of higher brain function and dramatically facilitate the development of effective treatment for human brain disorders" (Feng et al., 2020), it raises significant ethical concerns. Responding to the huge potential of genetically modified NHPs, Rommelfanger et al. (2018) raise the concern that researchers may be able to introduce cognitive capabilities that can contribute to blurring the boundaries of personhood and ultimately alter traditional perceptions of animal ethics. In 2019, a group of researchers from the Kunming Institute of Zoology in China claimed to have created transgenic monkeys with improved cognitive capacity (Shi et al., 2019). These modified monkeys were created with human MCPH1 genes and were not modeling any human diseases. They were simply modified to be phenotypically humanlike. During a series of cognitive tests, the researchers reported that these animals displayed better short-term memory than their counterparts in the wild. Basically, their brain development mirrored human brain development in many respects. The underlying logic behind this research, which is to manipulate monkeys to model humanlike capacities, presents a slippery slope concern. The question is where would the line be drawn in the path to generate human-like animals? To say the least, such research will not be permissible in many socio-cultural, ethical and legal contexts where the moral status of NHPs are hotly debated.

Given these differences in attitudes, values, principles, beliefs, regulations on the use of animals in research, Rommelfanger et al. (2018) further noted that, "sharing of brain data between countries that hold different ethical stances on what is considered appropriate animal experimentation raises additional questions." One of such questions is; "Should a country accept or use data collected elsewhere in a fashion that is not considered locally ethical?" (Ibid). This is a critical question at the heart of animal data governance consideration. It presents an ethical dilemma many scientists currently face in the neuroscience data sharing ecosystem (Eke D. O. et al., 2022). One project that has stated this as a concern is the PRIMatE Data Exchange (PRIME-DE) Consortium that has highlighted the lack of international standards and regulations as barriers to fostering international NHP data sharing and collaborations (Milham et al., 2020). This increased interest in neuroscience deserves more exploration and hence this paper.

5. Methodology

The issue of responsible animal data governance requires multi-stakeholder perspectives and insights (Rommelfanger et al., 2018; Eke D. et al., 2022). It calls for the appreciation of diverse cultural values and beliefs while respecting established ethical frameworks. Thus, a semi-structured interview was selected as the methodological choice. The underlying research philosophy, therefore, is to use social actors to provide in-depth and rich perspectives on the reality of animal data governance. The aim was to provide detailed and reasoned insights rather than objectively generalisable positions.

The target population included researchers around the globe who have conducted or are conducting animal experiments to answer diverse neuroscience questions. Participants were drawn from active research projects under the International Brain Initiative (IBI). The IBI is the umbrella body for all the large-scale brain research initiatives including the EU Human Brain Project, the US Brain Initiative, Japan Brain/MINDS, Australian Brain Alliance, Korean Brain Initiative, Canadian Brain Research Strategy, and China Brain Project. We also drew participants from Africa and Latin America. Initial list of 37 researchers was compiled and interview invitations extended to all of them. A total of 15 people accepted the invitation and 3 later withdrew citing busy schedules. A structured interview guide that aligns well with the principles of qualitative research design (Ragin and Amoroso, 2011) was developed and tested on two colleagues. Following feedback from these initial tests, the interview protocol was improved before the start of the interviews.

Ethics approval was obtained for this research from De Montfort University ethics review committee. Information sheet and an informed consent form were then emailed to participants. The information sheet contained comprehensive data on the research including but not limited to the research objectives, expectations from the participants and responsibilities of researchers, potential risks and benefits, the voluntariness of participation and how research data will be used. All participants returned the consent form before the interviews and further verbal consent was sought at the start of the interview to record the session. The interviews were conducted either in person or virtually via Skype or zoom and took approximately 40 min each to complete. These interviews occurred between January 2020 and November 2021. Overall, 13 interviews with 12 participants (10 males and 2 females) were conducted. One participant was interviewed twice because the first interview was cut short due to technical difficulties. A total of 13 interviews were considered sufficient to achieve saturation because according to Guest et al. (2006), theoretical saturation can be achieved even in 12 interviews and basic elements for metathemes can emerge as early as six interviews.

The interviews were transcribed and stored in NviVo 12 where they were inductively coded for themes by the first author (DE). The inductive coding process involved reading through the transcript and identifying common patterns and themes. Thematically, the coding tree included high level nodes that show participants' perspectives on why it is time to consider animal data governance or otherwise (such as cultural differences, legal differences etc.). Specific themes that align with high level themes are coded as sub-nodes. For example, *ethics dumping* was identified as a sub-node under the high-level node of *ethical differences*.

6. Key findings

One of the key results that emerged from the interviews was that all the participants agreed that due to the increasing

transfer of animal data across countries, it is time to consider animal data governance for a number of reasons. The reasons provided by the participants are diverse and include: differences in regulations, socio-cultural and ethical values and beliefs, as well as scientific quality (see Table 1).

6.1. Differences in regulations

A number of the participants pointed out legal differences as one of the major reasons why it is time to consider the governance of animal data. For instance, one participant stated that: "*although there is no regulation related to the use of animal data that I know of, there should be one since there are differences in regulations for the generation of animal data*"- (P2).

Another participant also confirmed that: "*Regulations inside laboratories around the world are not the same, why should the data we generate be treated the same way?*" (P1).

Furthermore, another participant presented these regulatory differences with an example of how it prevents collaborations: "*I have colleagues in the UK and we discussed collaborative projects but it turned out to be impossible because the Ethical Research Council's standards do not match Japanese standards. On another occasion, how to use data obtained from Japan, because it was monkey data, was the major problem*" (P5).

To reiterate this angle of NHPs, one participant stated that: "*It is not just that Africa has the animals or the research is cheaper here but it is because that they are allowed to do in some African countries, especially with NHPs, they are not legally allowed to do in their own countries*" (P11).

The mention of non-human primate (NHP) here is critical because as pointed out above, differences in animal welfare are more amplified when it involves NHPs. These legal differences lead to an ethical question of how to share and use data generated from experiments legally not permissible in certain regions of the world.

Some participants raised further issues and concerns related to power imbalance that may emerge due to non-harmonized regulations. One participant observed that: "*If one country has a very loose standard, this country is capable of doing many things that could not be done in some other places. This group will dominate its power in the science of new data maybe. I think this is happening now. If there cannot be harmonized welfare regulations due to many factors, then we should have some open-minded policy discussions on responsible sharing of the datasets*" (P1).

This observation shows that differences in regulation are giving researchers in certain countries (with less strict provisions) scientific advantage over others. In responding to this, another participant stated that: "*We ran into that with stem cells, in the United States. And the argument was, 'well, if you don't fund it, they're going to fund it over there, and we're going to fall behind.' but you cannot build your industries on the back of something that is ethically wrong*" (P3).

This was an important point since regulations are fundamentally shaped by societal values which are different in different regions. These findings align with evidence that has been demonstrated in literature. Vasbinder and Locke (2016) provided an overview of regulatory frameworks across the globe that demonstrated that whereas there are common standards across different jurisdictions, there are clear differences in how different countries regulate the use of animals in research. Most

TABLE 1 Summary of key findings.

Reasons why it is time to consider animal data governance	
Differences in regulations	• Animal welfare is regulated differently in different countries and regions. • While NHPs remain the most suitable animal for research in Neuroscience, their use in research are legally restricted in many countries. • There are no identifiable existing laws in these countries (with restrictive laws) against the use of animal data. *However, in countries where animal experiments are legally not allowed, it should not be legal to use data that emerge from such experiments conducted in other countries.*
Differences in socio-cultural values, beliefs, attitudes, and ethics	• Socio-cultural values, belief systems, attitudes and ethics shape regulations on the use of animals in research. • Whereas some cultures allow and even encourage the use of animals in research, other cultures actively work against the use of animals in research. • Strict animal welfare regulations and social activism against the use of animals in research are pushing researchers to outsource ethically questionable experiments to countries where there are little or no laws–ethics dumping. *Animal data governance will help to prevent ethics dumping and ensure that socio-cultural values and ethics are adhered to.*
Data quality concerns	• Good science requires good data from reliable experiments. • The quality of animal data depends on good animal care (overly stressed animals are bad subjects). *Animal data governance will help to ensure good animal care.*
Approaches to animal data governance	
• Requires a multi-stakeholder dialog for the co-creation of actionable frameworks (involving people from different cultures and animal). • Comparative study of existing regulations and the socio-cultural values and beliefs underlying them. • Can build on established minimum standards used for journal publications. • Requires regulations similar to human data regulations. • International research infrastructures where researchers meet data from different regions of the world will play in vital role in achieving effective animal data governance for research and innovation.	

importantly, they identified that animal research is "performed in African and Middle Eastern countries, but many of these countries have not yet enacted legislation nor established regulatory oversight, policies or guidelines" (Vasbinder and Locke, 2016 p. 263). Mitchell et al. (2021) have also provided detailed insights on the differences in regulatory provisions guiding the use of primates in neuroscience across countries which is the focus of this research. Some of the differences they pointed out include but are not limited to; how NHPs should be generated for use in research (e.g., the use of wild NHPs for research purposes are banned in the EU, in China prior to COVID-19 pandemic this was allowed). Another difference they pointed out is the sizing and flooring of the home enclosures and caging. The EU provides that NHPs should be housed in larger sized enclosures and cages while in China and other Asian countries the cages are smaller (Mitchell et al., 2021). These differences have implications such as forming barriers to effective international collaborations and global data sharing (Rommelfanger et al., 2018; Milham et al., 2020).

6.2. Differences in ethical principles, values and beliefs

It has been established in literature that the diverse socio-cultural values, beliefs, attitudes and ethics found in many regions of the world greatly influence the available diverse animal welfare regulations (Masiga and Munyua, 2005; Szucs et al., 2012; Garcia and McGlone, 2022). The participants highlighted the conflicting cultural and religious beliefs that make animal data governance an imperative. As one participant put it:

> "*I think policy of the animal ethics and welfare differ between countries. . .there are conflicting concepts and beliefs. . . and there should be respect for people's cultures*" (P6).

Another observation was;

> "*Our cultures are different. In our culture we believe that using animals for a scientific purpose is exactly the same as killing animals to eat to maintain our lives because the scientific research leads to the development of human medicine. That is culturally accepted. However, it is different in Europe*" (P5).

These opinions suggest that the acceptability of generation of animal research data is different in different societal contexts. There are differences in the understanding and conceptualization of ethical concerns associated with animal experiment. One of these concerns involves the idea of inflicting *pain* on animals. One participant stated that when scientists are inducing psychiatric disorders, they are creating suffering for the animals which is something that should be deeply examined. Reacting to pain associated with inducing psychiatric disorders in animals, a participant observed that; "*there are ethical concerns, and I think there should be a very deep and profound discussion about it*" (P7).

However, there was an opinion to always focus on the balance between costs and benefits of research experiments that cause pain. For this participant, "*cost and benefit balance is the most critical issue. . . such a kind of disease model can cause some painful situations in these animals and even this can also be an experiment for the chronic pain you know*" (P4).

Furthermore, the idea of pain and how it affects animals is amplified when NHPs are involved.

> *The general belief that monkeys are better animals for experiments because they are closer to humans in cognitive abilities suggests that they will feel more pain than rats and others. This is what we have come to know. . .I don't believe researchers in so many other places respect this. We know too much about primates, we know too much about their sociality, we know too much of this for us to use them as our own personal lab rats* (P8).

This is because NHPs are *sophisticated, capable entities* (P1). Therefore, we should be thinking about *minimizing suffering* (P1) rather than inflicting pain on them.

There are also heightened ethical concerns when transgenic NHPs are involved. According to a participant; *Technology has advanced and many tools are now available. You don't want to create monster-killing animals that can be used for military purposes and things like that* (P9).

Similarly, one participant observed that advancement in technologies are improving research but may be used in unethical ways. *We do have the tools in our hands to start playing dangerous games. That is definitely the case, not only in non-human primates; even in humans. There are possibilities of creating subjects that we don't want on earth. These experiments may be culturally allowed in some places and rejected in others. Governance of how the datasets that emerge from such dangerous experiments can deter such research* (P3).

For another participant responding to whether it is ethical to use data from transgenic NHPs, conditions under which such data is generated should be carefully examined because; *there are real concerns, especially as you get into transgenic models, to really consider whether or not you want to be seen endorsing a particular approach or not. That depends on where the researcher is working from* (P4).

To support this line of thought, another participant stated that: "*...this is a serious ethical concern. The use of the data from research that is not allowed in this country* (the participant's country) *can cause reputational damage. I have not thought about it this way but it can*" (P10).

Another argument presented by the participants was that the differences in ethical principles, values and regulations means that some researchers will move to other countries with less restrictive values and regulations or in some cases outsource the research. These are insights provided by the participants:

> *We already know that happens. We knew it happened with stem cells, we knew it happened with anything else, "...can't do it here, I*

> *will go and do it in that country, where there are no regulations." Some researchers can do the research in other countries and then bring back the data to be used here. This is data laundering and it is fraud, right? You are just outsourcing it to somebody else and taking advantage of the data* (P3).

> *To be fairly honest, I think at some scale that's already happening. There are people from Europe doing some kind of research in China for Example, which they would have a very difficult time to get easily approved in their own country. The same is true for the United States. So, in theory, yes, this opens the doors for that kind of stuff, but that's not where we want to go* (P1).

Another participant observed that animal data governance can *Prevent unnecessary animal research happening in Africa by researchers from Europe and America. So many of these do not follow the same ethical principles they are made to follow in their own countries* (P11).

These perspectives hint at an ethical concern often referred to as ethics dumping which will be explored further in sections below. It also means that while researchers in different regions of the world are required to comply with their regulations, some of these regulations may fall short of acceptable ethical standards in other regions.

6.3. Data quality concerns

Another reason the participants believe that it is time for animal data governance is to ensure that the quality of animal data shared is good for scientific purposes. The summary of the opinions here is that data governance can improve the quality of data because good scientific research relies upon high quality laboratory animal care (Friese, 2013, 2019). A harmonized animal data governance can help to improve animal welfare in a way that fewer confounding variables are introduced into research. Some of these opinions are as follows:

> *But we also know that overly-stressed animals are bad subjects, you don't get good data from them, you see mixed effects. So, I think you could make both the ethical and the scientific case that these animals need to be treated well. Governance can help harmonize best practices for animal welfare* (P4).

> *It will be a con-founding factor. If there are high stress levels on the animals, it will simply provide you with completely wrong measurements of whatever you are doing. Yes, that is what I think. Having some sort of universal principle for the derived data will improve the quality of data shared* (P12).

> *The quality of data is critical to this discussion. I think in science you want a certain degree of consistency and quality for effective research outcomes. Open data portals need to put mechanisms to ensure that the data they make available come from labs that comply with high standards of animal welfare* (P9).

These views suggest that animal data governance when implemented, particularly by open data repositories/archives, can help to improve the quality of animal data for research. However, without adequate governance mechanisms, low quality animal data may be allowed to permeate within research ecosystems.

6.4. Approaches to governance

Beyond providing insights into the reasons why it is time to consider animal data governance, the participants also gave their perspectives on how this can be done. One view that was shared by all the participants is that a harmonized governance framework for animal data requires inclusive discussions involving all relevant stakeholders. In order to appreciate and respect differences in regulations and social-cultural values, animal data governance is a multi-stakeholder endeavor. For instance, one view was that:

> *Discussions on governance approaches should be thoughtful and careful and involve researchers that are performing these experiments. It should also involve people from different cultures to understand what is an appropriate framework* (P12).

This is a very important view that can ensure that one region does not dictate for other regions as regards best practices. As a participant mentioned; *responsible data governance is an interesting concept but a complex one. Whose understanding of responsibility? Whose values are going to shape this? These are things that need further discussions and understanding* (P5).

Another participant also observed that exclusion or disregard of some cultural values and beliefs will make any developed mechanism unacceptable for many scientists. However, this does not mean that people working in regions with strict regulations should accept anything and everything.

There was also the feeling that already established standards used for publishing in journals can be a starting point–something to build on.

> *Minimum standards of acceptance and rejection for papers in neuroscience journals need to be studied and improved upon. For instance, the implementation of the 3Rs. Maybe a comparative literature about the different ethical standards used in different countries* (P2).

Another researcher further suggested that *auditing all data producing sites* can be a pragmatic way of understanding the *status quo*.

One critical view that was shared by many of the participants is that research infrastructures or open data archives, repositories or portals need to play important roles in providing efficient governance mechanisms for animal data. The argument was that animal data should be *governed through international research*

infrastructures where researchers and research data from different cultures meet (P10).

Some participants made a case for the establishment of regulations for the use of animal data similar to data protection laws. These can be new regulations or amendments to existing laws to cover the use or application of animal data, particularly NHP data. An EU participant puts it thus:

> *The rationale is, of course, that we wouldn't include data that has been acquired in a way that doesn't conform to the rules within the EU. That is important. Otherwise, it serves as an incentive, almost, for people to do something elsewhere and then hopefully get the data in* (P7).

7. Discussion and conclusion

The findings from our empirical work are broadly consistent with our insights from the existing literature. The international neuroscientists we talked to agreed that animal neuroscience research raises ethical concerns and that these concerns have consequences for the way the resulting data can and should be used. The very brief answer to the question in the title to this article whether it is time to consider animal data in data governance is thus a "yes."

Our empirical work highlights that this is not a purely theoretical problem but that questions of animal data governance do arise in practical neuroscience work. The global discussion of data governance in health-related research ostensibly has the purpose of facilitating collaboration and exchange of data with a view to support the creation of new knowledge and the resulting consequences. This logic can be extended to animal data which calls for animal data governance.

There remain, however, fundamental differences between human data and animal data. The very use of human data can raise ethical concerns, for example where patient record privacy is concerned. Animal data does not raise such intrinsic concerns. For animal data the core of most issues is the generation or collection of data. Data use is nevertheless important because a lack of attention to the use of data may facilitate the use of data which is deemed not to be ethically permissible in a particular jurisdiction or cultural context, thereby sidestepping agreed-upon ethical principles.

A key question is therefore which type of animal research is deemed to be permissible and on what grounds can such value judgments be made. Our interviews showed that neuroscientists are aware of differences with regards to these questions. There seems to be a continuum of ethical severity which starts with cell cultures, moves up via invertebrates, vertebrates, rodents, NHPs and may find its current summit in research on transgenic NHPs. The problem is that the evaluation of these different types of research differ between cultures and jurisdictions and there is no universally agreed position on these questions.

This lack of agreement points to the lively exchange of ideas between cultures which is probably a good thing. Ethics is a topic that often finds its expressions in dilemmas and disagreements, so the plurality of views is not surprising. It seems plausible that an ethical free-for-all is not desirable, neither in animal research, nor in research more generally. At the same time, a uniform ethical position would suppress legitimate positions and thus be likely unethical in itself. Ethical plurality is thus welcome and can stimulate academic debate, for example in the field of neuroethics, where questions are continually triggered by new technologies and methods. While we thus welcome ethical plurality on animal research, it raises practical questions with regards to what data can and should be used for which purposes.

This leaves us with the question of how ethical pluralism can be accommodated in data governance. One plausible response to this question could come from procedural approaches to ethics. What this means is that we should not expect to find agreement on the substance of complex ethical questions, but it may be possible to define procedures that support fruitful engagement on these questions. One can argue that most modern western approaches to philosophical ethics follow such a procedural approach. Elsewhere we have suggested that Discourse Ethics may provide a suitable avenue to pursue (Stahl et al., 2019). We believe, without having the space to make this argument in detail, that such a procedural approach could be applied to the problem of animal data governance.

In practice this would mean that data governance should be designed in a way to facilitate constructive debate about ethical issues and be suitable for supporting ethical consensus, where it exists. This implies that data governance should be used to highlight ethically contentious aspects of animal data. This means that the meta-data of neuroscientific animal research data should clearly show those aspects that we know to be ethically contentious. This would include the species, research question, type of intervention, whether transgenic animals are involved etc.

The result of such an approach to animal data would be that it would be easy to understand ethical agreements and disagreements. If, for example, a culture has a consensus view that *in vivo* experiments with NHPs or the use of transgenic animals is not ethically acceptable, then filtering out such data would be easy. Probably more importantly, a strong metadata schema would allow highlighting which aspects of the research and the resulting data may be contentious and thus foster communication around the reasons for disagreements and possible ways to shape research and data in ways that are more broadly deemed to be acceptable.

This proposal is of course not overwhelmingly novel. Data governance structures already routinely capture some of these items. As our interviews have shown, researchers see data quality as an (ethical) issue and good metadata is recognized as a means to increase transparency of data and to promote the FAIR principles. The novelty of our proposal is that ethics-related aspects of the metadata could be explicitly defined and collected. While many of these will already form part of data governance, a next step would be to more clearly define them in ways that support researchers who generate the scientific data in the first place and ensure that they are aware of relevant metadata requirements.

This proposal is of course no panacea. There are limitations of our research such as the limited number of respondents and a lack of statistical representativeness of our approach. While we believe that our methodological choices have ensured that we received relevant input, we cannot claim to have represented all possible angles and identified all ethical issues. This article should thus be read as an exploratory study that has confirmed that ethically informed animal data governance is called for. The real work of

shaping such a data governance approach will have to follow as a large-scale consultative exercise leading to the co-creation of animal data governance that truly captures ethical issues. This future research can include how these findings individually can or do shape animal data governance.

It is furthermore worth underlining that the existence of ethically sensitive data governance will not make the underlying ethical issues go away. Many of these issues touch on deep convictions of who we are and what we as humans can or should do. Such convictions do not change quickly and different views will remain. But, as indicated earlier, ethical pluralism does not need to be seen as a problem but can be celebrated as an expression of human diversity. What counts is that we find productive ways of dealing with issues. Ethically informed animal data governance can be one mechanism that allows us as researchers, as citizens of different countries, holders of different convictions to come together and have productive conversations about how to understand and deal with our different worldviews. If it achieves this, then this will not only strengthen neuroscience with all its concomitant benefits but also show how science can play an important role in tackling the broader ethical and social questions that our increasingly globalized world faces.

Data availability statement

The original contributions presented in this study are included in the article/supplementary material, further inquiries can be directed to the corresponding author.

Ethics statement

Ethics approval was obtained for this research from De Montfort University Ethics Review Committee. Information sheet and an informed consent form were then emailed to participants.

Author contributions

DE: conceptualization, writing, data analysis, original draft/review, and editing. GO: conceptualization, writing, and data collection. WK: conceptualization and writing. BS: conceptualization, writing, and funding. All authors contributed to the article and approved the submitted version.

Funding

This work was supported by the European Union's Horizon 2020 Framework Programme for Research and Innovation under the Specific Grant Agreements No. 720270 (HBP SGA1), No. 785907 (HBP SGA2), and No. 945539 (HBP SGA3).

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

References

Bailey, J., and Taylor, K. (2016). Non-human primates in neuroscience research: The case against its scientific necessity. *Altern. Lab. Anim.* 44, 43–69. doi: 10.1177/026119291604400101

Bjugstad, K. B., and Sladek, J. R. (2006). "Neural transplantation in the nonhuman primate model of Parkinson's disease," in *Cell therapy, stem cells, and brain repair*, eds C. D. Sanberg and P. R. Sanberg (Berlin: Springer), 61–82. doi: 10.1007/978-1-59745-147-5_3

Bystron, I., Rakic, P., Molnár, Z., and Blakemore, C. (2006). The first neurons of the human cerebral cortex. *Nat. Neurosci.* 9, 880–886. doi: 10.1038/nn1726

Capitanio, J. P., and Emborg, M. E. (2008). Contributions of non-human primates to neuroscience research. *Lancet* 371, 1126–1135. doi: 10.1016/S0140-6736(08)60489-4

Carvalho, C., Gaspar, A., Knight, A., and Vicente, L. (2018). Ethical and scientific pitfalls concerning laboratory research with non-human primates, and possible solutions. *Animals* 9:12. doi: 10.3390/ani9010012

Chan, A. W. (2014). "Production of transgenic nonhuman primates," in *Transgenic animal technology*, ed. C. A. Pinkert (Amsterdam: Elsevier), 359–385. doi: 10.1016/B978-0-12-410490-7.00014-1

Chimpanzee Sequencing and Analysis Consortium (2005). Initial sequence of the chimpanzee genome and comparison with the human genome. *Nature* 437, 69–87. doi: 10.1038/nature04072

Conlee, K. M., and Rowan, A. N. (2012). The case for phasing out experiments on primates. *Hastings Center Rep.* 42, S31–S34. doi: 10.1002/hast.106

Cyranoski, D. (2016). Monkey kingdom. *Nature* 532, 300–302. doi: 10.1038/532300a

Disotell, T. R., and Tosi, A. J. (2007). The monkey's perspective. *Genome biol.* 8, 1–4. doi: 10.1186/gb-2007-8-9-226

Eke, D., Ochang, P., Ogundele, T., Adimula, A., Borokini, F., and Akintoye, S. (2022). *Responsible Data Governance in Africa: Institutional gaps and capacity needs*. Abuja: Centre for the Study of African Economies (CSEA).

Eke, D. O., Bernard, A., Bjaalie, J., Chavarriaga, R., Hanakawa, T., Hannan, A. J., et al. (2022). International data governance for neuroscience. *Neuron* 110, 600–612. doi: 10.1016/j.neuron.2021.11.017

European Commission (2019). *EU statistical reports on the use of animals for scientific purposes*. Brussels: European Commission.

Feng, G., Jensen, F., Greely, H., Okano, H., Treue, S., Roberts, A., et al. (2020). Opportunities and limitations of genetically modified nonhuman primate models for neuroscience research. *Proc. Natl. Acad. Sci. U.S.A.* 117, 24022–24031. doi: 10.1073/pnas.2006515117

Fothergill, B. T., Fothergill, B., Knight, W., Stahl, B., and Ulnicane, I. (2019). Responsible data governance of neuroscience big data. *Front. Neuroinform.* 13:28. doi: 10.3389/fninf.2019.00028

Friedman, H., Ator, N., Haigwood, N., Newsome, W., Allan, J., Golos, T., et al. (2017). The critical role of nonhuman primates in medical research. *Pathogens Immunity* 2, 352–365. doi: 10.20411/pai.v2i3.186

Friese, C. (2013). Realizing potential in translational medicine: The uncanny emergence of care as science. *Curr. Anthropol.* 54, S129–S138.

Friese, C. (2019). Intimate entanglements in the animal house: Caring for and about mice. *Sociol. Rev.* 67, 287–298.

Garcia, A., and McGlone, J. J. (2022). Animal welfare and the acknowledgment of cultural differences. *Animals* 12:474.

Garner, J. P. (2014). The significance of meaning: Why do over 90% of behavioral neuroscience results fail to translate to humans, and what can we do to fix it? *ILAR J.* 55, 438–456. doi: 10.1093/ilar/ilu047

Guest, G., Bunce, A., and Johnson, L. (2006). How many interviews are enough?: An experiment with data saturation and variability. *Field Methods* 18, 59–82. doi: 10.1177/1525822X05279903

Guhad, F. (2005). Introduction to the 3Rs (Refinement, Reduction and Replacement). *J. Am. Assoc. Lab. Anim. Sci.* 44, 58–59.

Haesemeyer, M., and Schier, A. F. (2015). The study of psychiatric disease genes and drugs in zebrafish. *Curr. Opin. Neurobiol.* 30, 122–130. doi: 10.1016/j.conb.2014.12.002

Jones, B. (2021). *Feature: Why do we need to use animals in neuroscience research?* Belek: EARA.

Kaiser, T., and Feng, G. (2015). Modeling psychiatric disorders for developing effective treatments. *Nat. Med.* 21, 979–988. doi: 10.1038/nm.3935

Knight, A. (2007). The poor contribution of chimpanzee experiments to biomedical progress. *J. Appl. Anim. Welfare Sci.* 10, 281–308. doi: 10.1080/10888700701555501

Lankau, E. W., Turner, P., Mullan, R., and Galland, G. (2014). Use of nonhuman primates in research in North America. *J. Am. Assoc. Lab. Anim. Sci.* 53, 278–282.

Li, W.-H., and Saunders, M. A. (2005). The chimpanzee and us. *Nature* 437, 50–51.

Masiga, W. N., and Munyua, S. J. M. (2005). Global perspectives on animal welfare: Africa. *Rev. Sci. Tech. Off. Int. Des Épiz.* 24:579.

Mepham, T. B., Combes, R., Balls, M., Barbieri, O., Blokhuis, H., Costa, P., et al. (1998). The use of transgenic animals in the European Union: The report and recommendations of ECVAM workshop 28. *Altern. Lab. Anim.* 26, 21–43. doi: 10.1177/026119299802600108

Milham, M., Petkov, C. I., Margulies, D. S., Schroeder, C. E., Basso, M. A., Belin, P., et al. (2020). Accelerating the evolution of nonhuman primate neuroimaging. *Neuron* 105, 600–603. doi: 10.1016/j.neuron.2019.12.023

Mitchell, A. S., Hartig, R., Basso, M., Jarrett, W., Kastner, S., and Poirier, C. (2021). International primate neuroscience research regulation, public engagement and transparency opportunities. *Neuroimage* 229:117700. doi: 10.1016/j.neuroimage.2020.117700

Neuhaus, C. P. (2018). Ethical issues when modelling brain disorders in non-human primates. *J. Med. Ethics* 44, 323–327. doi: 10.1136/medethics-2016-104088

Ochang, P., Stahl, B. C., and Eke, D. (2022). The ethical and legal landscape of brain data governance. *PLoS One* 17:e0273473. doi: 10.1371/journal.pone.0273473

Okano, H., Sasaki, E., Yamamori, T., Iriki, A., Shimogori, T., Yamaguchi, Y., et al. (2016). Brain/MINDS: A Japanese national brain project for marmoset neuroscience. *Neuron* 92, 582–590. doi: 10.1016/j.neuron.2016.10.018

Poldrack, R. A., and Gorgolewski, K. J. (2014). Making big data open: Data sharing in neuroimaging. *Nat. Neurosci.* 17, 1510–1517. doi: 10.1038/nn.3818

Poo, M., Du, J., Ip, N., Xiong, Z., Xu, B., and Tan, T. (2016). China brain project: Basic neuroscience, brain diseases, and brain-inspired computing. *Neuron* 92, 591–596. doi: 10.1016/j.neuron.2016.10.050

Preuss, T. M. (2010). *Reinventing Primate Neuroscience for the Twenty-First Century*. Oxford: Oxford University Press. doi: 10.1093/acprof:oso/9780195326598.003.0022

Qiu, Z., and Li, X. (2017). Non-human primate models for brain disorders – Towards genetic manipulations via innovative technology. *Neurosci. Bull.* 33, 247–250. doi: 10.1007/s12264-017-0115-4

Ragin, C. C., and Amoroso, L. M. (2011). *Constructing social research: The unity and diversity of method*. Thousand Oaks, CA: Pine Forge Press.

Rommelfanger, K. S., Jeong, S., Ema, A., Fukushi, T., Kasai, K., Ramos, K., et al. (2018). Neuroethics questions to guide ethical research in the international brain initiatives. *Neuron* 100, 19–36. doi: 10.1016/j.neuron.2018.09.021

Russell, W. M. S., and Burch, R. L. (1960). The principles of humane experimental technique. *Med. J. Austr.* 1, 500–500. doi: 10.5694/j.1326-5377.1960.tb73127.x

Scientific Committee on Health, Environmental and Emerging Risks [SCHEER] (2017). *Final opinion on "The need for non-human primates in biomedical research, production and testing of products and devises"*. Saarbrücken: SCHEER.

Shi, L., Luo, X., Jiang, J., Chen, Y., Liu, C., Hu, T., et al. (2019). Transgenic rhesus monkeys carrying the human MCPH1 gene copies show human-like neoteny of brain development. *Natl. Sci. Rev.* 6, 480–493. doi: 10.1093/nsr/nwz043

Sikela, J. M. (2006). The jewels of our genome: The search for the genomic changes underlying the evolutionarily unique capacities of the human brain. *PLoS Genet.* 2:e80. doi: 10.1371/journal.pgen.0020080

Stahl, B. C., Akintoye, S., Fothergill, B., Guerrero, M., Knight, W., and Ulnicane, I. (2019). Beyond research ethics: Dialogues in Neuro-ICT research. *Front. Hum. Neurosci.* 13:105. doi: 10.3389/fnhum.2019.00105

Stahl, B. C., Rainey, S., Harris, E., and Fothergill, B. T. (2018). The role of ethics in data governance of large neuro-ICT projects. *J. Am. Med. Inform. Assoc.* 25, 1099–1107. doi: 10.1093/jamia/ocy040

Szucs, E., Geers, R., Jezierski, T., Sossidou, E. N., and Broom, D. M. (2012). Animal welfare in different human cultures, traditions and religious faiths. *Asian Aust. J. Anim. Sci.* 25, 1499–1506.

Tardif, S. D., Coleman, K., Hobbs, T., and Lutz, C. (2013). IACUC Review of Nonhuman Primate Research. *ILAR J.* 54, 234–245. doi: 10.1093/ilar/ilt040

Ting, J. T., and Feng, G. (2013). Development of transgenic animals for optogenetic manipulation of mammalian nervous system function: Progress and prospects for behavioral neuroscience. *Behav. Brain Res.* 255, 3–18. doi: 10.1016/j.bbr.2013.02.037

Tomioka, I., Nogami, N., Nakatani, T., Owari, K., Fujita, N., Motohashi, H., et al. (2017). Generation of transgenic marmosets using a tetracyclin-inducible transgene expression system as a neurodegenerative disease model. *Biol. Reprod.* 97, 772–780. doi: 10.1093/biolre/iox129

Varki, A. (2000). A chimpanzee genome project is a biomedical imperative. *Genome Res.* 10, 1065–1070. doi: 10.1101/gr.10.8.1065

Vasbinder, M. A., and Locke, P. (2016). Introduction: Global Laws, Regulations, and Standards for Animals in Research. *ILAR J.* 57, 261–265. doi: 10.1093/ilar/ilw039

Weatheall, D. (2015). *The use of non-human primates in research: A working group report*. London: Royal Society.

Windisch, M. (2014). We can treat Alzheimer's disease successfully in mice but not in men: Failure in translation? A perspective. *Neuro-Degener. Dis.* 13, 147–150. doi: 10.1159/000357568

Yeo, H.-G., Lee, Y., Jeon, C., Jeong, K., Jin, Y., Kang, P., et al. (2015). Characterization of cerebral damage in a monkey model of Alzheimer's disease induced by intracerebroventricular injection of streptozotocin. *J. Alzheimers Dis.* 46, 989–1005. doi: 10.3233/JAD-143222

Yun, H.-M., Choi, D., Oh, K., and Hong, J. (2015). PRDX6 exacerbates dopaminergic neurodegeneration in a MPTP mouse model of Parkinson's disease. *Mol. Neurobiol.* 52, 422–431. doi: 10.1007/s12035-014-8885-4

Zhao, H., Jiang, Y.-H., and Zhang, Y. Q. (2018). Modeling autism in non-human primates: Opportunities and challenges. *Autism Res.* 11, 686–694. doi: 10.1002/aur.1945

Zweier, C., de Jong, E., Zweier, M., Orrico, A., Ousager, L., Collins, A., et al. (2009). CNTNAP2 and NRXN1 are mutated in autosomal-recessive Pitt-Hopkins-like mental retardation and determine the level of a common synaptic protein in Drosophila. *Am. J. Hum. Genet.* 85, 655–666. doi: 10.1016/j.ajhg.2009.10.004

frontiers | Frontiers in Neuroinformatics

TYPE Original Research
PUBLISHED 31 August 2023
DOI 10.3389/fninf.2023.1215261

OPEN ACCESS

EDITED BY
Christian Haselgrove,
UMass Chan Medical School, United States

REVIEWED BY
Karsten Tabelow,
Weierstrass Institute for Applied Analysis and Stochastics (LG), Germany
Baris Evren Ugurcan,
Max Planck Institute for Human Cognitive and Brain Sciences, Germany

*CORRESPONDENCE
Lei Wang
✉ lei.wang@osumc.edu

RECEIVED 01 May 2023
ACCEPTED 01 August 2023
PUBLISHED 31 August 2023

CITATION
Wang L, Ambite JL, Appaji A, Bijsterbosch J, Dockes J, Herrick R, Kogan A, Lander H, Marcus D, Moore SM, Poline J-B, Rajasekar A, Sahoo SS, Turner MD, Wang X, Wang Y and Turner JA (2023) NeuroBridge: a prototype platform for discovery of the long-tail neuroimaging data.
Front. Neuroinform. 17:1215261.
doi: 10.3389/fninf.2023.1215261

NeuroBridge: a prototype platform for discovery of the long-tail neuroimaging data

Lei Wang[1]*, José Luis Ambite[2], Abhishek Appaji[3], Janine Bijsterbosch[4], Jerome Dockes[5], Rick Herrick[4], Alex Kogan[1], Howard Lander[6], Daniel Marcus[4], Stephen M. Moore[4], Jean-Baptiste Poline[5], Arcot Rajasekar[6,7], Satya S. Sahoo[8], Matthew D. Turner[1], Xiaochen Wang[9], Yue Wang[7] and Jessica A. Turner[1]

[1]Psychiatry and Behavioral Health Department, The Ohio State University Wexner Medical Center, Columbus, OH, United States, [2]Information Sciences Institute and Computer Science, University of Southern California, Los Angeles, CA, United States, [3]Department of Medical Electronics Engineering, BMS College of Engineering, Bangalore, India, [4]Department of Radiology, Washington University in St. Louis, St. Louis, MO, United States, [5]Department of Neurology and Neurosurgery, McGill University, Montreal, QC, Canada, [6]Renaissance Computing Institute, Chapel Hill, NC, United States, [7]School of Information and Library Science, University of North Carolina at Chapel Hill, Chapel Hill, NC, United States, [8]Department of Population and Quantitative Health Sciences, Case Western Reserve University, Cleveland, OH, United States, [9]College of Information Sciences and Technology, Pennsylvania State University, State College, PA, United States

Introduction: Open science initiatives have enabled sharing of large amounts of already collected data. However, significant gaps remain regarding how to find appropriate data, including underutilized data that exist in the long tail of science. We demonstrate the NeuroBridge prototype and its ability to search PubMed Central full-text papers for information relevant to neuroimaging data collected from schizophrenia and addiction studies.

Methods: The NeuroBridge architecture contained the following components: (1) Extensible ontology for modeling study metadata: subject population, imaging techniques, and relevant behavioral, cognitive, or clinical data. Details are described in the companion paper in this special issue; (2) A natural-language based document processor that leveraged pre-trained deep-learning models on a small-sample document corpus to establish efficient representations for each article as a collection of machine-recognized ontological terms; (3) Integrated search using ontology-driven similarity to query PubMed Central and NeuroQuery, which provides fMRI activation maps along with PubMed source articles.

Results: The NeuroBridge prototype contains a corpus of 356 papers from 2018 to 2021 describing schizophrenia and addiction neuroimaging studies, of which 186 were annotated with the NeuroBridge ontology. The search portal on the NeuroBridge website https://neurobridges.org/ provides an interactive Query Builder, where the user builds queries by selecting NeuroBridge ontology terms to preserve the ontology tree structure. For each return entry, links to the PubMed abstract as well as to the PMC full-text article, if available, are presented. For each of the returned articles, we provide a list of clinical assessments described in the Section "Methods" of the article. Articles returned from NeuroQuery based on the same search are also presented.

Conclusion: The NeuroBridge prototype combines ontology-based search with natural-language text-mining approaches to demonstrate that papers relevant to a user's research question can be identified. The NeuroBridge prototype takes a first step toward identifying potential neuroimaging data described in full-text papers. Toward the overall goal of discovering "enough data of the right kind," ongoing work includes validating the document processor with a larger corpus, extending the ontology to include detailed imaging data, and extracting information regarding data availability from the returned publications and incorporating XNAT-based neuroimaging databases to enhance data accessibility.

KEYWORDS

addiction, schizophrenia, experimental design, MRI, metadata, ontology, text-mining

Introduction

The unprecedented data revolution has generated an enormous amount of data, including biomedical imaging datasets. In 2022, the NIH funded over 7,000 neuroimaging-related projects, encompassing virtually every institute (National Institutes of Health, National Institutes of Health). Over 6,000 currently open clinical trials rely on imaging as a primary endpoint or other key dependency[1]. Much of the present efforts on reproducibility science are focused on annotation, processing, and to some extent analysis. The new NIH Data Management and Sharing Policy (National Institutes of Health, 2023) is encouraging the sharing of data and has pointed to repositories for depositing data. However, how to find data, and more importantly, how to find sufficient data that is appropriate to answering a specific research question, is currently left to the individual researcher to navigate. The facilitation of finding sufficient data of the right kind is a critical gap.

Currently, much of the data is not yet "findable." While organized, big neuroimaging data is being shared through mechanisms such as searchable archives (see an example list of the many different neuroimaging databases that are sharing data) (Eickhoff et al., 2016), and data are being reported and deposited with recently established resources such as data journals (Walters, 2020) and EuropePMC[2], an even larger number of smaller-sized datasets have been collected in day-to-day research by individual laboratories and reported in peer-reviewed publications: approximately 9,000 full text papers are available at *Frontiers in Psychology* and *Frontiers in Neuroscience* alone, and Neurosynth.org contains 10,000 fMRI papers. Many of these datasets are utilized once and never shared. These underutilized "gray data" along with the rest of the data that remain in the unpublished "darkness" form the "long tail of data" (Wallis et al., 2013; Ferguson et al., 2014). Finding, accessing, and reusing these data could greatly enhance their value and lead to improved reproducibility science.

Searching the scientific literature for data is a labor intensive endeavor. While researchers can search for papers on platforms such as PubMed Central (PMC) and Google Scholar, culling through the returned articles to identify which ones may contain relevant study populations and whether they include references to datasets is time consuming. One coauthor's Ph.D student wished to assess the reliability of automated tracing of the amygdala, and whether manual-vs-automated differences might account for disagreements in the literature. Through obtaining data directly from authors, she was able to definitively demonstrate that amygdala volumes were not a sensitive measure in the population she was researching, and that differences in tracing methodology did not account for the literature disagreements (Jayakar, 2017; Jayakar et al., 2018, 2020). However, this process took 18 months! A more efficient process by which researchers can find relevant data in the literature is needed.

To improve search efficiency, a large body of work has been done to annotate the research literature (Fox et al., 2005; Turner and Laird, 2012). PubMed, for example, tags papers with the Medical Subject Headings (MeSH) terms. In the neuroimaging community, the Neurosynth project has derived keywords and result tables from full text of functional MRI papers. The NeuroQuery platform developed a library of ~7,500 keywords to label fMRI activation coordinates in full text papers on psychiatric studies (Dockes et al., 2020). Many scientific domains, including neuroscience, extensively adopt ontologies to describe observations and organize knowledge (Moreau et al., 2008; Widom, 2008; Sahoo et al., 2019). Using these ontologies to annotate textual descriptions of datasets is therefore a key step toward effective data discovery and selection. Natural language processing (NLP) and machine learning approaches have the potential to automate this process. For example, the Brainmap Tracker used the Cognitive Paradigm Ontology to guide text-mining for tagging papers (Laird et al., 2005; Turner and Laird, 2012; Turner et al., 2013; Chakrabarti et al., 2014). Traditional machine learning algorithms often require training on a large number of annotated examples, where unstructured texts are manually annotated using a complex ontology. This is a labor-intensive process that requires highly specialized domain expertise. We have previously developed a deep-learning classification algorithm that obtained high accuracy without assuming large-scale training data (Wang et al., 2022), by exploiting pre-training deep neural language models on rich semantic knowledge in the ontology.

1 ClinicalTrials.gov: https://www.clinicaltrials.gov/.

2 Europe PubMed Central: https://europepmc.org/.

In this context, we launched the NeuroBridge project to facilitate the discovery and reuse of neuroimaging data described in peer-reviewed publications and searchable databases. It is important to note that while there are efforts on modeling provenance metadata during the design and implementation of studies prior to publication (Keator et al., 2013; Gorgolewski et al., 2016; Kennedy et al., 2019), the NeuroBridge is focused on completed studies that are described in papers.

The NeuroBridge project supports the FAIR data principles (Wilkinson et al., 2016) for improving findability, accessibility, interoperability and reusability of scientific data in the following ways. *Findability*: FAIR recommends that metadata and data should be easy to find. NeuroBridge enhances the findability of data through clinical ontology-based indexing for finding presence of data usage in publications. *Accessibility*: FAIR recommends that a user be given information on how data can be accessed once found. In NeuroBridge we provide the data availability statement and author contact information that we extract automatically from publication metadata. *Interoperability*: FAIR recommends common vocabulary and use of formal, accessible, shared, and broadly applicable language for representation of data and metadata. NeuroBridge provides mappings between metadata terms used by data providers and published studies to metadata schemas that conform to standard terms or ontology. *Reusability*: FAIR recommends data be richly described with a plurality of accurate and relevant metadata attributes. NeuroBridge provides metadata schemas that are annotated with common vocabulary and ontology. We have made all of our relevant data and tools freely available[3,4] to encourage the neuroscience community to produce data that can be legally and efficiently utilized by third party investigators.

Our long-term goal is to bridge the research question with data and scientific workflow, thereby significantly speeding up the cycle of hypothesis-based research. In the companion paper in this special issue, we describe the NeuroBridge ontology (Sahoo et al., 2023). In this paper, we report the NeuroBridge prototype platform that focused on neuroimaging studies of schizophrenia and addiction disorders as application domains. To extract metadata about study design and data collection from full-text papers, we leveraged a number of previous efforts on ontology development and machine-learning based natural-language processing.

The NeuroBridge prototype architecture

The design of the NeuroBridge architecture (**Figure 1**) was guided by our overall goal to find enough data of relevance to the user, and by the principle of identifying relevance by metadata that is harmonized by a common ontology. Within this principle, we first created an extensible NeuroBridge Ontology that was interoperable with other domain-specific terminological systems such as the Systematized Nomenclature of Medicine Clinical Terms (SNOMED CT), the Neuroimaging Data Model (NIDM) ontology (Maumet et al., 2016), and the RadLex ontology developed by the Radiological Society of North America. This ontology was then used to annotate a set of full-text peer-reviewed papers, which was then used to train a natural-language document processor to develop a deep neural network model to represent each paper with the ontological concepts. Finally, a user-friendly interface that contained an interactive query builder and integrated search across disparate data sources completed the prototype architecture.

We first established a document corpus of PMC papers to develop the NeuroBridge ontology and train our deep neural network document processor. The corpus contained 356 full-text articles from 2017 to 2020, available from the National Library of Medicine (NLM) BioC collection, reporting empirical studies of schizophrenia and substance-related disorders that have collected neuroimaging data on human subjects, excluding meta-analysis and review papers. The NLM BioC collection (Comeau et al., 2019) is a simple format designed for straightforward text processing, text mining and information retrieval research, e.g., using plain text or JSON. Details of queries performed on PMC are shown in **Table 1**. Of the 356 articles, 186 were used to annotate with the NeuroBridge ontology and train our deep neural network document processor, described below.

The NeuroBridge ontology

Full details of the ontology and its development process are described in the companion paper in this special issue (Sahoo et al., 2023). The NeuroBridge ontology was developed in the metadata framework called the S3 model that classified provenance metadata related to research studies into the categories of *study instrument*, *study data*, and *study method* (Sahoo et al., 2019), which extended the World Wide Web Consortium (W3C) PROV specification to represent provenance metadata for the biomedical domain. The NeuroBridge ontology was developed to be interoperable in annotating the neuroimaging literature and extensible to model additional study metadata such as subject recruitment and data collection methods. It incorporated our previous work on terminologies for data sharing in schizophrenia (Wang et al., 2016), and extended it to include metadata terms from the ENIGMA Addiction Project (Cao et al., 2021). It systematically and comprehensively modeled metadata information that described neuroscience experiments such as the number of participants in a diagnostic group, the type of experiment data collected (neuroimaging, neurophysiology etc.), and the clinical and cognitive assessment instruments.

The NeuroBridge ontology model included terms for neuroimaging data types for T1-weighted, task-based or resting-state functional imaging, a variety of clinical diagnoses such as neurodevelopmental disorder, mental disorders, and cognitive disorder. It also included various clinical and cognitive assessment instruments such as substance use scales, psychopathology scales, neurocognitive scales and mental health diagnosis scales. The ontology was integrated into the natural language processing pipeline and the NeuroBridge query interface, both described below, to allow use of metadata terms in composing user query expressions and identify relevant study articles.

3 NeuroBridge (Website): https://github.com/NeuroBridge/NeuroBridge1.0.

4 NeuroBridge (Ontology): https://github.com/NeuroBridge/neuro-ontologies/tree/main/neurobridge.

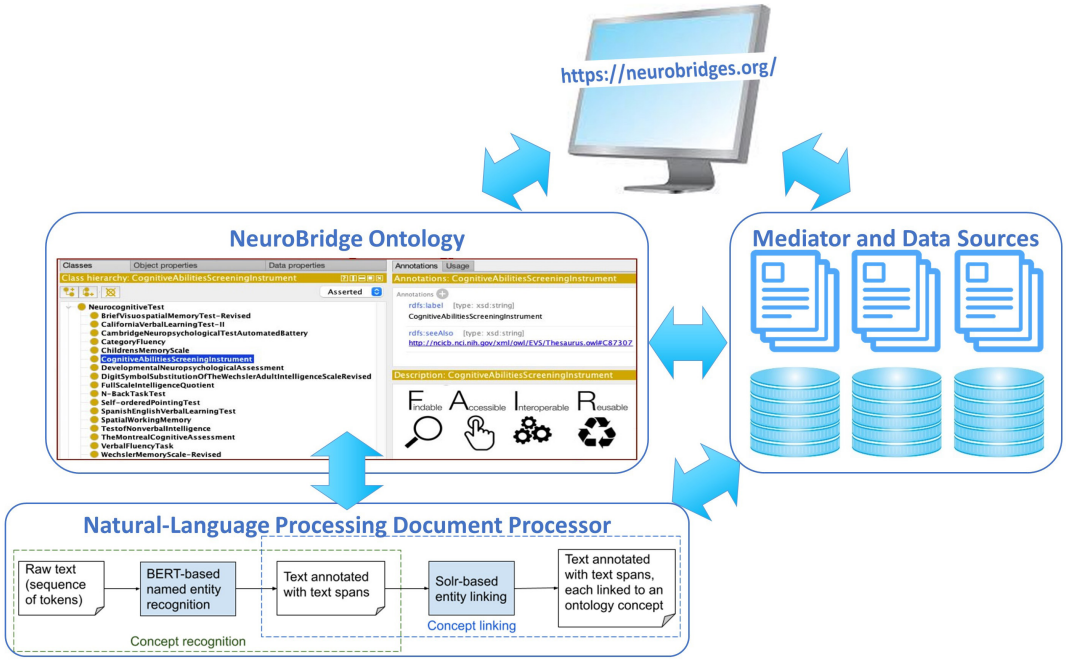


FIGURE 1
The NeuroBridge prototype architecture includes an extensible NeuroBridge ontology along with a set of full-text peer-reviewed papers annotated with the ontology, a natural-language processing document processor deep neural network model, and a user-friendly interface that contained an interactive query builder and integrated search across disparate data sources completed the prototype architecture.

The NeuroBridge ontology currently consists of 640 classes together with 49 properties that link the ontology classes. Using the ontology, we annotated 186 papers from our document corpus on the participant types, scanning, clinical and cognitive assessments. See the companion paper in this special issue for a more thorough presentation of the ontology and annotations (Sahoo et al., 2023), including the class hierarchy representing various diagnoses and assessment scales. The latest version of the NeuroBridge ontology is available on GitHub (NeuroBridge) (see text footnote 4) and will be made available on BioPortal soon.

Ontology-based natural language document processor

The goal of the document processor was to extract from full-text articles in our corpus any relevant metadata information regarding study design and data collection as modeled by the NeuroBridge ontology. A key element of the design was to represent each full-text article in the corpus as a collection of the ontological concepts, instead of the original representation as a sequence of words in the full text. This eliminated the need to generate synonyms, hypernyms and hyponyms that are common in text-based search platforms. For the prototype reported here, the sample size of our corpus of annotated full-text papers was small relative to the number of ontological concepts (186 vs. 640, respectively, see above). This small sample size did not lend itself to an end-to-end deep-learning model that would simultaneously tag and classify text spans into the ontology terms. Our prior research on low-resource named entity recognition showed that when the training set was small and entity tokens were sparse, fine-tuning a pre-trained large language model had a consistent performance advantage over training simpler models such as conditional random fields or bi-directional long short-term memory (Wang and Wang, 2022). This led to the development of a two-stage machine-learning model, described in detail in Wang et al. (2022) and briefly outlined here.

TABLE 1 The prototype document corpus.

PMC search	Schizophrenia	Substance-related disorder
Search string	["functional neuroimaging" (mh)] ["schizophrenia" (mh)] NOT [meta-analysis(pt) OR review(pt)] NOT [meta-analysis(ti) or review(ti)]	["functional neuroimaging" (mh)] ["substance-related disorders" (mh)] NOT [meta-analysis(pt) or review(pt)] NOT [meta-analysis(ti) or review(ti)]
Additional PMC filters applied to both searches	Free full text; Time In the last 5 years; Subjects: Humans; language: English	
# of returns on PMC	335	200
# of articles retrieved from BioC	196	162
# of articles used in document collection	196 + 162–2 = 356 (two articles are common between the above two sets)	

Stage 1 of the model was concept recognition, where text spans in the full text that may mention any ontological concept term were tagged. This was formulated as a binary sequence tagging task to determine whether a text span should be recognized as *any* concept or not, regardless which concept it is linked to. We employed the Bidirectional Encoder Representations from Transformers (BERT) with a conditional random fields (CRF) output layer as the binary sequence tagging model. BERT is a deep neural network model for natural language (Devlin et al., 2019) that learns from a corpus of documents to obtain the contextual representation of a word using information from all other words in a sentence. This makes BERT especially powerful in fine-grained natural language processing tasks (both at a sentence and at the word level) where nuanced syntactic and semantic understanding is critical.

Then in Stage 2, concept linking, the tagged texts were mapped to the most relevant concept in the ontology. For each concept, we constructed a "concept document" by concatenating its textual labels in the NeuroBridge ontology, its synonyms in the UMLS, and its associated text spans in the training data. We then calculated the textual similarity between the text span and the concept document by using Apache Solr to index all concept documents where a text span was treated as a free-text query and the BM25 relevance model (Amati, 2009) was used to rank concepts. The textual similarity provided a measure of relevance of a text span with respect to a concept, which was used to train and develop the model. In the case where Solr returned no result for a given text span, we used fuzzy string matching (i.e., Jaccard similarity of two sets of letter trigrams) between the text span and a concept as a fallback strategy to rank the relevance to the concepts.

For each of the articles in our corpus (except those used for training), we applied the trained two-stage document processor on the Sections "Abstract" and "Methods" to create a representation as a collection of machine-recognized ontological concepts. During queries performed in the NeuroBridge search portal (described below), these representations would be used to match against the query criteria.

Interactive search portal and integrated query across disparate sources

Overview

When the user comes to the NeuroBridge search portal website (see text footnote 1), a typical workflow begins in the query builder interface with the construction of a query by the user selecting a series of NeuroBridge ontology terms as search keywords. The query is then passed to the backend to search across the different data sources. Returns from each data source are then listed for further exploration by the user.

Query construction

In the Query Builder window, the user types in parts of the keyword that they want to query on, and the Query Builder will present a list of suggested ontology concept terms based on the spelling of the partial keyword. By default, all descendants of the selected ontology concept term will be included and the user can include and exclude individual descendants. The user can continue to add additional ontology concept terms to the query. An example query is shown in **Figure 2A**, constructed on the ontological concepts of "Schizophrenia," "FunctionalMagneticResonanceImaging," "NegativeSymptomScale," with all the descendants of these terms automatically included into the query.

The portal front-end will form the final query by joining the terms together with Boolean logics of "AND" and "OR," and represents it in a JSON format to preserve the ontology tree structure. A "View Query" option on the Query Builder portal allows the user to inspect the query syntax before submitting for execution. Upon user submission, the Boolean-represented query is then sent to the backend to be matched against the ontology representations of the full-text articles in the document corpus, as described above.

Query across disparate sources

For the same query the user constructed, we have also implemented mediation strategies to search additional data sources. In the current NeuroBridge prototype, in addition to the PMC articles corpus, we have incorporated NeuroQuery (Dockes et al., 2020) as a second data source and are currently working on incorporating XNAT (Marcus et al., 2007a) data sources. NeuroQuery is a platform that provides fMRI activation maps along with PubMed source articles (Dockes et al., 2020). It has a native search interface for user-input free texts and returns which terms, PMC publications, and brain regions are related to the query. The matching within NeuroQuery is based on its library of ~7,500 native terms and ~13,000 PMC neuroimaging articles.

We directed the NeuroBridge search to NeuroQuery by employing ElasticSearch and SapBERT (2023)[5] to semantically match terms in the NeuroBridge ontology to the native NeuroQuery terms so that terms being queried at NeuroBridge can be translated to NeuroQuery native terms. The translation process started by using SapBERT to create a floating-point vector of dimension 768 for each of NeuroQuery's native terms. These vectors represented the position in SapBERT's feature space of each of the terms. The vectors were then loaded into an ElasticSearch index that could be accessed by a Flask based API. To translate a NeuroBridge term to a NeuroQuery term, the API used SapBERT to create a corresponding vector for the NeuroBridge term. Then using the Cosine Similarity capability in ElasticSearch, the vector representing the NeuroBridge term was compared to the vector representing each of the NeuroQuery vectors to select the closest match. As an example, suppose the user has selected the NeuroBridge term "abstinent." Searching NeuroQuery using its native API did not return any data. Searching the ElasticSearch index for the closest match to abstinent selected the NeuroQuery term "abstinence." Using the NeuroQuery native API with this term returned several matches. The use of ElasticSearch and SapBERT

5 SapBERT, 2023: https://github.com/cambridgeltl/sapbert.

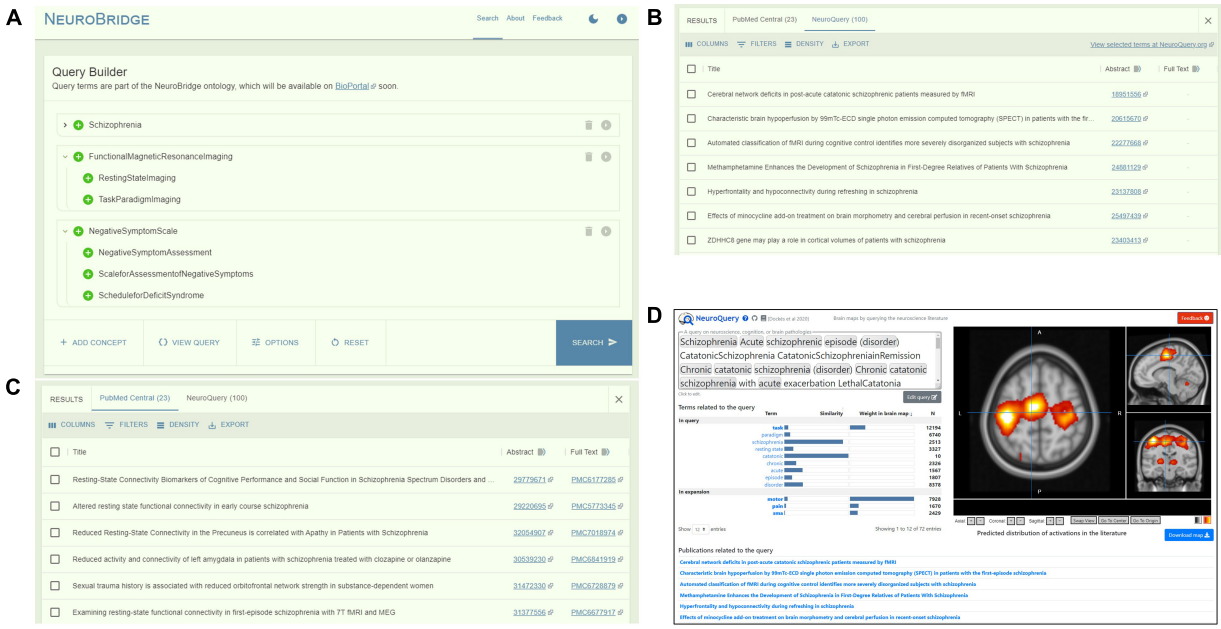


FIGURE 2
The NeuroBridge portal of an example query of "Schizophrenia" AND "FunctionalMagneticResonanceImaging" AND "NegativeSymptomScale." **(A)** Query builder interface showing query construction on the ontology concepts that automatically included all descendants. **(B)** Returns from PubMed Central with links to full-text articles. **(C)** Returns from NeuroQuery with links to PubMed abstracts. **(D)** User is directed to the NeuroQuery portal for direct interaction.

enabled searching the NeuroQuery API using its native term set while still enabling the user to search using the NeuroBridge ontology.

Return exploration

In the Results panel, returns of the query from each data source are presented to the user in its own tab. For returns from the PMC article corpus, the returns are sorted by relevance as computed above. Figure 2B shows that the query on the terms "Schizophrenia," "Functional Magnetic Resonance Imaging," "NegativeSymptomScale," and all their descendants resulted in a return of 23 PMC articles from the NeuroBridge corpus. For each return entry, links to the PubMed abstract as well as to the PMC full-text article, if available, are presented.

The same query resulted in a return of 100 articles from NeuroQuery (Figure 2C) (note: NeuroQuery by default returns 100 articles ranked by relevance from their corpus of ∼13,000 articles). A link to the NeuroQuery portal is also provided for users who are interested in interacting directly with NeuroQuery (Figure 2D).

We experimented with additional capabilities on the returned articles for providing useful information to the user. One kind of useful information is the set of clinical, behavioral and cognitive assessments that a study may have used. We first extracted a list of >4,400 names of common assessment instruments from the National Institute of Mental Health Data Archive (NDA). NDA is an informatics platform that supports data sharing across all mental health and other research communities. The list of assessment instruments thus spans across all mental health conditions[6]. The extracted list was in JSON format, where each assessment has a unique "title" (e.g., "Brief Psychiatric Rating Scale") and a unique "shortName" (e.g., "bprs01"). See Figure 3 for an example entry. We used the Apache Solr-based method we employed in the Document Processor (see previous section) to compute textual similarities between the assessment "title" and the texts in the Section "Methods" of the paper. Matched items were collated for each returned article and presented to the user. For example, for the returned article (PMCID PMC6177285) (Viviano et al., 2018), the assessments included "Brief Psychiatric Rating Scale," "Cumulative


```
{
    "shortName": "bprs01",
    "title": "Brief Psychiatric Rating Scale",
    "sources": [      "NDA"    ],
    "categories": [      "Questionnaire"    ],
    "dataType": "Clinical Assessments",
    "status": "Published",
    "publicStatus": "Submission Allowed",
    "publishDate": "2014-06-12T18:15:45.353+00:00",
    "modifiedDate": "2023-01-19T01:32:56.999+00:00"
},
```


FIGURE 3
One of the more than 4,400 common assessment instruments from the National Institute of Mental Health Data Archive (NDA), the assessment "Brief Psychiatric Rating Scale," in JSON format.

6 Nimh Data Archive [NDA]: https://nda.nih.gov/general-query.html?q=query=data-structure%20%7Eand%7E%20dataTypes=Clinical%20Assessments%20%7Eand%7E%20orderBy=shortName%20%7Eand%7E%20orderDirection=Ascending%20%7Eand%7E%20resultsView=table-view.

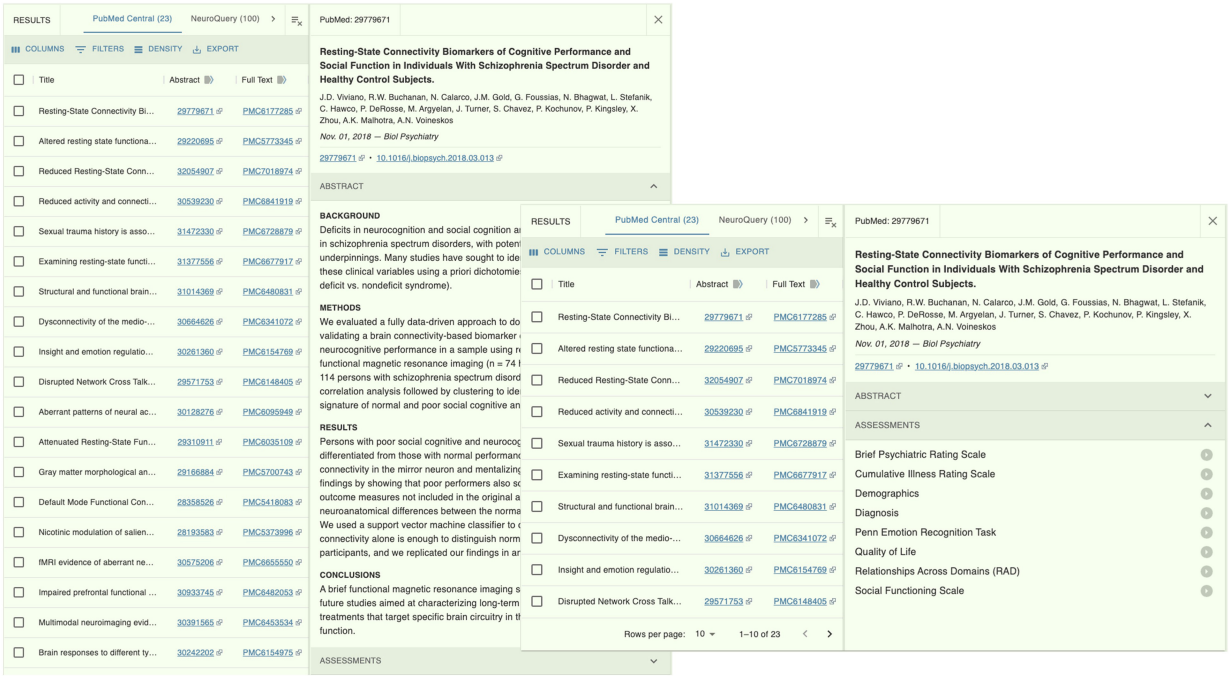


FIGURE 4
The NeuroBridge portal of returns on the example query in **Figure 2**, showing a list of clinical assessments described in the full-text article.

Illness Rating Scale," "Penn Emotion Recognition Task," and "Social Functioning Scale" (**Figure 4**).

As another example, we built a query using concepts "CannabisAbuse," "StructuralImaging," "NeurocognitiveTest" and their descendants (**Figure 5A**). **Figures 5B–D** show the returned PMC articles from the NeuroBridge corpus and results from NeuroQuery.

Power of ontology-based search

To demonstrate the power of ontology-based search, we compared query results of "Schizophrenia," "Resting-State Imaging," and "Young Mania Rating Scale" between NeuroBridge and a direct search on PMC. On the NeuroBridge search portal, two articles were returned: Lewandowski et al. (2019) "Functional Connectivity in Distinct Cognitive Subtypes in Psychosis" (PMC6378132) and Karcher et al. (2019) "Functional Connectivity of the Striatum in Schizophrenia and Psychotic Bipolar Disorder" (PMC6842092) (**Figure 6**). In Lewandowski et al. (2019), the Sections "Methods" included the terms "schizophrenia," "Young Mania Rating Scale (YMRS)," and "resting-state functional scans" (**Figure 7A**). In Karcher et al. (2019), the Sections "Methods" included the terms "schizophrenia," "Young Mania Rating Scale (YMRS)," and "resting-state fMRI" (**Figure 7B**). In comparison, the direct search on the PMC portal failed to return any entries. Additional synonyms such as "Resting-State fMRI" or "Resting-State functional" resulted in returns from the PMC. While the returns included the above articles, they also included many false positives. For example, the article by Gallucci et al. (2022) "Longer illness duration is associated with greater individual variability in functional brain activity in Schizophrenia, but not bipolar disorder" (PMC9723315) included the terms "schizophrenia" and "Young Mania Rating Scale (YMRS)" in the Section "Methods," the study did not utilize resting-state fMRI - subjects performed the N-back fMRI only.

Discussion

In this paper we describe the NeuroBridge: a project that takes a first step toward the discovery of gray neuroimaging data for reuse. The term "gray data" refers to data that has been gathered and used for analysis but is not publicly available. Reuse of these data is economic (i.e., compared with the large amount of funding required to collect new data) and can enhance reproducibility research (e.g., by facilitation of replication as well mega-analysis of aggregated data). Traditionally, finding data has been done mainly through professional networking and manually searching the literature[7]. However, much of the data mentioned in publications has not been shared yet through data links (such as DOI) or described in any searchable databases. Few resources currently exist that can help researchers find the right kind of data described in publications that are appropriate for their research questions.

Recent efforts have begun to facilitate these searches. For example, the field of life sciences requires papers to be deposited in domain repositories and uses DOIs to help to make data

7 Wageningen University & Research: https://www.wur.nl/en/Library/Researchers/Finding-sources/Finding-research-data.htm.

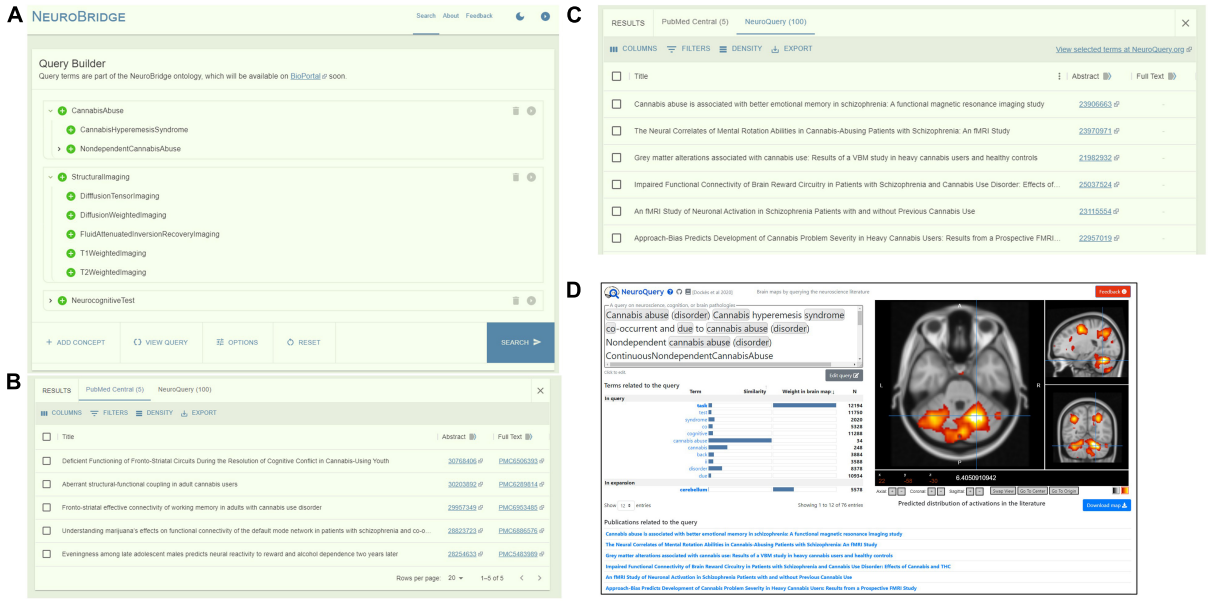


FIGURE 5
An example query of "CannabisAbuse" AND "StructuralImaging" AND "NeurocognitiveTest." **(A)** Query builder interface showing query construction on the ontology concepts that automatically included all descendants. **(B)** Returns from PubMed Central with links to full-text articles. **(C)** Returns from NeuroQuery with links to PubMed abstracts. **(D)** User is directed to the NeuroQuery portal for direct interaction.

search easier than before. Data journals publish peer-reviewed documentation of data and provide repositories for the data (Walters, 2020). The EuropePMC platform (see text footnote 2) provides links to data related to publications, but only data that has a citation or has been uploaded to one of the known 40 biological databases. Despite progress on improving data sharing and data curation, a wide gap exists in finding such data of interest for a user. Many of the legacy articles do not have explicit data citations and there is no standard in which articles detail the data gathering and data analysis operations. Finding whether there are any references to data of interest in these articles is a time-consuming manual process. Currently, finding data described in publications involves manually searching the literature, identifying the returned articles that are closely related to a research topic, and checking if the datasets described in the articles are appropriate and if the articles include references to the datasets.

The NeuroBridge prototype platform described in this paper aims to ease the burden for the user and takes a first step toward the discovery of gray neuroimaging data for reuse. NeuroBridge is powered by a machine learning system that is trained to identify clinical neuroscience metadata terms, including diagnosis, MRI scan types, and clinical assessments in a subset of articles that are accessible through PubMed Central. The current prototype is trained with an ontology in the domains of schizophrenia and substance-use disorders along with the clinical terms to facilitate discovery of relevant neuroimaging data described in peer-reviewed full-text journal papers. In the prototype platform, the user can perform a keyword-based search related to their research question, examine the returned papers for types of clinical assessment data, and pursue data access either via the data availability information or author contacts, both of which are provided in the NeuroBridge search returns.

Related work

The long-term goal of the NeuroBridge project is to provide researchers who are searching for neuroimaging data for a specific project (e.g., meta or mega analysis of a specific neuroimaging type in a specific clinical domain) with sufficient data of the right kind. Toward this goal, the NeuroBridge prototype reported here builds upon a number of previous and ongoing efforts on building ontologies of study design and text mining.

To discover studies that may contain relevant data for the user, we rely on the provenance metadata that can model study population, design and data collection. Extensive research on provenance metadata collection, storage, and querying has led to the emergence of relational databases, scientific workflow systems, sensor networks, and Semantic Web applications (Moreau et al., 2008; Widom, 2008; Sahoo et al., 2019). The biomedical research community has developed ontologies to model metadata information associated with clinical trials such as eligibility criteria (Tu et al., 2011; Sim et al., 2014). In the neuroscience research community, several recent initiatives have made significant progress toward identifying metadata information that can be used to describe the context of studies. These initiatives include the Neuroimaging Data Model (NIDM) (Keator et al., 2013), Reproducible Neuroimaging Computation (ReproNim) (Kennedy et al., 2019), and Brain Imaging Data Structure (BIDS) (Gorgolewski et al., 2016). These efforts, however, are focused on prospective studies on modeling provenance metadata prior to publication, while the goal of the NeuroBridge modeling approach is to help locate completed studies that are described in papers.

Many semantic search systems index unstructured content (usually text) using concepts or terms in a target ontology and allow users to query the content using these terms. The most prominent

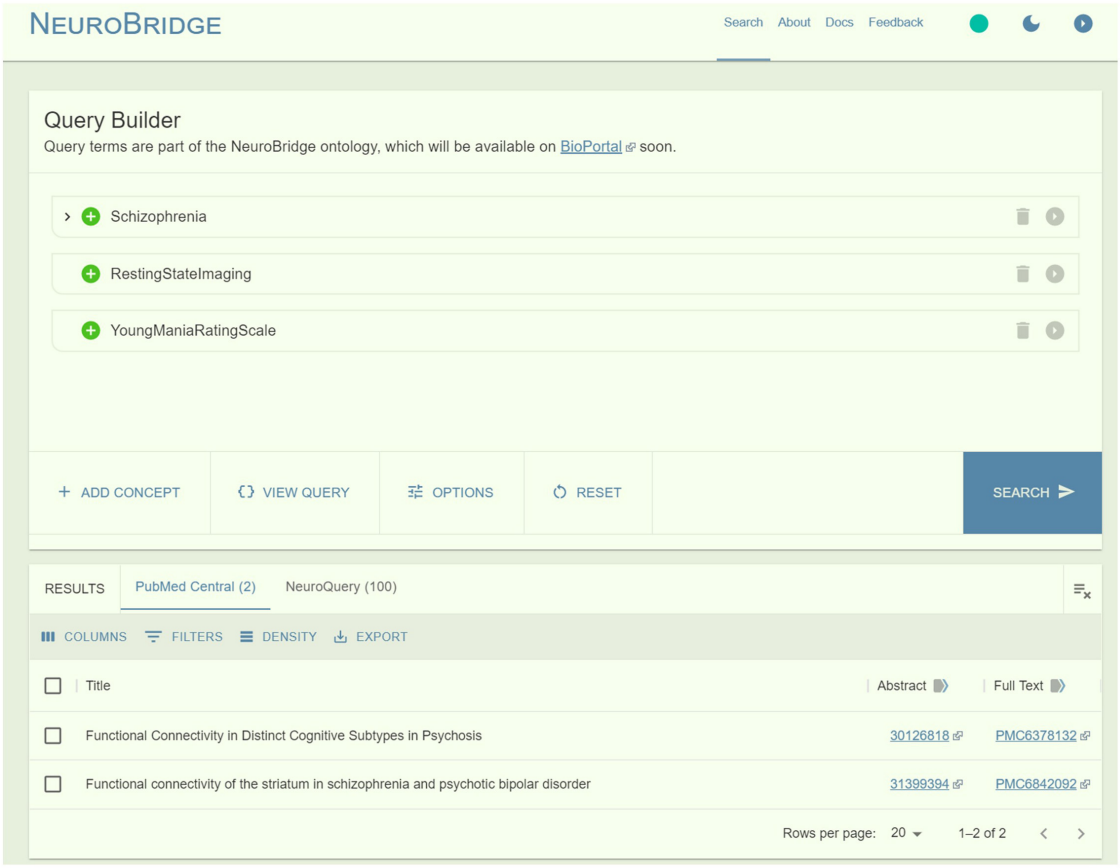


FIGURE 6
The power of ontology-based search, as demonstrated by a query of "Schizophrenia" AND "Resting-State Imaging" AND "Young Mania Rating Scale": On NeuroBridge, two articles were returned.

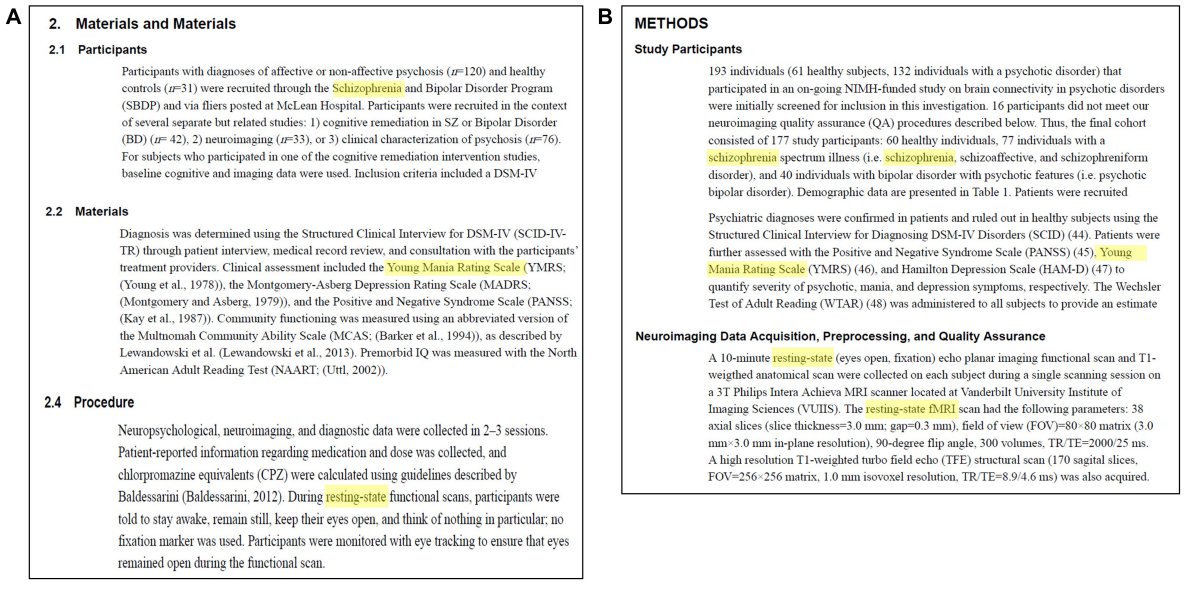


FIGURE 7
The power of ontology-based search, continued, as demonstrated by the query shown in Figure 6: Relevant text snippets in panel **(A)** Lewandowski et al. (2019) and **(B)** Karcher et al. (2019). In comparison, the direct search on the PMC portal failed to return any entries.

system is PubMed, which indexes the biomedical literature using terms in Medical Subject Headings (MeSH) and allows users to use MeSH terms in their search queries. The MeSH terms are currently automatically assigned to each PubMed paper using the MTI system[8,9], with a selected subset of papers reviewed by human indexers for quality control. Another system is LitCovid, which annotates and searches COVID-19-related research articles with medical terms such as genes, diseases, and chemical names (Chen et al., 2021). Other search engine prototypes such as SemEHR (Wu et al., 2018) and Thalia (Soto et al., 2019) assign terms in the Unified Medical Language System (UMLS) to documents and use these terms as search facets. The radiology image search engine prototype GoldMiner (Kahn and Thao, 2007) assigns terms in Systematized Nomenclature of Medicine Clinical Terms (SNOMED CT) and MeSH terms to image documents to facilitate image search. A key advantage of these systems compared to keyword-based search engines is that they allow users to directly use ontological concepts to express specific information needs that are otherwise challenging to precisely express through keywords.

A significant amount of research efforts has been dedicated to extracting semantic concepts from unstructured text. The problem is referred to as semantic indexing when the extracted concepts are used to represent texts in an information retrieval system (Reinanda et al., 2020). The problem is usually formulated as a natural language processing task, such as named entity recognition (Li et al., 2022), entity linking (Shen et al., 2015), or multi-label text classification (Mao and Lu, 2017). To solve these tasks, machine learning techniques are often employed. A machine learning system learns from a set of articles with human-assigned terms as training examples and generates a model that generalizes the term assignment procedure from the training articles to new unlabeled articles. Neural language models such as BERT (Devlin et al., 2019) can often deliver state-of-the-art performance on these tasks. These models learn rich prior knowledge from large-scale unlabeled text in their pre-training stage, which makes them easily adaptable to specific tasks by fine-tuning on a relatively small training dataset. A recent platform Elicit[10] uses Generative Pre-trained Transformer (GPT) to find papers related to a research question based on semantic similarity. For NeuroBridge, the ability to quickly learn from a small training dataset is important since it is expensive and time-consuming to curate even a moderate amount of biomedical research articles with concepts in a complex ontology. We have previously developed a deep-learning classification algorithm without large-scale training data (Wang et al., 2022). This was achieved by exploiting BERT that had been pre-trained on large unannotated text corpus and further fine-tuning it on annotated data that encoded rich semantic knowledge in the ontology. The technique could generalize to a wide range of biomedical text mining scenarios where the target ontological structure is complex but constructing large training data sets is too expensive and time-consuming.

Currently, a researcher can pursue the following ways to find data for their research question: utilize their professional network and institutional resources such as data search engines available at institutional libraries (e.g., University of Bath, 2023), search known data repositories such as ones listed in Eickhoff et al. (2016), search indices of datasets such as DataCite's Metadata Search[11]. The researcher can also search the literature. A number of journals in the field of biology, medicine and health sciences such as Scientific Data, Journal of Open Psychology Data, and Open Health Data are dedicated to the documentation and access of data created through research (Walters, 2020). While an increasing number of researchers are documenting their newly collected data in data journals, valuable, legacy data remain hidden in the literature. However, searches for data in the literature are performed by the researcher searching on literature databases such as PubMed Central, Open Science Framework then reading through each paper. There appears to be no current effort of systematically aiding this process. The abovementioned Elicit platform (see text footnote 10) offers advanced features such as extracting the number of participants and detailed study designs (e.g., case-control design, use of fMRI). To our knowledge, the NeuroBridge project is the first of its kind that is aimed at searching for relevant neuroimaging data described in peer-reviewed full-text papers.

Conclusion and future work

The NeuroBridge prototype we presented here uses an ontology-based approach to facilitate the search for relevant peer-reviewed journal papers. While limitations exist, such as the small sample size of our training and testing corpus, it nevertheless takes an important first step toward identifying potential neuroimaging data described in full-text papers that are relevant to a particular user's research interests. Work is ongoing to validate the document processor with a larger corpus, extend the ontology to include detailed imaging data, extract information regarding data availability from the returned publications to enhance data accessibility (FAIR), and measure semantic distances between studies based on assessment information to help identify relevance of studies to the user (Lander et al., 2019). Future work also involves extending the ontology and document corpus to include additional clinical domains (e.g., psychosis spectrum, dementia). These extensions will require similarly significant human effort including manually labeling a training set of papers with the ontology terms and careful review and curation of this work. See the companion paper in this issue for more detail of the labeling methods (Sahoo et al., 2023). As the system grows, the current iteration of the system supports this human labeling process by providing draft labels, and the entity-recognition, entity-linking, 2-stage natural language model will be retrained to complete the extension.

There is an increasing availability of multi-modal datasets in neuroscience research, especially as a result of the National Institutes of Health (NIH) Brain Research Through Advancing Innovative Neurotechnologies (BRAIN) initiative. NIH has developed large-scale data repositories such as the National

8 National Library of Medicine (NLM Medical Text Indexer): https://lhncbc.nlm.nih.gov/ii/tools/MTI.html.

9 National Library of Medicine (Automated Indexing FAQs): https://support.nlm.nih.gov/knowledgebase/article/KA-05326/en-us.

10 Elicit: https://elicit.org/.

11 DataCite: https://commons.datacite.org/.

Institute of Mental Health (NIMH) Data Archive (NDA) that contains datasets from structural and functional MRI, clinical phenotypes, and genomics. Eickhoff et al. (2016) described >40 neuroimaging data repositories across multiple clinical domains. A need exists to develop a metadata-based search and discovery platform on similar search criteria. Work is ongoing at the NeuroBridge project to incorporate XNAT-based (Marcus et al., 2007a) neuroimaging databases into our search. XNAT is a web-based software platform designed to facilitate common management and productivity tasks for imaging and associated data. It has been broadly adopted across domains of neuroscience, cardiology, cancer, and ophthalmology, supporting a wide range of many high impact data sharing initiatives, including OASIS (Marcus et al., 2007b, 2010), Dementia Platform UK, Human Connectome Project (Hodge et al., 2016), UK Biobank (Miller et al., 2016), NITRC Image Repository (Kennedy et al., 2015), and SchizConnect (Wang et al., 2016). These resources offer comprehensive data from deep phenotyping of subjects, including multiple imaging modalities and clinical, cognitive, behavior, and genomic data. As the number of datasets rapidly grows, often the problem is not finding datasets, but selecting enough data of the right kind from a large corpus of possible datasets.

Our long-term goal is to discover "enough data of the right kind" by providing a user-friendly portal for automatically searching multiple types of sources and identifying relevant datasets. We envision a scenario where a graduate student or a postdoctoral fellow from a small institution can use NeuroBridge to discover data for testing specific hypotheses. For example, she may have read an interesting paper on how changes in brain networks are modulated by cognitive demand but the effects are different by sex. She would like to design a study to test the hypothesis or replicate the paper's findings. However, her lab does not have the resources or budget for MR scanning or subject recruitment, and she can find only a very limited amount of data fitting her research needs in public databases. The student would then need to search through the literature to find data that are similar to the original study. It would take her an inordinate amount of time to comb through the details described in papers and decide whether they have the required data.

Additional future work of the NeuroBridge project includes: extracting detailed information on details of the study such as study design, sample demographic information as well as author contacts and data availability described in research papers, and identifying the location and links to such data if shared (through collaboration with platforms such as Brainlife (Avesani et al., 2019)[12] where shared data are associated with publications. In the not too distant future, researchers like this student would interact with the NeuroBridges.org and its APIs, describe a study, craft their hypothesis, and in a few steps discover how many studies and datasets contain subjects and data that can be used to answer their research question. Our platform will become a key component of the data sharing ecosystem that provides researchers with sustainable means of aggregating data–from discovery, to access and harmonization – that are directly relevant to their hypothesis, and compute on the data to test their hypotheses. It will enable more small-market scientists to do large-scale research and thus increase the findability, accessibility, and reusability of scientific data to a greater number of researchers. We believe our approach can become the prototype in other domains for bridging from the research question, to data, to scientific workflow, thereby significantly speeding up the cycle of hypothesis-based research.

12 https://brainlife.io/

Data availability statement

Publicly available datasets were analyzed in this study. This data can be found here: the NeuroBridge ontology: https://github.com/NeuroBridge/neuro-ontologies/tree/main/neurobridge.

Author contributions

LW, JA, HL, AR, and JT contributed to the conception and design of the study. SS, AA, AK, MT, XW, YW, and JT developed the document corpus and the ontology and its annotations. XW and YW developed the natural-language based document processor. JD, HL, and J-BP contributed to the connection with NeuroQuery. JA, JB, RH, DM, and SM contributed to data mediation. HL coordinated the development of the search portal. LW wrote the first draft of the manuscript. XW, YW, HL, AR, MT, and JT wrote sections of the manuscript. All authors contributed to manuscript revision and read and approved the submitted version.

Funding

The efforts described in this manuscript are funded by NIDA grant R01 DA053028 "CRCNS:NeuroBridge: Connecting big data for reproducible clinical neuroscience," the NSF Office of Cyberinfrastructure OCI-1247652, OCI-1247602, and OCI-1247663 grants, "BIGDATA: Mid-Scale: ESCE: DCM: Collaborative Research: DataBridge–A Sociometric System for Long Tail Science Data Collections," and by the NSF IIS Division of Information and Intelligent Systems grant number #1649397 "EAGER: DBfN: DataBridge for Neuroscience: A Novel Way of Discovery for Neuroscience Data," NIMH grant U01 MH097435 "SchizConnect: Large-Scale Schizophrenia Neuroimaging Data Mediation and Federation," NSF grant 1636893 SP0037646 "BD Spokes: SPOKE: MIDWEST: Collaborative: Advanced Computational Neuroscience Network (ACNN)." J-BP and JD were partially funded by the Michael J. Fox Foundation (LivingPark), the National Institutes of Health (NIH) NIH-NIBIB P41 EB019936 (ReproNim) NIH-NIMH R01 MH083320 (CANDIShare) and NIH RF1 MH120021 (NIDM), the National Institute Of Mental Health of the NIH under Award Number R01MH096906 (Neurosynth), as well as the Canada First Research Excellence Fund, awarded to McGill University for the Healthy Brains for Healthy Lives initiative and the Brain Canada Foundation with support from Health Canada. This project has been made possible by the Brain Canada Foundation, through the Canada Brain Research Fund, with the financial support of Health Canada and the McConnell Brain Imaging Centre.

Acknowledgments

We would like to thank Matthew Watson and his team for the technical contribution to the development of the NeuroBridge search portal.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

References

Amati, G. (2009). "BM25," in *Encyclopedia of database systems*, eds L. Liu and M. T. Özsu (Boston, MA: Springer).

Avesani, P., McPherson, B., Hayashi, S., Caiafa, C. F., Henschel, R., Garyfallidis, E., et al. (2019). The open diffusion data derivatives, brain data upcycling via integrated publishing of derivatives and reproducible open cloud services. *Sci. Data* 6:69. doi: 10.1038/s41597-019-0073-y

Cao, Z., Ottino-Gonzalez, J., Cupertino, R. B., Schwab, N., Hoke, C., Catherine, O., et al. (2021). Mapping cortical and subcortical asymmetries in substance dependence: Findings from the ENIGMA Addiction Working Group. *Addict. Biol.* 26:e13010. doi: 10.1111/adb.13010

Chakrabarti, C., Jones, T. B., Luger, G. F., Xu, J. F., Turner, M. D., Laird, A. R., et al. (2014). Statistical algorithms for ontology-based annotation of scientific literature. *J. Biomed. Semant.* 5:S2.

Chen, Q., Allot, A., and Lu, Z. (2021). LitCovid: An open database of COVID-19 literature. *Nucleic Acids Res.* 49, D1534–D1540.

Comeau, D. C., Wei, C. H., Islamaj, D., and Lu, Z. (2019). PMC text mining subset in BioC: About three million full-text articles and growing. *Bioinformatics* 35, 3533–3535. doi: 10.1093/bioinformatics/btz070

Devlin, J., Chang, M. W., Lee, K., and Toutanova, K. (2019). Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv* [Preprint]. doi: 10.48550/arXiv.1810.04805

Dockes, J., Poldrack, R. A., Primet, R., Gozukan, H., Yarkoni, T., Suchanek, F., et al. (2020). NeuroQuery, comprehensive meta-analysis of human brain mapping. *Elife* 9:e53385. doi: 10.7554/eLife.53385

Eickhoff, S., Nichols, T. E., Van Horn, J. D., and Turner, J. A. (2016). Sharing the wealth: Neuroimaging data repositories. *Neuroimage* 124, 1065–1068. doi: 10.1016/j.neuroimage.2015.10.079

Ferguson, A. R., Nielson, J. L., Cragin, M. H., Bandrowski, A. E., and Martone, M. E. (2014). Big data from small data: Data-sharing in the 'long tail' of neuroscience. *Nat. Neurosci.* 17, 1442–1447. doi: 10.1038/nn.3838

Fox, P. T., Laird, A. R., Fox, S. P., Fox, P. M., Uecker, A. M., Crank, M., et al. (2005). BrainMap taxonomy of experimental design: Description and evaluation. *Hum. Brain Mapp.* 25, 185–198. doi: 10.1002/hbm.20141

Gallucci, J., Pomarol-Clotet, E., Voineskos, A. N., Guerrero-Pedraza, A., Alonso-Lana, S., Vieta, E., et al. (2022). Longer illness duration is associated with greater individual variability in functional brain activity in Schizophrenia, but not bipolar disorder. *Neuroimage Clin.* 36:103269. doi: 10.1016/j.nicl.2022.103269

Gorgolewski, K. J., Auer, T., Calhoun, V. D., Craddock, R. C., Das, S., Duff, E. P., et al. (2016). The brain imaging data structure, a format for organizing and describing outputs of neuroimaging experiments. *Sci. Data* 3:160044. doi: 10.1038/sdata.2016.44

Hodge, M. R., Horton, W., Brown, T., Herrick, R., Olsen, T., Hileman, M. E., et al. (2016). ConnectomeDB–Sharing human brain connectivity data. *Neuroimage* 124, 1102–1107. doi: 10.1016/j.neuroimage.2015.04.046

Jayakar, R. (2017). *Amygdala volume and social anxiety symptom severity: A mutli-method Study, psychology*. Atlanta, GA: Georgia State University.

Jayakar, R., Tone, E. B., Crosson, B., Turner, J. A., Anderson, P. L., Phan, K. L., et al. (2020). Amygdala volume and social anxiety symptom severity: Does segmentation technique matter? *Psychiatry Res. Neuroimaging* 295:111006.

Jayakar, R., Tone, E. B., Crosson, B. A., Turner, J. A., Anderson, P. L., Phan, K. L., et al. (2018). "Association between amygdala volume and social anxiety symptom severity: A multi-method study," in *46th Annual Meeting of the International Neuropsychological Society*, (Washington, DC).

Kahn, C. E. Jr., and Thao, C. (2007). GoldMiner: A radiology image search engine. *AJR* 188, 1475–1478.

Karcher, N. R., Rogers, B. P., and Woodward, N. D. (2019). Functional connectivity of the striatum in schizophrenia and psychotic bipolar disorder. *Biol. Psychiatry Cogn. Neurosci. Neuroimaging* 4, 956–965.

Keator, D. B., Helmer, K., Steffener, J., Turner, J. A., Van Erp, T. G., Gadde, S., et al. (2013). Towards structured sharing of raw and derived neuroimaging data across existing resources. *Neuroimage* 82, 647–661. doi: 10.1016/j.neuroimage.2013.05.094

Kennedy, D. N., Abraham, S. A., Bates, J. F., Crowley, A., Ghosh, S., Gillespie, T., et al. (2019). The repronim perspective on reproducible neuroimaging. *Front. Neuroinform* 13:1. doi: 10.3389/fninf.2019.00001

Kennedy, D. N., Haselgrove, C., Riehl, J., Preuss, N., and Buccigrossi, R. (2015). The three NITRCs: a guide to neuroimaging neuroinformatics resources. *Neuroinformatics* 13, 383–386. doi: 10.1007/s12021-015-9263-8

Laird, A. R., Lancaster, J. L., and Fox, P. T. (2005). BrainMap: The social evolution of a human brain mapping database. *Neuroinformatics* 3, 65–78. doi: 10.1385/ni:3:1:065

Lander, H., Alpert, K., Rajasekar, A., Turner, J., and Wang, L. (2019). "Data Discovery for Case Studies: The DataBridge for Neuroscience Project," in *Proceeding of the 13th International Multi-Conference on Society, Cybernetics and Informatics*, (Orlando, FL), 19–25.

Lewandowski, K. E., McCarthy, J. M., Ongur, D., Norris, L. A., Liu, G. Z., Juelich, R. J., et al. (2019). Functional connectivity in distinct cognitive subtypes in psychosis. *Schizophr. Res.* 204, 120–126.

Li, J., Sun, A., Han, J., and Li, C. (2022). A survey on deep learning for named entity recognition. *IEEE Trans. Knowl. Data Eng.* 34, 50–70.

Mao, Y., and Lu, Z. (2017). MeSH Now: Automatic MeSH indexing at PubMed scale via learning to rank. *J. Biomed. Semant.* 8:15. doi: 10.1186/s13326-017-0123-3

Marcus, D. S., Fotenos, A. F., Csernansky, J. G., Morris, J. C., and Buckner, R. L. (2010). Open access series of imaging studies: Longitudinal MRI data in nondemented and demented older adults. *J. Cogn. Neurosci.* 22, 2677–2684. doi: 10.1162/jocn.2009.21407

Marcus, D. S., Olsen, T. R., Ramaratnam, M., and Buckner, R. L. (2007a). The Extensible Neuroimaging Archive Toolkit: An informatics platform for managing, exploring, and sharing neuroimaging data. *Neuroinformatics* 5, 11–34. doi: 10.1385/ni:5:1:11

Marcus, D. S., Wang, T. H., Parker, J., Csernansky, J. G., Morris, J. C., and Buckner, R. L. (2007b). Open Access Series of Imaging Studies (OASIS): Cross-sectional MRI data in young, middle aged, nondemented, and demented older adults. *J. Cogn. Neurosci.* 19, 1498–1507. doi: 10.1162/jocn.2007.19.9.1498

Maumet, C., Auer, T., Bowring, A., Chen, G., Das, S., Flandin, G., et al. (2016). Sharing brain mapping statistical results with the neuroimaging data model. *Sci. Data* 3:160102.

Miller, K. L., Alfaro-Almagro, F., Bangerter, N. K., Thomas, D. L., Yacoub, E., Xu, J., et al. (2016). Multimodal population brain imaging in the UK Biobank prospective epidemiological study. *Nat. Neurosci.* 19, 1523–1536.

Moreau, L., Ludascher, B., Altintas, I., Barga, R. S., Bowers, S., Callahan, S., et al. (2008). The provenance challenge. *Concurr. Comput. Pract. Exper.* 20, 409–418.

National Institutes of Health. *NHI Reporter*. Vienna, VA: NHI.

National Institutes of Health (2023). *Data management and sharing policy*. Vienna, VA: NHI.

Reinanda, R., Meij, E., and de Rijke, M. (2020). Knowledge graphs: An information retrieval perspective. *Found. Trends Inform. Retrieval* 14, 289–444.

Sahoo, S. S., Turner, M. D., Wang, L., Ambite, J. L., Appaji, A., Rajasekar, A., et al. (2023). NeuroBridge ontology: Computable provenance metadata to give the long tail of neuroimaging data a FAIR chance for secondary use. *Front Neuroinform*. 17:1216443.

Sahoo, S. S., Valdez, J., Kim, M., Rueschman, M., and Redline, S. (2019). ProvCaRe: Characterizing Scientific Reproducibility of Biomedical Research Studies using Semantic Provenance Metadata. *Int. J. Med. Inform.* 121, 10–18.

Shen, W., Wang, J., and Han, J. (2015). Entity linking with a knowledge base: Issues, techniques, and solutions. *IEEE Trans. Knowl. Data Eng.* 27, 443–460.

Sim, I., Tu, S. W., Carini, S., Lehmann, H. P., Pollock, B. H., Peleg, M., et al. (2014). The Ontology of Clinical Research (OCRe): an informatics foundation for the science of clinical research. *J. Biomed. Inform.* 52, 78–91. doi: 10.1016/j.jbi.2013.11.002

Soto, A. J., Przybyla, P., and Ananiadou, S. (2019). Thalia: Semantic search engine for biomedical abstracts. *Bioinformatics* 35, 1799–1801. doi: 10.1093/bioinformatics/bty871

Tu, S. W., Peleg, M., Carini, S., Bobak, M., Ross, J., Rubin, D., et al. (2011). A practical method for transforming free-text eligibility criteria into computable criteria. *J. Biomed. Inform.* 44, 239–250.

Turner, J. A., and Laird, A. R. (2012). The cognitive paradigm ontology: Design and application. *Neuroinformatics* 10, 57–66.

Turner, M. D., Chakrabarti, C., Jones, T. B., Xu, J. F., Fox, P. T., Luger, G. F., et al. (2013). Automated annotation of functional imaging experiments via multi-label classification. *Front. Neurosci.* 7:240. doi: 10.3389/fnins.2013.00240

University of Bath (2023). *Finding and reusing research datasets: Finding Data Home*. Bath: University of Bath.

Viviano, J. D., Buchanan, R. W., Calarco, N., Gold, J. M., Foussias, G., Bhagwat, N., et al. (2018). Initiative in neurobiology of the schizophrenia, resting-state connectivity biomarkers of cognitive performance and social function in individuals with schizophrenia spectrum disorder and healthy control subjects. *Biol. Psychiatry* 84, 665–674. doi: 10.1016/j.biopsych.2018.03.013

Wallis, J. C., Rolando, E., and Borgman, C. L. (2013). If we share data, will anyone use them? Data sharing and reuse in the long tail of science and technology. *PLoS One* 8:e67332. doi: 10.1371/journal.pone.0067332

Walters, W. H. (2020). Data journals: Incentivizing data access and documentation within the scholarly communication system. *Insights UKSG J.* 33:18.

Wang, L., Alpert, K. I., Calhoun, V. D., Cobia, D. J., Keator, D. B., King, M. D., et al. (2016). SchizConnect: Mediating neuroimaging databases on schizophrenia and related disorders for large-scale integration. *Neuroimage* 124, 1155–1167. doi: 10.1016/j.neuroimage.2015.06.065

Wang, X., and Wang, Y. (2022). *Sentence-Level Resampling for Named Entity Recognition*. Seattle, US: Association for Computational Linguistics.

Wang, X., Wang, Y., Ambite, J. L., Appaji, A., Lander, H., Moore, S. M., et al. (2022). Enabling Scientific Reproducibility through FAIR Data Management: An ontology-driven deep learning approach in the NeuroBridge Project. *AMIA Annu. Symposium Proc.* 2022, 1135–1144.

Widom, J. (2008). "Trio: A System for Data, Uncertainty, and Lineage," in *Managing and Mining Uncertain Data*, ed. C. Aggarwal (Berlin: Springer).

Wilkinson, M. D., Dumontier, M., Aalbersberg, J., Appleton, G., Axton, M., Baak, A., et al. (2016). The FAIR Guiding Principles for scientific data management and stewardship. *Sci. Data* 3:160018.

Wu, H., Toti, G., Morley, K. I., Ibrahim, Z. M., Folarin, A., Jackson, R., et al. (2018). SemEHR: A general-purpose semantic search system to surface semantic data from clinical notes for tailored care, trial recruitment, and clinical research. *J. Am. Med. Inform. Assoc.* 25, 530–537. doi: 10.1093/jamia/ocx160

frontiers | Frontiers in Neuroinformatics

TYPE Brief Research Report
PUBLISHED 14 September 2023
DOI 10.3389/fninf.2023.1250260

OPEN ACCESS

EDITED BY
Christian Haselgrove,
UMass Chan Medical School, United States

REVIEWED BY
Rania Mohamed Hassan Baleela,
University of Khartoum, Sudan
Candido Cabo,
The City University of New York, United States

*CORRESPONDENCE
Peter Konradi
✉ peter.konradi@rwth-aachen.de

†These authors have contributed equally to this work

RECEIVED 29 June 2023
ACCEPTED 28 August 2023
PUBLISHED 14 September 2023

CITATION
Konradi P, Troglio A, Pérez Garriga A, Pérez Martín A, Röhrig R, Namer B and Kutafina E (2023) PyDapsys: an open-source library for accessing electrophysiology data recorded with DAPSYS.
Front. Neuroinform. 17:1250260.
doi: 10.3389/fninf.2023.1250260

PyDapsys: an open-source library for accessing electrophysiology data recorded with DAPSYS

Peter Konradi[1]*, Alina Troglio[2], Ariadna Pérez Garriga[1], Aarón Pérez Martín[3], Rainer Röhrig[1], Barbara Namer[2,4,5†] and Ekaterina Kutafina[1†]

[1]Institute of Medical Informatics, Medical Faculty, RWTH Aachen University, Aachen, Germany, [2]Research Group Neuroscience, IZKF, RWTH Aachen, Aachen, Germany, [3]Simulation and Data Lab Neuroscience, Jülich Supercomputing Centre (JSC), Institute for Advanced Simulation, JARA, Forschungszentrum Jülich GmbH, Jülich, Germany, [4]Department for Neurophysiology, University Hospital RWTH Aachen, Aachen, Germany, [5]Institute of Physiology and Pathophysiology, Friedrich-Alexander-Universität Erlangen-Nürnberg, Erlangen, Germany

In the field of neuroscience, a considerable number of commercial data acquisition and processing solutions rely on proprietary formats for data storage. This often leads to data being locked up in formats that are only accessible by using the original software, which may lead to interoperability problems. In fact, even the loss of data access is possible if the software becomes unsupported, changed, or otherwise unavailable. To ensure FAIR data management, strategies should be established to enable long-term, independent, and unified access to data in proprietary formats. In this work, we demonstrate PyDapsys, a solution to gain open access to data that was acquired using the proprietary recording system DAPSYS. PyDapsys enables us to open the recorded files directly in Python and saves them as NIX files, commonly used for open research in the electrophysiology domain. Thus, PyDapsys secures efficient and open access to existing and prospective data. The manuscript demonstrates the complete process of reverse engineering a proprietary electrophysiological format on the example of microneurography data collected for studies on pain and itch signaling in peripheral neural fibers.

KEYWORDS

interoperability, open data, FAIR, data management tools, reverse-engineered, microneurography, electrophysiology, pain

1. Introduction

Many commercial software solutions use custom proprietary formats to store their data. Reasons vary from dealing with special use cases to trying to lock users into a vendor-specific ecosystem. While there is a trend in the general IT space to open-source custom solutions and establish cross-vendor standards (Kilamo et al., 2012), in the scientific world, the focus is put on FAIR data principles (Wilkinson et al., 2016). FAIR principles consist of a number of requirements for data to be findable, accessible, interoperable, and reusable. Proprietary formats naturally obstruct the adoption of these principles. In some research domains, large efforts are put into building solutions to convert proprietary formats into open standards, while simultaneously lobbying companies to use open formats. Examples

TABLE 1 The size difference between CSV files created by the DAPSYS export and the files created by using PyDapsys with the NIX-exporter of the Neo library.

Original file size [MiB]	CSV file size [MiB]	NIX/H5 file size [MiB]	Size increase CSV [%]	Size increase NIX/H5 [%]
44.5	229.5	45.1	415.73	1.35
109.0	576.2	109.7	428.62	0.64
124.9	664.5	126.1	432.83	0.96
165.8	889.2	167.4	436.31	0.97

of such formats are DICOM[1] for storing, managing, and exchanging medical images and EDF (Kemp et al., 1992) for biosignals, including EEG systems.

Progress has also been made in the field of neuroscience, where the Neuroscience Information Exchange format (NIX) (Stoewer et al., 2014) and the Neurodata Without Borders (NWB) (Rübel et al., 2022) projects are aiming to establish community standards for sharing neuroscientific data. Both projects specify a storage layout, which is implemented on top of the Hierarchical Data Format (HDF5) but use different approaches to model data. NWB uses a stricter and more standardized data model, whereas NIX allows for a comparatively flexible structure and can describe the file contents using the open metadata Markup Language (odML) (Grewe et al., 2011).

However, smaller fields of neuroscience are facing challenges to fully adopt FAIR data principles, as vendors may not have the resources to address the specific wishes of such small user-bases. The "Data Acquisition Processor System" (DAPSYS)[2] is a general-purpose neurophysiological data acquisition system (DAS) for recording and processing neural signals, which is, among other places, used in the microneurography (MNG) lab of the University Hospital RWTH Aachen. MNG is an electrophysiological technique to record activity from single nerve fibers of the peripheral nervous system using a single microelectrode (Vallbo and Hagbarth, 1968; Torebjork and Hallin, 1974; Ackerley and Watkins, 2018). Due to the small size of the electrode, the method causes only minimal discomfort and does not require anesthetics. This means that the volunteer stays awake and cooperates during the recording, making it possible to correlate nerve fiber signals with individual sensations. Thus, MNG is a unique translational method in sensory research in humans, especially in chronic pain and itch.

DAPSYS uses a proprietary format to store data and only offers manual (file-by-file) export of the recordings to CSV files. However, the CSV exports produce comparatively large files (see Table 1) and take a long time (see Table 2). In addition, some minor precision loss due to the fixed number of decimals in the exported CSV is observed. Our recent works on establishing data-sharing standards in the MNG community and developing a computational pipeline for spike analysis in MNG data (Schlebusch et al., 2021; Kutafina et al., 2022; Troglio et al., 2023) has raised the urgency for an efficient way to read DAPSYS recordings and store them in more suitable data formats, such as HDF5.

While there are many commercial applications for reverse engineering, most of them target computer science professionals and the primary use-case of reverse engineering software, not file formats. The MARBLE project[3] is to our best knowledge the first research-oriented solution to reverse engineer file formats with the aim of making the process as accessible as possible. However, at the time of the reported work, MARBLE was still in development and the usage required problem-specific adjustments.

Therefore, in this paper, we show our approach to reverse engineering the DAPSYS file format and implement a Python library to gain open access to our own data recorded in the microneurography lab. By providing functionality to load data into the structure defined by the Neo library (Garcia et al., 2014), it can be simply exported to multiple data formats used in electrophysiology, including NIX. This ensures full access to the data even if DAPSYS is unavailable.

The primary aim of our work is to ensure the accessibility and interoperability of DAPSYS-recorded data sets. The secondary aim is to share the steps of our reverse engineering solution with the neuroscience community to support building FAIR access to rare data formats.

2. Method

2.1. Data

We used four DAPSYS files, recorded at the microneurography labs of the University Hospital RWTH Aachen and Friedrich-Alexander-University of Erlangen-Nürnberg. The studies involving human participants were reviewed and approved by the Ethics Boards of those two institutions with the corresponding numbers EK141-19 and 4361. The participants provided their written informed consent, and the studies were conducted according to the Declaration of Helsinki.

2.2. Reverse engineering method

For the reverse engineering process, we used the hex editor "ImHex"[4] to open and analyze the DAPSYS files. A hex editor shows the binary contents of a file in hexadecimal representation. A value of a single byte can be represented by only two characters, making it

1 DICOM: Digital Imaging and Communications in Medicine, Medical Imaging Technology Association (MITA), https://www.dicomstandard.org.

2 Data Acquisition Processor System (DAPSYS), Brian Turnquist, http://dapsys.net.

3 MARBLE software project, Steffen Brinckmann et al., https://gitlab-public.fz-juelich.de/marble.

4 ImHex, Nikolaij "WerWolv" Sägesser, https://github.com/WerWolv/ImHex.

TABLE 2 The time comparison for exporting the continuous recording.

Original file size [MiB]	CSV export time* [s]	PyDapsys export to NIX/H5 time* [s]	Speedup PyDapsys vs. CSV export*	PyDapsys total time [s]
44.5	35	0.36	97.2	0.77
109.0	91	0.46	197.8	0.84
124.9	102	0.68	150.0	1.03
165.8	133	0.98	135.7	1.49

*Time required for processing and writing, excluding user interaction.

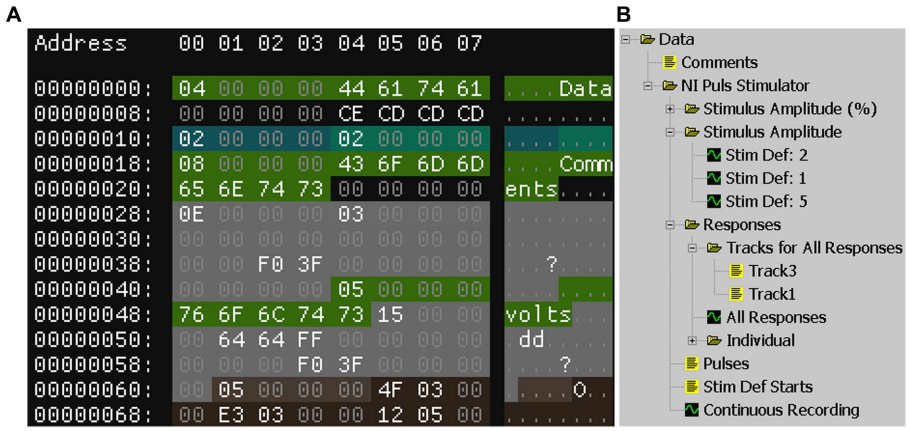


FIGURE 1
(A) "ImHex" showing the start of a DAPSYS files' table of contents section. Shown are the hexadecimal byte-values of the respective address and the interpretation of that byte as characters. Data fields of structures are shown in different colors. The address shown is in relation to the start of the table of contents. **(B)** Structure of the same file from panel **(A)** shown by the GUI of DAPSYS.

easier to recognize patterns (see Figure 1A for an example). Since we knew what values the file should contain, we were able to search for them and identify related fields. From there on, we identified structures based on repeating patterns.

The functions of data fields in the structures were then identified by using the following workflow:

1. Make changes to the file using the DAPSYS GUI (for example: changing the plot configuration, removing data points, etc.).
2. Track these changes DAPSYS made to the binary file and identify changed fields in the hex editor.
3. Open a different recording in the hex editor, identify the known fields, and change their values using the hex editor.
4. Open the changed file from step 3 in DAPSYS and verify that the changes made to the recording fit with the assumed function of the field.

This process was substantially supported by built-in "ImHex" functions like the pattern language that can be used to specify the layout of structures in the binary file. These structures can be utilized to highlight and verify known structures and fields in the file.

2.3. Concept of the library implementation

Based on the results from the reverse-engineering process, we implemented a Python library capable of opening and processing recordings. The library also offers a method to export data from DAPSYS recordings into HDF5 files using the NIX structure (abbreviated as NIX/H5) for easier data exchange between labs and software.

2.3.1. Verification

To verify the implementation of the file format in PyDapsys, we read each of the four DAPSYS files (see 2.1) with PyDapsys. The read values were then compared to the CSV files. As the values in the exported CSV files only have limited precision (6 or 4 decimal places, depending on the type of data exported), we first rounded the values read from the file to the same precision before comparing them. Comparison of floating-point values was done by comparing the absolute difference of two values to the system epsilon for 64-bit floating point (f64) values. Numeric values from the CSV were converted to f64 values using built-in Python functions. When comparing f64 with 32-bit floating point (f32) values, the f32 values were first converted to f64. Texts were compared with built-in Python functions.

2.3.2. Performance testing

We also compared the performance (duration and file sizes) of the CSV export of DAPSYS and the export to NIX/H5 using PyDapsys. To achieve comparable measurements, we only looked at the time each system required to write the continuous recording to their respective target format, without the time required for user interactions or loading the data. We had to focus on a single data stream, as DAPSYS would require user interactions in between

exporting multiple streams. We chose to focus on the continuous recording, as it makes up the largest part of a file's size. We also excluded loading times, as there was no reliable way to measure them for DAPSYS. Times for PyDapsys were measured using the wall-clock time directly in the Python program, whereas DAPSYS times were taken by a stopwatch. All measurements were performed on the same system.

3. Results

3.1. Analysis of the DAPSYS file structure

The DAPSYS user interface displays the contents of a file in a hierarchical structure, composed of folders, text streams, and data streams (see Figure 1B). DAPSYS binary files store data in a flat structure that can be split into 4 parts:

1. Header. Files begin with a header with a fixed length. Information in the header is not required to read the file contents.
2. Data Pages. DAPSYS stores data in discontinuous chunks, which we call "pages." All pages have a unique ID in the context of the file and can hold either data of a waveform or textual data.
3. Table of Contents (ToC). After the last data page, the ToC begins. It defines the hierarchical structure shown in the GUI and comprises of folders, which can have additional child elements and streams. Streams contain an array of data page IDs.
4. Footer. After the ToC, there comes a small footer consisting of a string holding the version and the serial number of the DAPSYS program used to create the file.

3.1.1. Data pages

As seen in Figure 2, DAPSYS uses two types of pages: one for waveform data and one for textual data. Both types start with the same fields that store metadata, such as their ID, which is unique among all pages in a file, an identifier for their type (text or waveform), and an optional reference to another page. Waveform pages store the amplitude of the waveform as an array of 32-bit floating point (f32) values, and corresponding timestamps as an array of 64-bit floating point (f64) values. For regularly sampled waveforms, only the first timestamp is saved in the array, while an additional f64 value is used for the regular sampling interval. Text pages are used to store comments as well as sorted spikes. They consist of a string containing the text, and two f64 values. The first f64 value is used to store the timestamp. The second one is used for sorted spikes to indicate the timestamp of the automatically recognized spike. For normal comments, it is set to the same value as the first timestamp. From our observations, DAPSYS writes pages in the order they occur during the recording. If, for example, a comment is entered during a recording, DAPSYS will save the recorded data up to that point in a waveform page, append it to the list of pages followed by the text page containing the comment, and then begin a new waveform page with the new data.

3.1.2. Table of contents

The ToC defines the logical structure of a DAPSYS file. As seen in Figure 2, its elements consist of folders and streams, all of which have an ID unrelated to the IDs used for pages and a string containing their display name. Folders can have several other elements as children. Streams contain multiple fields for storing the configuration of the plot used to visualize their data and most importantly, contain an array of the page IDs belonging to that stream. A stream may either reference text pages or waveform pages, making it a text or data stream, respectively.

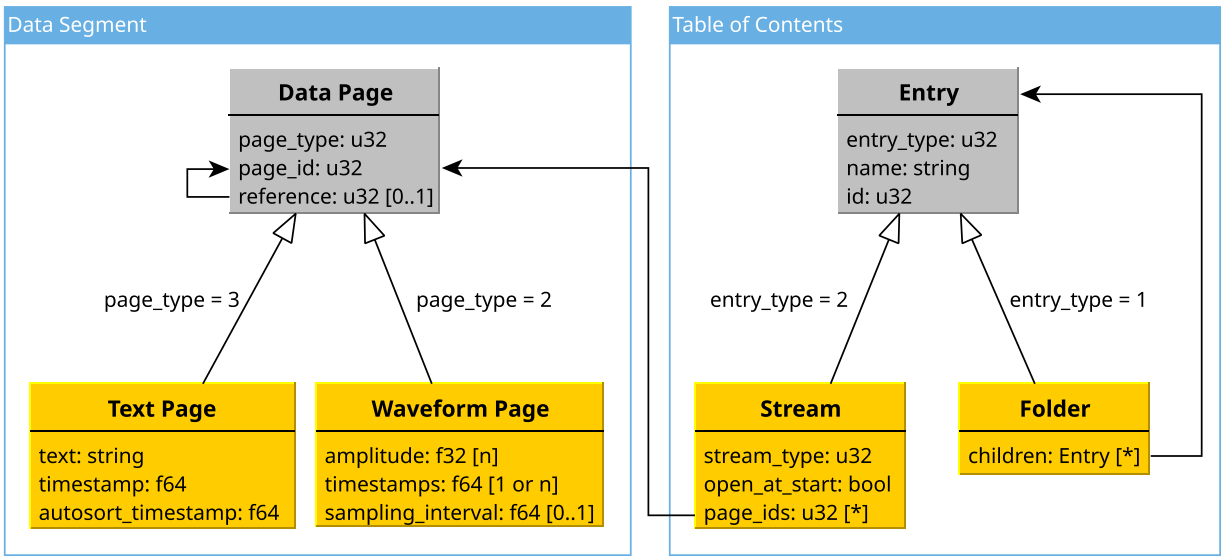


FIGURE 2
Simplified logical model of a DAPSYS file. Black arrows show to which other fields the value refers. Large unfilled arrows indicate that an entry extends another one. Gray types define shared header fields for a group of types, with extending arrows indicating the field and the value to identify the group of fields that follow the header (e.g., a "Folder" and a "Stream" both start with the fields of an generic "Entry." Depending on the value of the field "entry_type," the fields following the "Entry" will be either those of a "Stream" or a "Folder").

3.2. Development of the Python library "PyDapsys"

The functionality of PyDapsys [see (Konradi et al., 2023) for the repository containing the source code. The package is also available on PyPI as "pydapsys"] focuses on accessing data stored in a DAPSYS file. Pages are read into a dictionary that maps the page IDs to an object storing the metadata (type of the page, ID, optional ID of the referenced page) and data, i.e., text and timestamps for text pages of the corresponding page. The ToC is represented by folder and stream objects. The folder objects offer dictionary-like access to their children, while stream objects store the IDs of the pages belonging to them. The library uses NumPy (Harris et al., 2020) to improve the reading speed and memory efficiency of the arrays storing page IDs, amplitudes, and timestamps. To keep the library portable, NumPy is the only required dependency. The functionality to convert a recording to the Neo structure is implemented as an optional dependency. As different experiment set-ups may produce different structures in the DAPSYS file, there is no "universal" converter. Instead, the library provides an abstract base class for Neo converters, which offers functions for common conversions (i.e., text stream to event). Based on this class, additional converters may be implemented for different ToC structures.

3.2.1. Verification

As described in section 2.3.1, we compared the CSV data exported by DAPSYS with the data read by PyDapsys. Depending on the type of stream being exported to CSV, the resulting file contains different values:

- Waveform streams: Contain both the timestamps for each data point and its signal value. Both timestamp and signal values have a precision of 6 decimals.
- Text streams: Contain the timestamps for each text with a precision of 4 decimals and the text itself.

Across all files used for testing, 284,453,786 individual floating-point values were compared, of which 3,009,074 values differed. The maximum difference was 0.00001. As this is exactly the precision offered by waveform CSV-exports, it is most likely a result from rounding errors and not a systemic error in the PyDapsys implementation. There were no differences in the text data.

3.2.2. Performance testing

As seen in Table 1, storing data in NIX/H5 with Neo had no significant impact on file sizes compared to the original file, whereas the CSV increased the file size by factor 4. PyDapsys reliably outperformed DAPSYS in the time required for exporting a file by more than factor 97 (see Table 2).

4. Discussion

In order to make electrophysiological recordings obtained with the DAPSYS DAS available to other systems in our lab, we implemented the open-source Python library "PyDapsys." The library has functionality for reading data from DAPSYS files and offers built-in functions to automatically load read data into the structure defined by the Neo library, from where it can be exported to NIX and other data formats, which are used by the neuroscience community and can be read by various other software solutions. By offering direct access to the data stored in DAPSYS files, rounding errors that may occur when exporting the data to CSV are avoided, thus improving the accuracy and quality of subsequent analyses. The library outperforms the DAPSYS CSV export, both in export duration and size of the exported files, while additionally not being dependent on DAPSYS itself. Currently, the usage of the PyDapsys library requires a certain level of programming experience. To make the library available for a more general audience, we are working on implementing a GUI (graphical user interface).

While DAPSYS is not used very commonly, it should be seen as a representative of many domain-specific proprietary formats, which are used in neuroscientific research. FAIR data handling principles require the accessibility and interoperability of data, and the opening of proprietary formats is a necessary step to ensure those qualities (Berens and Ayhan, 2019). We expect the presented process of analyzing the files with the "ImHex" software and modifying the parameters to understand their internal structure to be useful for other research groups, who are facing similar challenges. It is important to note that the DAPSYS file format does not utilize any compression or encryption. Reverse engineering compressed or encrypted data would have made the process significantly more difficult.

In general, our case highlights the importance of proper procedures to ensure long-term access to experimental data. In the microneurography community, the experiments are complex, and many data sets are unique due to rare genetic mutations of the patients. Moreover, guaranteeing reliable access and unification of data also simplifies collaboration between research groups. Therefore, ensuring FAIR principles allows us to optimize the research benefit derived from the data.

The appropriate processes should ideally be put in place early on to ensure that data is available in open formats. For example, if the formats cannot be read using open software, this could include manual exporting new data to open formats once a week to avoid forming a backlog and potentially losing access to large quantities of non-exported data if the original software is not available anymore.

Open science and FAIR principles are becoming more and more widely accepted in academia and in neuroscience in particular. However, at the current stage of ongoing works, it is important to include smaller communities in the discussion, as the popularity of the specific software and hardware solution influences the motivation of the vendors to provide open off-the-shelf solutions. PyDapsys alongside more general emerging approaches, such as MARBLE, serves as an example of a possible solution for these research communities.

Data availability statement

The data analyzed in this study is subject to the following licenses/restrictions: the hospital regulations limit open data sharing. Data is available upon reasonable request. Requests to access these datasets should be directed to BN, bnamer@ukaachen.de.

Ethics statement

The studies involving humans were approved by the Ethics Boards of the University Hospital RWTH Aachen and Friedrich-Alexander-University of Erlangen-Nürnberg. The studies were conducted in accordance with the local legislation and institutional requirements. The participants provided their written informed consent to participate in this study.

Author contributions

PK developed the software and drafted the manuscript. AT supervised the work on microneurography data. APG and APM supervised the software-development work. RR, BN, and EK supervised the project. All authors substantially revised the manuscript.

Funding

This work was partially funded by the Excellence Initiative of the German Federal and State Governments G:(DE-82) EXS-SF-SFDdM013 and also supported by the IZKF TN1-6/IA 532006. BN was supported by a grant from the Interdisciplinary Center for Clinical Research within the Faculty of Medicine at the RWTH Aachen University and the German Research Council DFG NA 970 3-1, DFG FOR 2690 project 6.

Acknowledgments

The authors would like to thank Abigail Morrison for her support and many insightful discussions. Also, they would like to thank Dagmar Krefting for the discussion on interoperability in biosignals.

Conflict of interest

APM was employed by Forschungszentrum Jülich GmbH.

The remaining authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

References

Ackerley, R., and Watkins, R. H. (2018). Microneurography as a tool to study the function of individual C-Fiber afferents in humans: responses from nociceptors, thermoreceptors, and mechanoreceptors. *J. Neurophysiol.* 120, 2834–2846. doi: 10.1152/jn.00109.2018

Berens, P., and Ayhan, M. S. (2019). Proprietary data formats block Health Research. *Nature* 565:429. doi: 10.1038/d41586-019-00231-9

Garcia, S., Guarino, D., Jaillet, F., Jennings, T., Pröpper, R., Rautenberg, P. L., et al. (2014). Neo: an object model for handling electrophysiology data in multiple formats. *Front. Neuroinform.* 8:10. doi: 10.3389/fninf.2014.00010

Grewe, J., Wachtler, T., and Benda, J. (2011). A bottom-up approach to data annotation in neurophysiology. *Front. Neuroinform.* 5:16. doi: 10.3389/fninf.2011.00016

Harris, C. R., Jarrod Millman, K., Van Der Walt, S. J., Gommers, R., Virtanen, P., Cournapeau, D., et al. (2020). Array programming with NumPy. *Nature* 585, 357–362. doi: 10.1038/s41586-020-2649-2

Kemp, B., Värri, A., Rosa, A. C., Nielsen, K. D., and Gade, J. (1992). A simple format for exchange of digitized Polygraphic recordings. *Electroencephalogr. Clin. Neurophysiol.* 82, 391–393. doi: 10.1016/0013-4694(92)90009-7

Kilamo, T., Hammouda, I., Mikkonen, T., and Aaltonen, T. (2012). From proprietary to open source—growing an open source ecosystem. *J. Syst. Softw.* 85, 1467–1478. doi: 10.1016/j.jss.2011.06.071

Konradi, P., Troglio, A., Namer, B., and Kutafina, E. (2023). Digital-C-Fiber/PyDapsys. *Zenodo*. doi: 10.5281/ZENODO.7970520

Kutafina, E., Troglio, A., De Col, R., Röhrig, R., Rossmanith, P., and Namer, B. (2022). Decoding neuropathic pain: can we predict fluctuations of propagation speed in stimulated peripheral nerve? *Front. Comput. Neurosci.* 16:899584. doi: 10.3389/fncom.2022.899584

Rübel, O., Tritt, A., Ly, R., Dichter, B. K., Ghosh, S., Niu, L., et al. (2022). The Neurodata without Borders ecosystem for neurophysiological data science. *eLife* 11:e78362. doi: 10.7554/eLife.78362

Schlebusch, F., Kehrein, F., Röhrig, R., Namer, B., and Kutafina, E. (2021). openMNGlab: data analysis framework for microneurography – a technical report. *Stud. Health Technol. Inform.* 283, 165–171. doi: 10.3233/SHTI210556

Stoewer, A., Kellner, C., Benda, J., Wachtler, T., and Grewe, J. (2014). File format and library for neuroscience data and metadata. *Front. Neuroinform.* 8:15. doi: 10.3389/conf.fninf.2014.18.00027

Troglio, A., Schlebusch, F., Röhrig, R., Dunham, J., Namer, B., and Kutafina, E. (2023). odML-tables as a metadata standard in microneurography. *Stud. Health Technol. Inform.* 302, 368–369. doi: 10.3233/SHTI230144

Torebjork, H. E., and Hallin, R. G. (1974). Responses in Human A and C Fibres to Repeated Electrical Intradermal Stimulation. *J. Neur. Neurosurgery Psych.* 37, 653–64. doi: 10.1136/jnnp.37.6.653

Vallbo, Å. B., and Hagbarth, K.-E. (1968). Activity from skin mechanoreceptors recorded percutaneously in awake human subjects. *Exp. Neurol.* 21, 270–289. doi: 10.1016/0014-4886(68)90041-1

Wilkinson, M. D., Dumontier, M., Aalbersberg, I. J., Appleton, G., Axton, M., Baak, A., et al. (2016). The FAIR guiding principles for scientific data management and stewardship. *Sci. Data* 3:160018. doi: 10.1038/sdata.2016.18

frontiers | Frontiers in Neuroinformatics

TYPE Review
PUBLISHED 05 January 2024
DOI 10.3389/fninf.2023.1276407

OPEN ACCESS

EDITED BY
Maaike M. H. Van Swieten,
Netherlands Comprehensive Cancer Organisation (IKNL), Netherlands

REVIEWED BY
Leonardo Candela,
National Research Council (CNR), Italy
Alexandre Rosa Franco,
Nathan Kline Institute for Psychiatric Research, United States

*CORRESPONDENCE
Maryann E. Martone
✉ mmartone@ucsd.edu

RECEIVED 11 August 2023
ACCEPTED 31 October 2023
PUBLISHED 05 January 2024

CITATION
Martone ME (2024) The past, present and future of neuroscience data sharing: a perspective on the state of practices and infrastructure for FAIR.
Front. Neuroinform. 17:1276407.
doi: 10.3389/fninf.2023.1276407

The past, present and future of neuroscience data sharing: a perspective on the state of practices and infrastructure for FAIR

Maryann E. Martone[1,2]*

[1]Department of Neurosciences, University of California, San Diego, CA, United States, [2]San Francisco Veterans Administration Hospital, San Francisco, CA, United States

Neuroscience has made significant strides over the past decade in moving from a largely closed science characterized by anemic data sharing, to a largely open science where the amount of publicly available neuroscience data has increased dramatically. While this increase is driven in significant part by large prospective data sharing studies, we are starting to see increased sharing in the long tail of neuroscience data, driven no doubt by journal requirements and funder mandates. Concomitant with this shift to open is the increasing support of the FAIR data principles by neuroscience practices and infrastructure. FAIR is particularly critical for neuroscience with its multiplicity of data types, scales and model systems and the infrastructure that serves them. As envisioned from the early days of neuroinformatics, neuroscience is currently served by a globally distributed ecosystem of neuroscience-centric data repositories, largely specialized around data types. To make neuroscience data findable, accessible, interoperable, and reusable requires the coordination across different stakeholders, including the researchers who produce the data, data repositories who make it available, the aggregators and indexers who field search engines across the data, and community organizations who help to coordinate efforts and develop the community standards critical to FAIR. The International Neuroinformatics Coordinating Facility has led efforts to move neuroscience toward FAIR, fielding several resources to help researchers and repositories achieve FAIR. In this perspective, I provide an overview of the components and practices required to achieve FAIR in neuroscience and provide thoughts on the past, present and future of FAIR infrastructure for neuroscience, from the laboratory to the search engine.

KEYWORDS

data sharing, neuroinformatics, data bases, FAIR (findable accessible interoperable and reusable) principles, data management, INCF

Introduction

The transformation of neuroscience from a closed to an open science, where the entirety of research products like data and code produced during a study are routinely made available, has accelerated in recent years. Data sharing requires that the necessary human and technical infrastructure be in place to make these data broadly available. The first Human Brain Project,

funded by the US National Institute of Mental Health in the 1990s, launched some of the first efforts to "database the brain," envisioning a "paradigm shift in which primary data are openly shared with the worldwide neuroscience community" (Koslow, 2000). Despite this early optimism, neuroscience had a rocky history with open data sharing. Unlike the genomics and structural biology communities where the mechanisms and value of sharing primary sequence and structural data were agreed upon fairly early, the how and why of sharing the more diverse and complex data types of neuroscience was met with early resistance (Whose Scans Are They, Anyway?, 2000). In these early days, before the spotlight was shown on reproducibility problems facing neuroscience (Ioannidis, 2007; Button et al., 2013) and before "big data" became a buzzword in neuroscience and across biomedicine, there were few motivations or incentives for researchers to share their data openly. Like other areas of biomedicine (Nelson, 2009), neuroscience archives were largely underpopulated relative to the amount of data generated in Table 1 (Ferguson et al., 2014).

Neuroscience started to put its first big stake in the ground for open data sharing with the commissioning of large prospective data sharing efforts where large, comprehensive data sets were collected by large teams of scientists with the goal of making them openly available. Some of early efforts include the Alzheimer's Disease Neuroimaging Initiative (ADNI; Weiner et al., 2010) launched in 2004 and Allen Brain Atlas launched in 2005, followed by large consortia such as the Human Connectome Project (2011) and the Big Brain (2013; Amunts et al., 2013) among many others. The large national and international brain projects launched in the second decade of the 21st century articulated a strong commitment to the open sharing of data and tools. The European Human Brain Project (HBP) was launched in 2013, followed by the US Brain Research through Advancing Innovative Neurotechnologies (BRAIN) Initiative (2014), the Korean Brain Initiative (2016), Canadian Brain Research Strategy (2017), Japan BRAIN/Minds (2018), and the China (2021) and Australian Brain Projects (International Brain Initiative, 2020; Quaglio et al., 2021). These projects have provided a significant infusion of resources to develop the next generation infrastructures necessary to house the sizes and complexity of data developed through new imaging, genomic, and physiological techniques.

An updated analysis of the repositories listed in Ferguson et al. (2014) provides some data on the current state of data sharing. Table 1 shows that data sharing has increased overall, but it is uneven, with explosive growth in some repositories, e.g., NeuroMorpho.org and FigShare, and more modest growth in others. But with the release of the data sharing mandates by funding agencies around the globe (Funders' Policies, 2015; Eke et al., 2022), neuroscience-whether practiced by large consortia or individual labs-is now expected to be "open by default and open by design" (National Academies of Sciences, Engineering, and Medicine, 2018). So the question is no longer whether neuroscience as a whole will share data, it is how effectively? We are seeing some real success stories emerging in neuroscience from the reuse of data, e.g., (Torres-Espín et al., 2021; Almeida et al., 2022) and the ability for multiple groups to analyze the same datasets are providing new insights into notions of reproducibility and robustness (Botvinik-Nezer et al., 2020), but public data are still often difficult to find and use. Effective data sharing, that is, data sharing that views data as a public product of research meant to be reused, referenced, and respected requires the infrastructure, skills, tools, and willingness on the part of the neuroscience community to value data as a research product (Martone and Nakamura, 2022).

Effective data sharing starts with the FAIR data principles (Wilkinson et al., 2016) which grew out of frustrations experienced when trying to use open data on the web in the early days of sharing data. Through the Neuroscience Information Framework (NIF), started in 2008 (Gardner et al., 2008), we were tasked with cataloging all the neuroscience-relevant digital products that were being created (Cachat et al., 2012). NIF was also tasked with developing a strategy to query across the dozens of neuroscience data-and knowledge bases and the 100's of biomedical databases with neuroscience-relevant information that were coming on-line. In these early days of on-line databases, the problems with accessing the data were legion: broken links, insufficient metadata, non-standardized vocabularies and nomenclature, non-actionable data formats, cryptic variables, and proprietary formats to name a few.

FAIR states the minimum set of requirements for digital data for it to be useful: data should be findable, accessible, interoperable, and reusable. FAIR then lays out a set of practices that would make it more likely that data will meet these requirements. The FAIR data principles were formulated in a workshop in Leiden in 2014 (Wilkinson et al., 2016), and were first released through FORCE11, the Future of Research Communications and e-Scholarship. The paper came out 2 years later in 2016. When our group participated in the 2017 kick off meeting for the BRAIN Initiative Cell Census Network (BICCN), a large consortium designed to use multimodal data techniques to determine the major cell types in the brain, few hands were raised when we asked how many people had heard of FAIR. Fortunately, FAIR eventually made its way to neuroscience and found a natural home in the International Neuroinformatics Coordinating Facility (INCF.org), an international organization devoted to developing standards and coordinating infrastructures for neuroscience. INCF incorporated FAIR into its mission statement and has served as a coordinating center for introducing neuroscience to FAIR through its role as a standards organization for neuroscience, its training programs, and other resources (Abrams et al., 2021).

The FAIR partnership

The FAIR acronym itself is now likely better known among practicing neuroscientists, as funders and journals have started to support FAIR in their data sharing policies; but the details of FAIR as elaborated in the detailed recommendations are fairly arcane. Anyone outside the field of informatics is likely to look at these and scratch their head. Persistent identifiers? Knowledge representation languages? A plurality of relevant attributes? Thus, while the practicing neuroscientist may understand what FAIR stands for, they are often at a loss to explain exactly how to achieve it. In reality, no one can create fully FAIR data alone; it requires the interplay of data acquisition and documentation practices, infrastructure, informatics, and community consensus. FAIR is therefore best thought of as a partnership between investigators, data repositories, data aggregators and community organizations (Figure 1). Navigating the landscape of FAIR data sharing and neuroscience infrastructure requires understanding the roles, responsibilities, and interfaces between each of these stakeholder groups. In the following I discuss the different components and some of the tasks required for FAIR and provide information and resources to help navigate the different components required for fully FAIR neuroscience.

TABLE 1 State of population of selected data repositories 2014 vs. 2023.

Resource name	Country / region	Type of data	Date started	Data elements 2014	Update to resource (Feb 2023)	Data elements 2023	Datasets added since 2014	Provenance
NDAR	USA	Demographics, imaging, genetic, phenotypic	2009 (oldest news archives)	>108,000 subjects (from 157 labs)	Now NDA; no longer restricted to autism	–	–	Not comparable as new data types were added
NeuroMorpho.Org	USA	digitally reconstructed neurons	2006	11,335 (reconstructions from 1,339 publications)	Still in existence under same stewardship	298,387 reconstructions 2,103 publications	287,052 reconstructions 764 publications	https://neuromorpho.org/LS_availability.jsp Feb 25 2023
Cell Centered Database/ CIL-Cell Image Library	USA	images, videos, and animations of cell	2002 CCDB/2010 CIL	10,360 image datasets	Still in existence under same stewardship	13,990	3,630	http://www.cellimagelibrary.org/images?k=&simple_search=Search copied number of results Feb 25, 2023
FigShare	International	Various	–	> 8,000 datasets (query: neuroscience)	Still in existence under same stewardship	182,542	174,542	query: neuroscience with dataset filter Feb 252,023
ModelDB	USA	computational neuroscience models	1996	875 available datasets	Same stewardship; transition of leadership	1787	912	https://tinyurl.com/37z5p88f Feb 252,023
Open Source Brain	United Kingdom	Models	2014	47 available datasets	Still in existence under same stewardship	99	52	https://www.opensourcebrain.org/projects
CRCNS	USA	computational neuroscience	2008	38 available datasets	Under same stewardship; not clear if still active	140	102	documented through NIF; Feb 2023
XNAT Central	USA	Neuroimaging	2010	34 available datasets	Will be decommissioned in Oct 2023	510	300	https://central.xnat.org/ project number on home page; accessed Feb 252,023
1,000 Functional Connecto mes Project/IN DI	International (USA, China, Germany, Spain)	fMRI, DTI, MPRAGE, psychological assessements, behavioral phenotype, demographic	2009	28 datasets	Under same stewardship; also 1,000 Functional Connectomes INDI	33	5	
OpenfMRI	USA	fMRI	2012	24 datasets	Under same stewardship; changed name to Open Neuro	805	781	https://openneuro.org/ Feb 26 2023
BIRN	USA	Imaging, histology	–	21 datasets	No longer in service		–	
LONI Image Data Archive	USA	Imaging	–	18 (atlas), 9 databases	Under same stewardship; changed location; hard to compare as atlases and databases are not provided	144	135	https://ida.loni.usc.edu/login.jsp

(Continued)

TABLE 1 (Continued)

Resource name	Country / region	Type of data	Date started	Data elements 2014	Update to resource (Feb 2023)	Data elements 2023	Datasets added since 2014	Provenance
BrainLiner	Japan	ECoG, EEG, fMRI, MEG, Microelect rode, NIRS, Optical Imaging, PET, Other	2011	10 available datasets	Platform there but does not look like it has been updated recently	23	13	http://brainliner.jp/search/showall/1
Open Connecto me Project	USA	Serial electron Microscopy	2011	9 available datasets	Now NeuroData	24	15	https://neurodata.io/project/ocp/ Manually counted Feb 252,023
CARMEN	United Kingdom	neurophysiology	2006	–	No longer in service according to NIF	–	–	
FITBIR	USA	Common data elements	2011	–	Same stewardship	–	–	
INCF Dataspace	International	Various	2012	–	No longer in service	–	–	
UCSF DataShare	USA	biomedical including neuroimaging, MRI, cognitive impairment, dementia, aging	2011	18 datasets	No longer in service	–	–	

Update of Supplementary Table 1 from Ferguson et al. (2014): A sample of Neuroscience centered data repositories available to the community. Only data repositories that accept outside data are included in the update. This table provided the number of data elements (usually equivalent to datasets) in each repository in 2014 (Data elements 2014). We include an update on the status of the resource (Update to Resource Feb 2023 column), the number of data elements found in Feb 2023 (Data elements 2023), the total number added since 2014 (Datasets added since 2014), and how these numbers were derived if the repository did not provide the number of datasets directly. Data repositories that are no longer in service are colored in light orange.

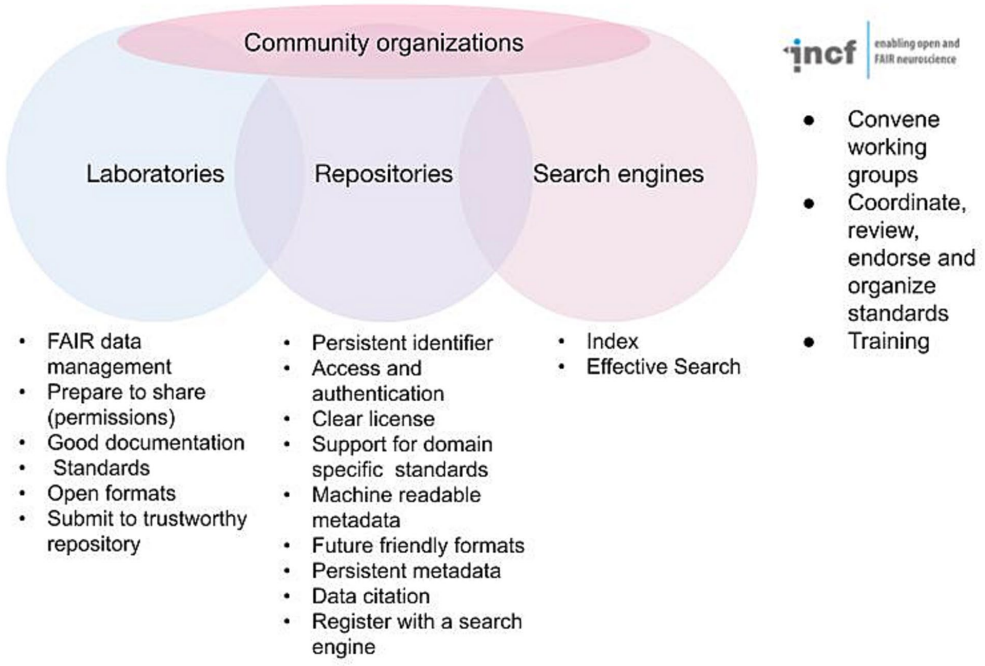


FIGURE 1
Major stakeholders involved in defining and implementing FAIR. Some of the major requirements for achieving FAIR are listed under each stakeholder group. The INCF is given as an example of a community organization supporting FAIR for neuroscience.

TABLE 2 Some FAIR laboratory data management practices.

FAIR goal	Principle	FAIR practices	Reference
Findable	Unique identifiers	1. Create identifiers that are globally unique within the lab for all key entities in the lab, e.g., subjects, experiments, reagents, via the creation of a central registry or use of an existing system, e.g., RRIDs for reagents and tools. Globally unique = no two objects have the same ID, no ID may be reused.	Fouad et al. (2023)
	Rich metadata	Each identifier in the registry is accompanied by rich metadata that provides key details, e.g., for experiments: dates, experimenter, description, collaborators, techniques etc.; for subjects: species/strain, age, weight, etc.	Fouad et al. (2023)
		Use unique identifier for file names, folder names, to label physical objects like slides or slide boxes, so that all entities associated with the lab can be tied unambiguously to metadata	
Accessible	Authentication and authorization	Create a centralized, accessible store for data and code under a lab-wide account for lab data to ensure that files are not scattered around multiple systems or accessible only via personal accounts that may not be available after someone has left the lab.	
Interoperable	FAIR vocabularies	Move away from idiosyncratic naming of variables and annotations towards standards like Common Data Elements and the use of community-based ontologies, atlases, and controlled vocabularies. Consistent lab, wide terminology ensures that lab members can understand what the data are about, and aids in search across and combining files.	Bush et al. (2022)
		Consider creating a lab-wide data dictionary where all variables used across experiments are clearly defined	Bush et al. (2022); Fouad et al. (2023)
Reusable	Documentation	Create a "Read me" file for each dataset where notes can be captured and helpful information provided for reuse of the data	
	Community Standards	All files should be collected and stored in well supported open formats ideally to ensure long term availability.	
		Adopt community standards within the lab where possible; a good place to identify relevant standards is to look at repositories where the data may end up. Specialized repositories usually have a list of required or recommended standards. Some repositories are providing help with developing a data management and sharing plan for grant proposals, e.g., INCF, SPARC and ODC-SCI/TBI.	Bush et al. (2022); FAIRsharing.org, INCF Standards Portfolio
	Provenance	Datasets should be clearly versioned and differences between them documented. Depending on the system used for storing data, formal support for versioning may be available, e.g., Google Docs, but if not, implement a file naming convention so that versions can be tracked	
		Always keep a version of record that can be reverted to if necessary. Often when one is working with data, different versions are created rapidly and it is easy to lose track of which version is which. It is good practice to have stable versions that are easily retrievable so that there are stable points to which to return if provenance is lost.	
		Datasets should also be accompanied by detailed experimental protocols that describe how the data were acquired and computational workflows that detail the processing steps. Use of tools designed for this purpose, e.g., protocols.io, NeuroShapes (Neuroshapes, n.d.) and ReproNIM (Kennedy et al., 2019).	
	Licenses	Prepare to share: Make sure that how and when the data are to be shared is agreed upon with all collaborators early on. For clinical datasets, make sure that the consents are in place for open sharing of de-identified data.	

Examples of laboratory data management practices based on the FAIR principles.

Laboratories

FAIR data management

In the US National Academies of Science, Engineering and Medicine workshop on "Changing the Culture on Data Management and Sharing (Martone and Nakamura, 2022), one of the main takeaways was that the focus of data sharing efforts should not be targeted toward the individual investigator, but the laboratory. As one participant noted: "If you can share data with people in your lab, you are much more likely to have something worthwhile to share outside the lab." FAIR data management is therefore an intentional lab-wide strategy that ensures that data can be shared with lab mates, the PIs, and other colleagues, your future self and eventually with the broader scientific community. Across all stages of the data lifecycle, the management strategy puts in place processes so that data can be found, accessed, combined when necessary, and reused. By paying attention to FAIR in the laboratory throughout the life cycle, benefits start to accrue to the data creator, the laboratory, PI, and collaborators well before data flows out to the wider scientific community (Bush et al., 2022; Dempsey et al., 2022).

Examples of lab management practices built on the FAIR principles are given in Table 2.

We are starting to see neuroscience researchers sharing their experiences with developing and utilizing lab-centric data management systems. They range from tightly integrated digital infrastructures (Bush et al., 2022; Dempsey et al., 2022) to a set of

practices that can be implemented using "off the shelf" components for an average neuroscience wet lab (Fouad et al., 2023).

Choosing a repository

One of the most important steps for a researcher in ensuring that their data is FAIR for the long term is to submit their data to a trustworthy repository that supports FAIR. The new NIH data sharing policy requires researchers to indicate where they will be sharing their data as part of the data management and sharing plan. As recommended in Table 2, knowing in what repository the data will be published allows the researcher to understand what standards are required so they can be built into the laboratory management workflow. With its growing ecosystem of specialized databases, researchers have a choice about where to publish their data.

Understanding how the neuroscience repository landscape is organized may help in finding the right repository. Repositories are generally specialized by data type (Figure 2). However, repositories also exist that are specialized for a domain, e.g., the SPARC database accepts all data associated with the peripheral nervous system, or serve researchers within a particular region, e.g., CONP, or institution, e.g., BrainCode and the Donders Repository. Often, multiple repositories may be appropriate, in which case there are additional features that may make a given repository more or less attractive. These include tool support, curation services, support for data citation, choice of license, size of data allowed, help with data management plans (see Table 2) and possible costs (Murphy et al., 2021). A functioning neuroscience ecosystem also requires open neuroscience repositories that have few restrictions on data types, regions, or subdisciplines to ensure that all data has a home. The EU EBRAIN infrastructure is an example of such a repository, as it takes multiple types of data regardless of discipline or geographical location, although there may be issues with transferring certain types of data across international borders (Eke et al., 2022).

Supplementing the specialist repository landscape are the generalist repositories, data repositories that span scientific disciplines and data types (Assante et al., 2016). These repositories are often useful for publishing smaller supplemental datasets that are required for a publication (Stall et al., 2023). Specialist repositories generally provide more standards, tools and services for harmonizing and using data, and make it easier for researchers to find data of a particular type. To aid researchers in choosing an appropriate neuroscience data repository, the INCF has a searchable infrastructure catalog, where each repository is described according to the checklist developed by Sandström et al. (2022). Other repository finder tools include NITRC for neuroimaging related repositories, re3data, the catalog of open data repositories maintained by the National Library of Medicine, and the NIF listing of BRAIN Initiative Repositories.

Repositories

The central role of community repositories

While the investigator takes the central role in acquiring data in a manner that supports FAIR, the community repository is arguably

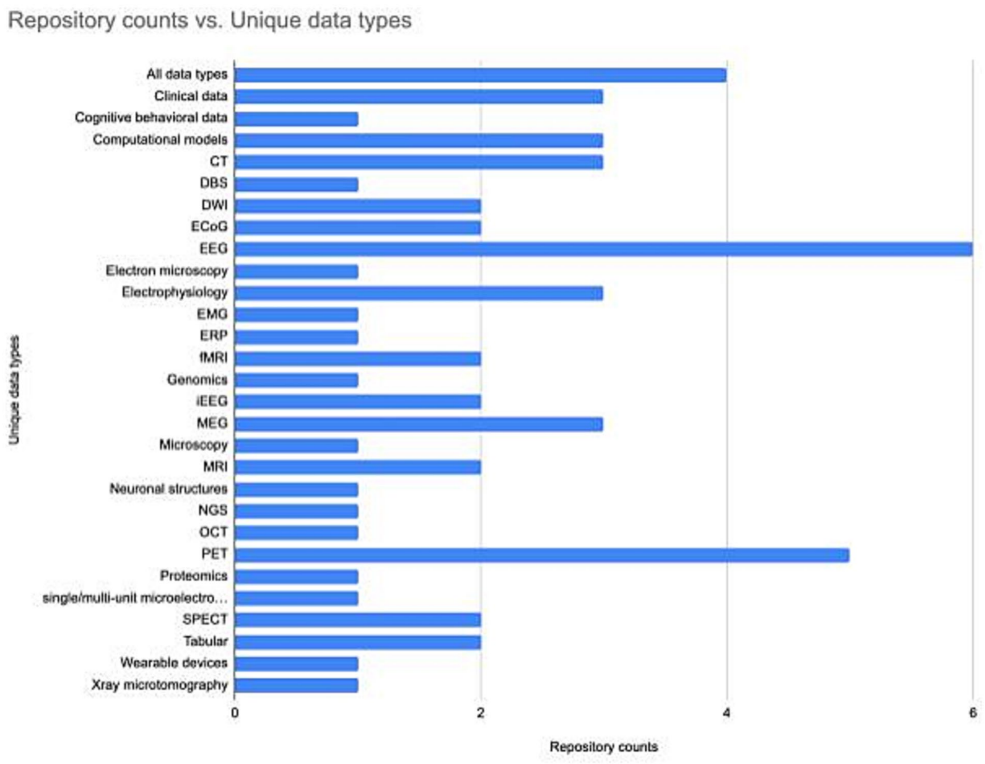


FIGURE 2
The number of neuroscience specialist repositories supporting different data types. The repository list and associated data types was assembled using information available through the INCF Infrastructure Portfolio and the SciCrunch Registry. The data underlying the figure is available at Zenodo, DOI: 10.5281/zenodo.8239845.

the central player in implementing the basic requirements for achieving FAIR for the long term (Figure 1). We are using the term "community repository" here to designate infrastructures that are designed to accept primary data contributed by outside researchers, rather than a single data set produced by a given project (e.g., the Allen Brain Atlas) or a knowledge base that aggregates information about a particular entity (e.g., CoCoMac).[1] As shown in Figure 1, the repositories have critical responsibilities for ensuring that submitted data are made available according to the FAIR principles (Lin et al., 2020). These practices include issuing and maintaining persistent identifiers, tying those identifiers to rich metadata, providing access and any necessary access controls, enforcing or supporting annotation with FAIR vocabularies, enforcing or supporting community standards, supporting data versioning, providing links to other critical products like experimental protocols and code, and provisioning a clear data license for each data set. Repositories also have the critical role of ensuring that data is available for the long term.

From the earliest days of neuroinformatics, it was envisioned that neuroscience would likely best be served by a decentralized system of federated databases (Koslow, 2000). Due to the variety and complexity of neuroscience data, a single large repository like Genbank or the Protein Data Bank was likely not going to be feasible. The early investments in neuroinformatics by the US Human Brain Project and the success of the International Neuroinformatics Coordinating Facility in growing the field of neuroinformatics globally, led to the first generation of neuroscience databases. These databases were largely organized around data type, e.g., structural neuroimaging (XNAT), functional neuroimaging (fMRI Data Center; Open fMRI), neurophysiology (CARMEN; Neurodatabase.org," GNode), EEG (open EEG, iEEG), neuronal morphology (NeuroMorpho), microscopic images (Cell Centered Database), neuromodeling (ModelDB). Some examples are shown in Table 1.

When the first generation of neuroscience databases were started, there were few standard practices for designing web-accessible databases. As documented by NIF, each database had a different mode of access, different data structure, and the use of standards was very limited. It was a time of tremendous technological fluidity, with standard features we take for granted today (e.g., RESTful web APIs) still being invented. The cloud did not exist, and attempts to build resources on the early version of a cloud-like system ("the grid") met with considerable challenges (Grethe et al., 2005). With today's emphasis on data sharing, increased attention is starting to be paid to these critical infrastructures and how they are constructed, operated, and evaluated (Nelson, 2022). Various recommendations on desired characteristics for data repositories have been issued by different groups (Sansone et al., 2020; Shearer, n.d.), including NIH (Selecting a Data Repository) and additional sets of principles, e.g., the TRUST principles (Lin et al., 2020) and principles for open infrastructures (Bilder et al., 2015) have been formulated to help further guide how these critical infrastructures should operate. The Elixir project, a large scale bioinformatics consortium in the EU, has developed a maturity model for evaluating the success of repositories which is designed to be used by funders to determine the criticality of various infrastructures (Bahim et al., 2020). The INCF Infrastructure working group recently issued a set of guidelines from a neuroscience perspective, that provide a mix of technical and "customer service" recommendations for operating repositories (Sandström et al., 2022). Although these various lists of desiderata do not overlap completely (Murphy et al., 2021), over time we will likely converge on a core set of functions and expectations for these critical infrastructures, balancing the often dual requirement for these infrastructures to serve as both publishing platforms and dynamic scientific gateways (Sandström et al., 2022).

INCF has served as an important conduit by which the FAIR principles have permeated the construction of neuroscience data repositories and gateways. Investigators who have been active in INCF through governance, committees and working groups are involved with several of the next generation neuroscience infrastructures including EBrains, CONP, SPARC, DANDI, Open Neuro, and BRAIN/Minds. Table 3 lists and compares some of the key ways that these infrastructures implement FAIR and "FAIR-adjacent" practices. Following consistent design principles that support FAIR provides a level of common functionality and services that make it easier to work across these databases for an individual user or an automated agent. The more similar FAIR practices are across repositories, the more likely it is that the repositories themselves are interoperable.

Standards: role of repositories

A significant and positive change that is accelerating progress toward FAIR is the emergence of a set of robust standards for neuroscience data types that are starting to gain adoption. The INCF was created to help with this process of standardization and produced some early successes, e.g., the Waxholm space for registration of mouse and rat brain data (Hawrylycz et al., 2011; Papp et al., 2014), the Neuroimaging Data Model (Keator et al., 2013) the Brain Imaging Data Structure (Gorgolewski et al., 2016) were produced with support from INCF. Over the last few years, a set of standards has emerged for major neuroscience data types that can accommodate the increased size and complexity of neuroscience data through additional investments by funders and the efforts of the large brain projects, e.g., NWB, 3D-MMS (Ropelewski et al., 2022). Repositories serve as important stakeholders in ensuring that standards are followed by supporting or requiring them for data submission (Figure 2). Data uploaded to OpenNeuro, for example, must be validated against BIDS before it is accepted. The INCF has implemented an open community review and endorsement process to help improve the quality, usability, interoperability and awareness of these standards (Abrams et al., 2021). They have made available a searchable Standards and Best Practices Portfolio[2] where researchers can learn about each standard and how it can be used. FAIRsharing.org more broadly aggregates standards from across biomedicine and makes them available through a searchable catalog.

As neuroscience standards become more mature, better supported, and more widely used, they provide the seeds for knitting the landscape of neuroscience data repositories into a true data ecosystem, where (meta)data can flow from the laboratory to repositories and from repositories to computational tools and back

1 http://cocomac.g-node.org/

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

TABLE 3 FAIR practices across data repositories.

Principle	Function	EBRAINS	SPARC		DANDI	CONP Portal	OpenNeuro
F1. Globally unique identifier	Basic core	DOI	DOI		DOI	ARK, DOI	DOI
F2. Rich metadata		Y	DataCite		Y	DATS	Y
A1. Retrievable by identifier		Y	Y		Y	Y	Y
A1.1 Free, open, universal retrieval protocol	Enhanced access	Y	Y		Y	Y	Y
F4. Registered in a searchable resource		KS, GDS	KS, GDS		KS, GDS	KS	KS, GDS
A1.2: Authentication and authorization		Y	Y		Y	Y	Y
R1.1: Clear data usage license		Y	CC-BY		CC-BY, CC0	Y	CC0
R1.3: Community standards	Use of standards	Multiple	SDS, MIS		NWB, BIDS	Y*	BIDS
F3: Metadata contains identifier		Y	Y		Y	Y	Y
I1: Formal knowledge representation language		Y	Y		N	Y	
R1: Plurality of relevant attributes	Rich(er) metadata	OpenMinds	OpenMinds, MIS		NWB	DATS	Y
I2: FAIR vocabularies		Y	Y		Y	Y	N
I3: Qualified references to other metadata		Y	Y		Y	Y	Y
R1.2: Provenance	Provenance and context		Exp Protocol			Y	N
A2: Metadata persistence			Y		Y		
Landing page	Additional features	Y	Y		Y	Y	Y
CCFs		Y	Y*		N	N	N
Data citation		Y	Y		Y	Y	Y
Curation		Y	Y		N	Y	N

Comparison of FAIR features across five large brain repositories where the principal investigators have been active through the INCF. The principles are organized according to the functions they support based on an organization proposed by Hodson et al. (2018). Highlighted in purple are additional features that are relevant for FAIR although they are not mentioned explicitly in the FAIR principles, e.g., the use of landing pages and support for data citation. KS, INCF Knowledge Space; GDS, Google Dataset Search; DOI, Digital object identifier; NWB, NeuroData Without Borders; BIDS, Brain Imaging Data Structure; DATS, Data tag suite.

again. Figure 3 shows a graph illustrating the connections between standards (light gray) and infrastructures that support them (dark gray). The data was assembled from the INCF Infrastructure Catalog, FAIRsharing, the SciCrunch Registry (Subash et al., 2023) and examination of repository websites. As shown in Figure 3, multiple repositories and infrastructures are connected via these standards. For example, the Brain Imaging Data Structure (BIDS; Gorgolewski et al., 2016) links 10 different repositories and computational platforms. The success of BIDS has led to extensions of BIDS for other modalities through a formal governance process (Governance, n.d.). The adoption of these BIDS-based standards starts to create a degree of interoperability across data types.

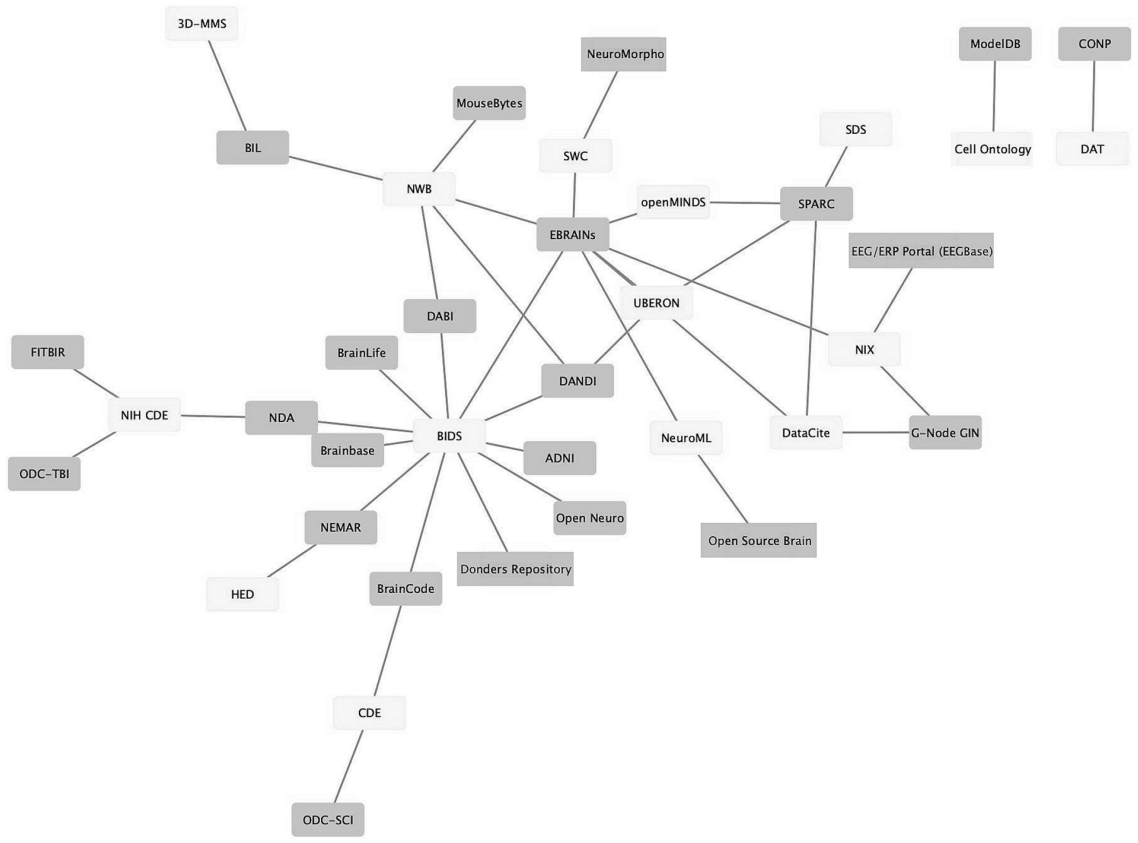


FIGURE 3
Ecosystem of neuroscience resources emerging around standards. Network graph of neuroscience data repositories and gateways (purple) and some of the standards they support (yellow). The graph shows repositories/gateways connected via the use of a common standard. A description of how standards were determined is given in the text.

As tool support grows, standards are also making their way into the laboratory. BIDS, for example, has been estimated to have been used to organize over 100,000 datasets containing millions of images, indicating significant uptake by the research community (Poldrack et al., 2023). In a recent paper that outlined a neuroimaging center's implementation of BIDS, Bush et al. (2022) stated: "*Learning the BIDS specification, implementing software pipelines to map the data, and validating that the resultant mappings met the BIDS standard consumed many months of effort across multiple imaging center team members… The benefits of mapping our data to BIDS, however, far exceed the costs.*" (Bush et al., 2022). These benefits included access to BID-APPS, a set of containerized analysis tools and pipelines that run on validated BIDS data, as well as improved code sharing within the lab and with colleagues, as well as a reduced barrier to publishing the data in OpenNeuro. Similarly, the electrophysiology standard, NWB, has made inroads in tackling one of the most challenging data types in neuroscience, evidenced by uptake in laboratories (Rübel et al., 2022) and support by platforms such as DANDI and EBRAINs.

Standards: use of FAIR vocabularies and common coordinate frameworks

Interoperability across neuroscience data has always been hampered by the multiplicity of nomenclatures and parcellation schemes from brain regions and nerve cells (Martone et al., 2004). Although slow, progress has been made. Some repositories are starting to map generic neuroanatomical structures to community ontologies like UBERON (Mungall et al., 2012). Mapping data to a common coordinate framework (CCF) allows more precise localization independent of labels applied to them (Hawrylycz et al., 2023). Encouraging signs are emerging, as CCFs for multiple species are in use or in development for the major species across the international brain projects. For example, both the BICCN/BICAN and EBrains are utilizing the Allen Institute Common Coordinate Framework v3 for mouse (Hawrylycz et al., 2009, 2023). To help manage the different versions and components that go into these atlas-based environments, a new standard for describing and versioning brain atlases was recently proposed (Kleven et al., 2023).

Standardized nomenclature for cellular taxonomies and transcriptionally defined cell types are also emerging from projects like the BICCN/BICAN to help deal with the plethora of new cell types that are emerging from new transcriptomics-based approaches (Miller et al., 2020; Tan et al., 2023). Over the years, there have been proposals for naming neurons that can bridge the multiplicity of phenotypes generated by multiple experimental techniques (Hamilton et al., 2012; Shepherd et al., 2019; Gillespie et al., 2022). However, these approaches have had difficulty in handling the complex expression patterns coming out of transcriptomics. The BICCN/BICAN recently developed the Brain Standards Data Ontology, providing a model for providing data-driven definitions of taxonomic classes (Hawrylycz et al., 2023; Tan et al., 2023). BICCN has recently introduced Cell

Cards to provide a tool for exploring the BICCN taxonomic cell types for human, marmoset, and mouse primary motor cortex, including linking them to primary data sets (Hawrylycz et al., 2023). As new technologies are allowing us to derive wider scale, more complete representations of the molecular, morphological, physiological, and connectional phenotypes of neurons than was possible in the past, it is time for the global neuroscience community to come together around a common nomenclature for naming populations of cells that will aid in comparison across studies.

Services for accessing ontologies and building them into annotation and metadata pipelines have improved significantly over the past decade, with tools such as BioPortal[3] and the Ontology Look Up Service[4] providing programmatic access to community ontologies. Nevertheless, neuroscience is still a cutting edge science where many new terms are needed, particularly for annotating experimental data. For this reason, NIF and INCF had developed the NeuroLex Wiki (Larson and Martone, 2013) that lowered the barrier for creating new ontology terms. When the semantic wiki technology underlying NeuroLex was no longer available, the approach and content were ported to the Interlex on-line vocabulary management system by NIF (Surles-Zeigler et al., 2021). Interlex mints a unique identifier for each term (URI) when it is entered and allows the addition of basic metadata for each term, e.g., definition, synonyms. It also provides basic knowledge engineering functions, e.g., parent–child and other relationships, annotations. Interlex also provides various review and curation functions. These specialized terms can be used as controlled vocabularies or further engineered into ontologies as needed. Surles-Zeigler et al. (2021) provide a description of how Interlex is being used to enhance anatomical annotation of SPARC data, models and knowledge base, allowing new anatomical terms to be minted, curated, linked to existing ontologies and contributed as necessary to augment community ontologies.

On the sustainability of neuroscience data repositories

As most neuroscience infrastructure is researcher-led and grant-supported, questions often arise about long-term sustainability when choosing a repository, or indeed, any infrastructure. Sustainability of individual resources remains a challenge, not just for neuroscience but for all research-led infrastructures that rely on grant funding for their operation. Of the data repositories listed in Table 1 taken from Ferguson et al. (2014), 4/18 are no longer in service and 3/18 are moribund (i.e., not taking data). Three were rebranded and expanded their scope, and one merged with another database. The good news is that the majority of this first generation of neuroscience databases are still in existence, indicating a degree of stability. We can also see movement in the ecosystem, with databases merging with others, or moving across institutions indicating a degree of dynamism that keeps the ecosystem healthy. Looking at a larger sample using the SciCrunch Registry (formerly the NIF Registry; Ozyurt et al., 2016) out of a total of 563 neuroscience data resources (including data repositories, databases, data sets, atlases and knowledge bases), 71 appear to be out of service (~13%). These numbers compare favorably to a study done on the longevity of bioinformatics biological databases founded in the late 20th century, 63% of which were defunct by 2015 (Attwood et al., 2015). In 2016 NIF began to track the usage of these neuroscience resources within the scientific literature (Ozyurt et al., 2016), revealing interesting patterns including the creation of thousands of data repositories across biomedicine. A recent analysis showed that only a handful of these repositories are actively used, with many of the neuroscience repositories referenced here among them, suggesting that neuroscience is coalescing around a set of core resources (Piekniewska et al., 2023). Thus, while sustainability is always a concern, neuroscience repositories have generally been good stewards of their data, utilizing a variety of strategies to keep data safe and accessible.

As neuroscience data and repositories start to align around the FAIR principles, the ecosystem should become more robust as it will make it easier for other repositories to absorb data if a repository loses its funding. Merging of similar resources also makes the ecosystem more efficient. The 'professionalization" of scientific data repositories also means that researchers are taking their role as an archive more seriously. The INCF recommendations for neuroscience infrastructure include that repositories should have an exit plan and they should clearly state their persistence policy (Sandström et al., 2022). For example, some repositories are partnering with institutional libraries or other resources to ensure that data remain available, even if funding is lost (e.g., EBRAINS). Another promising development is the repurposing of infrastructure components. Rather than building a separate data repository, two computational and analytic platforms, Brainlife and NEMAR, utilize Open Neuro as their data platform, even as they field their own portals with their own branding. The ODC-SCI and ODC-TBI share the same infrastructure (SciCrunch; Surles-Zeigler et al., 2021), but each have their own separate community portal where they can access data and establish their own governance rules. The more that neuroscience infrastructure can be repurposed for new projects, the less funding needs to go to building and maintaining new infrastructures.

Search engines

In tandem with the vision of a distributed system of databases laid out by the NIH HBP was the creation of a neuroscience portal where data could be accessed via a *"a smart 'neuroscience browser' instructed to look for a particular variable or set of variables and import the data back to the user's computer"* (Koslow, 2000). For the distributed ecosystem to work effectively, users would have to be able to issue dynamic queries across these databases and be able to retrieve the necessary subsets of data. And, in fact, FAIR states that data should be registered with an appropriate index (F4). NIF set up one of the first searches across neuroscience databases by creating an index over the contents of distributed databases. At its height, NIF queried over 200 data sources across biomedicine comprising over 8 million data records (Cachat et al., 2012). NIF used the NIFSTD to help mediate across the different vocabularies and relationships that were needed to link across databases. NIF was able to align different databases covering the same content

3 https://bioportal.bioontology.org/

4 https://www.ebi.ac.uk/ols/index

across a core set of variables, but did not have the resources to harmonize the content, especially given the lack of standards at that time. NIF was designed to allow researchers to understand what was in a given database by providing limited views of the data, but not to perform deep structured queries of the content. So you could use NIF to identify a database that had relevant data, but for more structured queries and to retrieve the complete data, users needed to visit the source database. The INCF Knowledge Space and currently performs a similar type of search over 16 major neuroscience databases (KnowledgeSpace, n.d.).

The more that repositories enforce consistent standards for metadata and data formats, the closer neuroscience gets toward achieving true federated search and retrieval across the entirety of the neuroscience repository ecosystem (Koslow, 2000). The Canadian Open Neuroscience Portal was recently launched that allows users to search across data hosted in multiple data repositories. It is currently deployed across 17 Canadian institutions and also integrates select specialist and generalist repositories. All the high level metadata is aligned to the DATS standard, developed by the NIH-funded BioCADDIE Big Data to Knowledge project (Alter et al., 2020), allowing for a unified dataset search. The portal implements some uniform functions that can be executed directly from the portal. Some data are available for download via DataLad and containerized workflows that work across these distributed data are available via Boutiques (Poline et al., 2023).

New tools are also becoming available that lower the barrier to making content available to search engines. For example, multiple neuroscience databases have marked up their content with schema.org so that their datasets are searchable through Google Dataset Search (Table 3). Neuroscience, like other domains, is building knowledge graphs that link neuroscience concepts to each other and to datasets to aid in search.EBrains, CONP and the SPARC projects are making their data available via a knowledge graph. CONP uses the Nexus knowledge graph developed by the Blue Brain Projects which provides a set of tools and resources for searching, linking and viewing data.[5]

Community organizations

The FAIR data principles delegate a good amount of responsibility to individual communities to define what is FAIR for their domain. Community organizations play an important role as coordinators by serving as conveners to allow researchers to come to consensus about best practices and recommendations for their community. International neuroscience is currently supported by two community organizations, the INCF and the IBI. IBI is principally focused on coordination of the large international brain projects, focusing on data sharing among these projects, as well as issues such as data governance and ethics. INCF works across all neuroscience efforts, whether individual or team based, and focuses on standards, infrastructure coordination and training. Both organizations provide support for working groups that come together to tackle issues such as the development of international data governance (IBI), standards and best practices (INCF, IBI), training (INCF), and coordination of infrastructures (INCF, IBI). Any member of INCF can propose a working group and membership is open to the community, while IBI working groups are set by the Strategy Committee. The two organizations work together and with other organizations such as the IEEE Neuro Standards working group and the Global Brain Consortium.[6] In this way, there is a level of coordination across these international organizations. Eke et al. (2022) raised the issue of whether neuroscience needs an umbrella organization modeled after the Global Alliance for Genomic Health, to more effectively address data reuse at the technical, ethical, sociological and political level.

Is neuroscience FAIR yet?

Neuroscience has made tremendous progress over the first two decades of the 21st century in establishing the infrastructure, standards, expertise and tools for moving neuroscience significantly toward FAIR. It is now served by a set of robust international data repositories and scientific gateways specialized for neuroscience data, implementing the vision laid out in the dawn of neuroinformatics for a distributed ecosystem of repositories. The first inroads have been made in establishing FAIR practices and supporting infrastructure in the lab to manage data in a way that smooths the transition between private, semi-private, and public sharing. As best practices for FAIR are articulated, tested, and shared, we can expect that the quality of both the databases and the data will continue to improve.

A federated system allows neuroscience infrastructure to respond more rapidly to new data types and technologies as they are developed. While there are more resources to be sustained, there are also more resources from which to draw should a repository need to be decommissioned. We see from the last 20 years that there is movement in the repository landscape, with some resources ceasing operations, but others merging or changing ownership. As repositories start to align around sets of core features, both interoperability and flexibility will be increased, providing some measure of stability in an otherwise dynamic ecosystem.

While the distributed nature of neuroscience infrastructure brings many benefits, there are concomitant challenges it imposes on both those who submit their data and those that wish to use it. As the motivations and incentives for these two user groups can differ (Subash et al., 2023), balancing the efforts required to submit vs. reuse data will need to be a priority. Until these are addressed, neuroscience will not be considered a fully FAIR discipline:

- **Findable**: We still do not have an effective query system over the ecosystem of neuroscience data, that allows for aggregation relevant data distributed across multiple repositories. Important steps have been taken by IBI, INCF and CONP, but these efforts will need support if they are to be fully realized.
- **Accessible**: Users are increasingly acquiring multimodal datasets that may require deposition in multiple repositories.

5 https://bluebrainnexus.io/

6 https://globalbrainconsortium.org/

Currently, that requires a user to navigate multiple repositories, set up multiple accounts, entering the same metadata repeatedly and creating the necessary linkages across the different parts of the dataset (Subash et al., 2023). Some work is underway in the US BRAIN Initiative BICCN and BICAN projects to create a more unified workflow including a centralized registry, but such a service would be useful across all neuroscience. Many repositories are starting to implement login and authorization via ORCID, making it easier for users to work across multiple repositories.

- **Interoperable**: In a distributed system, interoperability is not just about the data but also about the infrastructures. Working across multiple repositories means working across multiple front ends, back ends and data access policies. As core sets of features are described for data repositories, neuroscience infrastructure may also start to converge on certain design patterns that make it easier for users to work across them. A term was introduced in an NIH Workshop on a FAIR Data Ecosystem for Generalist Repositories: coopetition (NIH workshop on the role of generalist repositories to enhance data discoverability and reuse: Workshop summary, 2012). Repositories can compete on certain features to encourage innovation, but there should be a set of features that are shared across repositories and work similarly.
- At the same time, competition among different data providers also can lead to a decrease in data interoperability, as repositories must compete for users. Thus, many repositories lower their requirements for standards compliance (Subash et al., 2023) recommending rather than requiring standards so as to lower the barrier of data submission. Instead of making compliance optional, neuroscience repositories should work on improving their customer service, providing both human and tool support to make it easier for researchers to comply with standards. SPARC has taken this approach, employing customer-oriented curators who assist researchers to comply with SPARC standards. SPARC also developed the SODA tool directed toward researchers with few computational skills to guide and support them in organizing and uploading their files according to the SPARC SDS (Bandrowski et al., 2021). In this way, the burden on the submitter is lessened, while data quality and standards compliance are not sacrificed.
- **Reusable**: Despite FAIR, most neuroscience data is still very difficult to use. Different projects have devoted different amounts of resources to curation of data and quality control. Generally curated data is of higher quality because it is more completely documented and some QC is performed (Gonçalves and Musen, 2019). Particularly with the push to make data AI/ML ready, funders should be prepared to support curation services for the near future, to ensure that high quality data are available. Such investments will likely not be needed forever; indeed, labs are at this moment experimenting with tools such as ChatGPT to help with query and harmonization. However, investments now in well curated data can help to accelerate training of these types of algorithms, while at the same time, making high quality data immediately available for discovery science.

Finally, usability is not simply a matter of technology or documentation. As Eke et al. (2022) and (Fothergill et al., 2019) have noted, the international nature of neuroscience infrastructure also means that issues of transferring data across national borders, i.e., international data governance, also must be addressed. Federation across distributed databases provides a model that can minimize data governance issues, as the data can remain in place, while compute is brought to the data (Poline et al., 2023).

The good news is that routine data sharing, if not exactly easy, is now at least possible across the sizes and complexities of neuroscience data. Islands of interoperation are starting to emerge among these different resources promoting federated search and shared computational platforms and services. Those of us who were involved from the beginning in attempts to "database the brain" cannot help but be impressed with how far neuroscience sharing and infrastructure has come, even as there is still quite a way to go. As the paradigm continues to shift toward open and effective data sharing in neuroscience, we will fulfill the early vision of neuroinformatics as a driver for *"..a new depth of understanding of how the nervous system works in both health and disease."* (Koslow, 2000).

Author contributions

MM: Writing – original draft, Writing – review & editing.

Funding

The author(s) declare financial support was received for the research, authorship, and/or publication of this article. MM is supported by grants from NIH Office of the Director OT2OD030541 for the SPARC Knowledge Management and Curation Core and the US BRAIN Initiative grant U24MH130919.

Acknowledgments

I would like to thank my colleagues Anita Bandrowski and Mathew Abrams for their helpful comments.

Conflict of interest

MM is a founder and board member of SciCrunch Inc., which develops tools and services around rigor and reproducibility.

The author(s) declared that they were an editorial board member of Frontiers, at the time of submission. This had no impact on the peer review process and the final decision.

Publisher's note

References

Abrams, M. B., Bjaalie, J. G., Das, S., Egan, G. F., Ghosh, S. S., Goscinski, W. J., et al. (2021). A standards Organization for Open and FAIR neuroscience: the international Neuroinformatics coordinating facility. *Neuroinformatics.* 20, 25–36. doi: 10.1007/s12021-020-09509-0

Almeida, C. A., Abel Torres-Espin, J., Huie, R., Sun, D., Noble-Haeusslein, L. J., Young, W., et al. (2022). Excavating FAIR data: the case of the multicenter animal spinal cord injury study (MASCIS), blood pressure, and neuro-recovery. *Neuroinformatics* 20, 39–52. doi: 10.1007/s12021-021-09512-z

Alter, G., Gonzalez-Beltran, A., Ohno-Machado, L., and Rocca-Serra, P. (2020). The data tags suite (DATS) model for discovering data access and use requirements. *GigaScience* 9. doi: 10.1093/gigascience/giz165

Amunts, K., Lepage, C., Borgeat, L., Mohlberg, H., Dickscheid, T., Rousseau, M.-É., et al. (2013). BigBrain: an ultrahigh-resolution 3D human brain model. *Science* 340, 1472–1475. doi: 10.1126/science.1235381

Assante, M., Candela, L., Castelli, D., and Tani, A. (2016). Are scientific data repositories coping with research data publishing? *Data Sci. J.* 15, 1–24. doi: 10.5334/dsj-2016-006

Attwood, T. K., Agit, B., and Ellis, L. B. M. (2015). Longevity of biological databases. *EMBnet.journal* 21:803. doi: 10.14806/ej.21.0.803

Bahim, C., Casorrán-Amilburu, C., Dekkers, M., Herczog, E., Loozen, N., Repanas, K., et al. (2020). The FAIR data maturity model: an approach to harmonise FAIR assessments. *Data Sci. J.* 19, 1–7. doi: 10.5334/dsj-2020-041

Bilder, Geoffrey, Lin, Jennifer, and Neylon, Cameron. (2015). "Principles for open scholarly infrastructures." Science in the Open. Available at: https://cameronneylon.net/blog/principles-for-open-scholarly-infrastructures/.

Bandrowski, A., Grethe, J. S., Pilko, A., Gillespie, T., Pine, G., Patel, B., et al. (2021). SPARC Data Structure: Rationale and Design of a FAIR Standard for Biomedical Research Data. *bioRxiv*. doi: 10.1101/2021.02.10.430563

Botvinik-Nezer, R., Holzmeister, F., Camerer, C. F., Dreber, A., Huber, J., Johannesson, M., et al. (2020). Variability in the analysis of a single neuroimaging dataset by many teams. *Nature* 582, 84–88. doi: 10.1038/s41586-020-2314-9

Bush, K. A., Calvert, M. L., and Kilts, C. D. (2022). Lessons learned: a neuroimaging research Center's transition to open and reproducible science. *Front. Big Data* 5:988084. doi: 10.3389/fdata.2022.988084

Button, K. S., Ioannidis, J. P. A., Mokrysz, C., Nosek, B. A., Flint, J., Robinson, E. S. J., et al. (2013). Power failure: why small sample size undermines the reliability of neuroscience. *Nature Rev. Neurosci. Advan.* 14, 365–376. doi: 10.1038/nrn3475

Cachat, J., Bandrowski, A., Grethe, J. S., Gupta, A., Astakhov, V., Imam, F., et al. (2012). A survey of the neuroscience resource landscape: perspectives from the neuroscience information framework. *Int. Rev. Neurobiol.* 103, 39–68. doi: 10.1016/B978-0-12-388408-4.00003-4

Dempsey, W., Foster, I., Fraser, S., and Kesselman, C. (2022). Sharing begins at home: how continuous and ubiquitous FAIRness can enhance research productivity and data reuse. *Harvard Data Sci. Rev.* 4. doi: 10.1162/99608f92.44d21b86

Eke, D. O., Bernard, A., Bjaalie, J. G., Chavarriaga, R., Hanakawa, T., Hannan, A. J., et al. (2022). International data governance for neuroscience. *Neuron* 110, 600–612. doi: 10.1016/j.neuron.2021.11.017

Ferguson, A. R., Nielson, J. L., Cragin, M. H., Bandrowski, A. E., and Martone, M. E. (2014). Big data from small data: data-sharing in the 'long tail' of neuroscience. *Nat. Neurosci.* 17, 1442–1447. doi: 10.1038/nn.3838

Fothergill, B. T., Knight, W., Stahl, B. C., and Ulnicane, I. (2019). Responsible data governance of neuroscience big data. *Front. Neuroinform.* 13:28. doi: 10.3389/fninf.2019.00028

Fouad, K., Vavrek, R., Surles-Zeigler, M. C., Huie, J. R., Radabaugh, H., Gurkoff, G. G., et al. (2023). A practical guide to data management and sharing for biomedical laboratory researchers. Zenodo. doi: 10.5281/zenodo.8206341,

"Funders' Policies." (2015). Available at: https://www.data.cam.ac.uk/funders.

Gardner, D., Akil, H., Ascoli, G. A., Bowden, D. M., Bug, W., Donohue, D. E., et al. (2008). The neuroscience information framework: a data and knowledge environment for neuroscience. *Neuroinformatics* 6, 149–160. doi: 10.1007/s12021-008-9024-z

Gillespie, T. H., Tripathy, S. J., Sy, M. F., Martone, M. E., and Hill, S. L. (2022). The neuron phenotype ontology: a FAIR approach to proposing and classifying neuronal types. *Neuroinformatics.* 20, 793–809. doi: 10.1007/s12021-022-09566-7

Gonçalves, R. S., and Musen, M. A. (2019). The variable quality of metadata about biological samples used in biomedical experiments. *Scientific Data* 6:190021. doi: 10.1038/sdata.2019.21

Gorgolewski, K. J., Auer, T., Calhoun, V. D., Cameron Craddock, R., Das, S., Duff, E. P., et al. (2016). The brain imaging data structure, a format for organizing and describing outputs of neuroimaging experiments. *Scientific Data* 3:160044. doi: 10.1038/sdata.2016.44

"Governance." (n.d.). Brain imaging data structure. Available at: https://bids.neuroimaging.io/governance.

Grethe, J. S., Baru, C., Gupta, A., James, M., Ludaescher, B., Martone, M. E., et al. (2005). Biomedical informatics research network: building a National Collaboratory to hasten the derivation of new understanding and treatment of disease. *Stud. Health Technol. Inform.* 112, 100–109.

Hamilton, D. J., Shepherd, G. M., Martone, M. E., and Ascoli, G. A. (2012). An ontological approach to describing neurons and their relationships. *Front. Neuroinform.* 6:15. doi: 10.3389/fninf.2012.00015

Hawrylycz, M., Baldock, R. A., Burger, A., Hashikawa, T., Johnson, G. A., Martone, M. E., et al. (2011). Digital Atlasing and standardization in the mouse brain. *PLoS Comput. Biol.* 7:e1001065. doi: 10.1371/journal.pcbi.1001065

Hawrylycz, M., Boline, J., Burger, A., Hashikawa, T., Johnson, G. A., Martone, M. E., et al. (2009). "The INCF digital Atlasing program: report on digital Atlasing standards in the rodent brain." Available at: http://precedings.nature.com/documents/4000/version/1.

Hawrylycz, M., Martone, M. E., Ascoli, G. A., Bjaalie, J. G., Dong, H.-W., Ghosh, S. S., et al. (2023). A guide to the BRAIN initiative cell census network data ecosystem. *PLoS Biol.* 21:e3002133. doi: 10.1371/journal.pbio.3002133

Hodson, Simon, Jones, Sarah, Collins, Sandra, Genova, Françoise, Harrower, Natalie, Laaksonen, Leif, et al. (2018). *Turning FAIR data into reality: Interim report from the European Commission expert group on FAIR data*. Amsterdam, Washington, DC: IOS Press.

International Brain Initiative (2020). International brain initiative: an innovative framework for coordinated global brain research efforts. *Neuron* 105, 212–216. doi: 10.1016/j.neuron.2020.01.002

Ioannidis, J. P. A. (2007). Why Most published research findings are false: Author's reply to Goodman and Greenland. *PLoS Med.* 4:2. doi: 10.1371/journal.pmed.0040215

Keator, D. B., Helmer, K., Steffener, J., Turner, J. A., Van Erp, T. G. M., Gadde, S., et al. (2013). Towards structured sharing of raw and derived neuroimaging data across existing resources. *NeuroImage* 82, 647–661. doi: 10.1016/j.neuroimage.2013.05.094

Kennedy, D. N., Abraham, S. A., Bates, J. F., Crowley, A., Ghosh, S., Gillespie, T., et al. (2019). Everything matters: the ReproNim perspective on reproducible neuroimaging. *Front. Neuroinform.* 13:1. doi: 10.3389/fninf.2019.00001

Kleven, H., Gillespie, T. H., Zehl, L., Dickscheid, T., Bjaalie, J. G., Martone, M. E., et al. (2023). AtOM, an ontology model to standardize use of brain atlases in tools, workflows, and data infrastructures. *Scientific Data* 10:486. doi: 10.1038/s41597-023-02389-4

"KnowledgeSpace." (n.d.). Available at: https://knowledge-space.org/.

Koslow, S. H. (2000). "Should the Neuroscience Community Make a Paradigm Shift to Sharing Primary Data?" *Nature Neuroscience* 3: 863–65.

Larson, S. D., and Martone, M. E. (2013). NeuroLex.org: an online framework for neuroscience knowledge. *Front. Neuroinform.* 7:18. doi: 10.3389/fninf.2013.00018

Lin, D., Crabtree, J., Dillo, I., Downs, R. R., Edmunds, R., Giaretta, D., et al. (2020). The TRUST principles for digital repositories. *Scientific Data* 7:144. doi: 10.1038/s41597-020-0486-7

Martone, Maryann E., and Nakamura, Richard. (2022). "Changing the culture on data management and sharing: overview and highlights from a workshop held by the National Academies of sciences, engineering, and medicine." Available at: https://hdsr.mitpress.mit.edu/pub/p1xu0son/release/1?readingCollection=b697ca32.

Martone, M. E., Gupta, A., and Ellisman, M. H. (2004). E-neuroscience: challenges and triumphs in integrating distributed data from molecules to brains. *Nat. Neurosci.* 7, 467–472. doi: 10.1038/nn1229

Miller, J. A., Gouwens, N. W., Tasic, B., Collman, F., van Velthoven, C. T. J., Bakken, T. E., et al. (2020). Common cell type nomenclature for the mammalian brain. *elife* 9. doi: 10.7554/eLife.59928

Mungall, C. J., Torniai, C., Gkoutos, G. V., Lewis, S. E., and Haendel, M. A. (2012). Uberon, an integrative multi-species anatomy ontology. *Genome Biol.* 13:R5. doi: 10.1186/gb-2012-13-1-r5

Murphy, F., Bar-Sinai, M., and Martone, M. E. (2021). A tool for assessing alignment of biomedical data repositories with open, FAIR, citation and trustworthy principles. *PLoS One* 16:e0253538. doi: 10.1371/journal.pone.0253538

National Academies of Sciences, Engineering, and Medicine (2018). *Open Science by design* National Academies of Sciences, Engineering, and Medicine. Washington DC. doi: 10.17226/25116

Nelson, Alondra. (2022). Office of Science and Technology Policy (OSTP). Letter to Heads of us Executive Departments and Agencies. "08-2022-OSTP-public-access-memo.Pdf," August 25, 2022.

Nelson, B. (2009). Data sharing: empty archives. *Nature* 461, 160–163. doi: 10.1038/461160a

"Neuroshapes." (n.d.). Accessed August 4, 2023. Available at: http://neuroshapes.org/.

"NIH workshop on the role of generalist repositories to enhance data discoverability and reuse: Workshop summary." (2012). NIH. Available at: https://datascience.nih.gov/data-ecosystem/nih-data-repository-workshop-summary.

Ozyurt, I. B., Grethe, J. S., Martone, M. E., and Bandrowski, A. E. (2016). Resource Disambiguator for the web: extracting biomedical resources and their citations from the scientific literature. *PLoS One* 11:e0146300. doi: 10.1371/journal.pone.0146300

Papp, E. A., Leergaard, T. B., Evan Calabrese, G., Johnson, A., and Bjaalie, J. G. (2014). Waxholm space atlas of the Sprague Dawley rat brain. *NeuroImage* 97, 374–386. doi: 10.1016/j.neuroimage.2014.04.001

Piekniewska, A., Haak, L. L., Henderson, D., McNeill, K., Bandrowski, A., and Seger, Y. (2023). Establishing an early Indicator for data sharing and reuse. Piekniewsk, doi: 10.31219/osf.io/ryxg2,

Poldrack, Russell A., Markiewicz, Christopher J., Appelhoff, Stefan, Ashar, Yoni K., Auer, Tibor, Baillet, Sylvain, et al. (2023). "The past, present, and future of the brain imaging data structure (BIDS)." arXiv [q-bio.OT]. arXiv. Available at: http://arxiv.org/abs/2309.05768.

Poline, J.-B., Das, S., Glatard, T., Madjar, C., Dickie, E. W., Lecours, X., et al. (2023). Data and tools integration in the Canadian open neuroscience platform. *Scientific Data* 10:189. doi: 10.1038/s41597-023-01946-1

Quaglio, G., Toia, P., Moser, E. I., Karapiperis, T., Amunts, K., Okabe, S., et al. (2021). The international brain initiative: enabling collaborative science. *Lancet Neurol.* 20, 985–986. doi: 10.1016/S1474-4422(21)00389-6

Ropelewski, A. J., Rizzo, M. A., Swedlow, J. R., Huisken, J., Osten, P., Khanjani, N., et al. (2022). Standard metadata for 3D microscopy. *Scientific Data* 9:449. doi: 10.1038/s41597-022-01562-5

Rübel, O, Andrew, T, Ryan, Ly, Benjamin, K., et al. (2022). "The Neurodata Without Borders Ecosystem for Neurophysiological Data Science." eLife 11 (October). doi: 10.7554/eLife.7836

Sandström, Malin, Abrams, Mathew, Bjaalie, Jan, Hicks, Mona, Kennedy, David, et al. (2022). "Recommendations for repositories and scientific gateways from a neuroscience perspective." arXiv [cs.CY]. arXiv. Available at: http://arxiv.org/abs/2201.00727.

Sansone, S.-A., McQuilton, P., Cousijn, H., Cannon, M., Chan, W. M., Callaghan, S., et al. (2020). Data repository selection: criteria that matter. doi: 10.5281/zenodo.4084763,

Shearer, Kathleen. (n.d.). "COAR community framework for best practices in repositories." Accessed April 3, 2021. Available at: https://www.coar-repositories.org/news-updates/coar-community-framework-for-best-practices-in-repositories/.

Shepherd, G. M., Marenco, L., Hines, M. L., Migliore, M., McDougal, R. A., Carnevale, N. T., et al. (2019). Neuron names: a gene-and property-based name format, with special reference to cortical neurons. *Front. Neuroanat.* 13:25. doi: 10.3389/fnana.2019.00025

Stall, S., Martone, M. E., Chandramouliswaran, I., Federer, L., Gautier, J., Gibson, J., et al. (2023). Generalist Repository Comparison Chart. doi: 10.5281/zenodo.7946938,

Subash, P., Gray, A., Boswell, M., Cohen, S. L., Garner, R., Salehi, S., et al. (2023). A comparison of Neuroelectrophysiology databases. *ArXiv* 10:719. doi: 10.1038/s41597-023-02614-0

Surles-Zeigler, M. C., Sincomb, T., Gillespie, T. H., de Bono, B., Bresnahan, J., Mawe, G. M., et al. (2021). Extending and Using Anatomical Vocabularies in the Stimulating Peripheral Activity to Relieve Conditions (SPARC) Project. *bioRxiv*. doi: 10.1101/2021.11.15.467961

Tan, S. Z., Kai, H. K., Aevermann, B. D., Gillespie, T., Harris, N., Hawrylycz, M. J., et al. (2023). Brain data standards - a method for building data-driven cell-type ontologies. *Scientific Data* 10:50. doi: 10.1038/s41597-022-01886-2

Torres-Espín, A., Haefeli, J., Ehsanian, R., Torres, D., Almeida, C. A., Russell Huie, J., et al. (2021). Topological network analysis of patient similarity for precision Management of Acute Blood Pressure in spinal cord injury. *elife* 10. doi: 10.7554/eLife.68015

Weiner, M. W., Aisen, P. S., Jack Jr, C. R., Jagust, W. J., Trojanowski, J. Q., Shaw, L., et al. (2010). The Alzheimer's Disease Neuroimaging Initiative: Progress report and future plans. *Alzheimers Dement.* 6, 202–11.e7. doi: 10.1016/j.jalz.2010.03.007

"*Whose Scans Are They, Anyway?*" (2000). Nature publishing group: UK. *406*, 2000.

Wilkinson, M. D., Michel Dumontier, I., Aalbersberg, J. J., Appleton, G., Axton, M., Baak, A., et al. (2016). The FAIR guiding principles for scientific data management and stewardship. *Scientific Data* 3:160018. doi: 10.1038/sdata.2016.18

Glossary

3D-MMS	Metadata for 3D microscopy standard
ADNI	Alzheimer's Disease Neuroimaging Initiative
BICAN	BRAIN Initiative Cell Atlas Network
BICCN	BRAIN Initiative Cell Census Network
BIDS	Brain Imaging Data Structure
BIL	Brain Image Library
BRAIN Initiative	Brain Research through Advancing Innovative Neurotechnologies
CDE	Common data element
CONP	Canadian Open Neuroscience Platform
CT	Computed tomography
DANDI	Distributed Archives for Neurophysiology Data Integration
DATs	Data tag suite
DBS	Deep brain stimulation
DOI	Digital Object Identifier
ECOG	Electrocorticography
EEG	Electron encephalography
EMG	Electromyography
ERP	Event-related potential
fMRI	Functional magnetic resonance imaging
FORCE11	Future of Research Communications and e-Scholarship
HBP	Human Brain Project
HED	Hierarchical event descriptor
IBI	International Brain Initiative
iEEG	Intracranial electroencephalography
INCF	International Neuroinformatics Coordinating Facility
MEG	Magnetoencephalography
MIS	SPARC minimal information standard
MRI	Magnetic resonance imaging
NEMAR	NeuroElectroMagnetic data Archive
NIF	Neuroscience Information Framework
NIH	National Institutes of Health
NWB	Neurodata Without Borders
ODC-SCI	Open Data Commons for Spinal Cord Injury
ODC-TBI	Open Data Commons for Traumatic Brain Injury
PET	Positron emission tomography
SDS	SPARC dataset structure
SPARC	Stimulating Peripheral Activity to Relieve Conditions
SPECT	Single-photon emission computed tomography
URI	Uniform Resource Identifier

frontiers | Frontiers in Neuroinformatics

TYPE Original Research
PUBLISHED 09 February 2024
DOI 10.3389/fninf.2024.1284107

OPEN ACCESS

EDITED BY
Christian Haselgrove,
UMass Chan Medical School, United States

REVIEWED BY
Adam Tyson,
University College London, United Kingdom
Yongsoo Kim,
Penn State Milton S. Hershey Medical Center, United States

*CORRESPONDENCE
Trygve B. Leergaard
✉ t.b.leergaard@medisin.uio.no

RECEIVED 27 August 2023
ACCEPTED 24 January 2024
PUBLISHED 09 February 2024

CITATION
Blixhavn CH, Reiten I, Kleven H, Øvsthus M, Yates SC, Schlegel U, Puchades MA, Schmid O, Bjaalie JG, Bjerke IE and Leergaard TB (2024) The Locare workflow: representing neuroscience data locations as geometric objects in 3D brain atlases.
Front. Neuroinform. 18:1284107.
doi: 10.3389/fninf.2024.1284107

The Locare workflow: representing neuroscience data locations as geometric objects in 3D brain atlases

Camilla H. Blixhavn[1], Ingrid Reiten[1], Heidi Kleven[1], Martin Øvsthus[1], Sharon C. Yates[1], Ulrike Schlegel[1], Maja A. Puchades[1], Oliver Schmid[2], Jan G. Bjaalie[1], Ingvild E. Bjerke[1] and Trygve B. Leergaard[1]*

[1]Neural Systems Laboratory, Department of Molecular Medicine, Institute of Basic Medical Sciences, University of Oslo, Oslo, Norway, [2]EBRAINS AISBL, Brussels, Belgium

Neuroscientists employ a range of methods and generate increasing amounts of data describing brain structure and function. The anatomical locations from which observations or measurements originate represent a common context for data interpretation, and a starting point for identifying data of interest. However, the multimodality and abundance of brain data pose a challenge for efforts to organize, integrate, and analyze data based on anatomical locations. While structured metadata allow faceted data queries, different types of data are not easily represented in a standardized and machine-readable way that allow comparison, analysis, and queries related to anatomical relevance. To this end, three-dimensional (3D) digital brain atlases provide frameworks in which disparate multimodal and multilevel neuroscience data can be spatially represented. We propose to represent the locations of different neuroscience data as geometric objects in 3D brain atlases. Such geometric objects can be specified in a standardized file format and stored as location metadata for use with different computational tools. We here present the Locare workflow developed for defining the anatomical location of data elements from rodent brains as geometric objects. We demonstrate how the workflow can be used to define geometric objects representing multimodal and multilevel experimental neuroscience in rat or mouse brain atlases. We further propose a collection of JSON schemas (LocareJSON) for specifying geometric objects by atlas coordinates, suitable as a starting point for co-visualization of different data in an anatomical context and for enabling spatial data queries.

KEYWORDS

3D brain atlas, FAIR data, interoperability, rat brain, mouse brain, standardization, data integration

1 Introduction

Experimental brain research in animal models generates large amounts of disparate data of different modality, format, and spatial scale (Sejnowski et al., 2014). To manage and exploit the growing resource of neuroscience data it is now widely recognized that the data must be shared in accordance with the FAIR principles (Wilkinson et al., 2016), ensuring that data are findable, accessible, interoperable and reusable for future analyses (see e.g., Abrams et al.,

2022). This trend has resulted in a growing volume of neuroscience data being made accessible through various data repositories and infrastructures (Ferguson et al., 2014; Jorgenson et al., 2015; Ascoli et al., 2017; Amunts et al., 2019). While free-text searches based on structured metadata are typically implemented in such databases (Clarkson, 2016), possibilities for more sophisticated queries, visualizations, and analysis depend on a harmonization across data files with different formats, scales, and organization (Zaslavsky et al., 2014; Abrams et al., 2022).

Anatomical information is widely used to provide a common context for harmonizing and comparing neuroscience data (Martone et al., 2004; Bassett and Sporns, 2017). The availability of open-access 3D rodent brain reference atlases (Oh et al., 2014; Papp et al., 2014; Wang et al., 2020; Kleven et al., 2023a) has opened up new opportunities for combining and analyzing data that have been aligned to a common spatial framework (Leergaard and Bjaalie, 2022). This allows researchers to integrate and analyze data from different sources within a common anatomical context more easily. For example, spatial registration procedures allow image data to be directly compared and analyzed based on atlas coordinates or annotated brain structures (Puchades et al., 2019; Tappan et al., 2019; Tyson and Margrie, 2022; Kleven et al., 2023b), e.g., through use of computational analyses of features of interest in atlas-defined regions of interest (Kim et al., 2017; Bjerke et al., 2018b, 2023; Yates et al., 2019; Kleven et al., 2023a,b). For other data types, such as locations of electrode tracts, 3D reconstructed neurons, or other features of interest, procedures and tools have been developed to represent the data as coordinate-based points of interest allowing validation or visualization of locations (Bjerke et al., 2018b; Fiorilli et al., 2023).

Atlases, tools, and resources for building, viewing, and using collections of spatially registered data have also proven to be fundamental for digital research infrastructures, such as the Allen Brain Map data portal[1] and to some extent also the EBRAINS Research Infrastructure.[2] But while the Allen institute provides collections of systematically generated homogenous and standardized image data spatially integrated in a 3D atlas, EBRAINS allows the research community to share a wide variety of data. These data may be related to anatomical locations using anatomical terms, reference to stereotaxic coordinates, or spatial registration to atlases. Thus, the location documentation provided with published data is as disparate as the data themselves—ranging from coordinate-based documentation defining the position of data in an atlas, to anatomical terms, illustrations, and unstructured descriptions (Bjerke et al., 2018a). The specification of such location metadata varies considerably, and a common standard for storing them is lacking in neuroscience. This poses a challenge to effectively utilize the metadata for spatial queries, co-visualization, and other analytic purposes. To achieve the ambitions of the community to accumulate and re-use neuroscience research data in agreement with the FAIR principles, it is necessary to represent metadata describing anatomical locations (spatial metadata) in a standardized and machine-readable format.

To address this challenge, we developed the Locare workflow (from *locāre*, latin: to place) for representing disparate neuroscience data in a simplified and standardized manner. The workflow was developed based on a large collection of diverse experimental data from mouse and rat brains shared via the EBRAINS Knowledge Graph.[3] The available location documentation, specifying data location through points of interest, images, or semantic descriptions determines the starting point of the workflow, which through different workflow routes outputs geometric objects. We here present Locare as a generic workflow for specifying interoperable spatial metadata for neuroscience data, and exemplify how it can be used to specify anatomical locations for different data types as geometric objects in atlas space using a JSON format. The LocareJSON schemas allow representation of data in a simplified and standardized format that can enable spatial search, co-visualization, and analyses of otherwise disparate neuroscience data. The Locare workflow provides a solution for defining heterogeneous neuroscience data as atlas-defined geometric objects in a machine-readable format, which in turn can be utilized to represent data as interoperable objects in a 3D anatomical atlas and develop spatial query functionalities. The workflow is here presented in context of the EBRAINS Research Infrastructure but is generally applicable to any infrastructure of databases holding neuroscience data.

2 Materials and methods

The Locare workflow builds on several years of experience with assisting researchers to share and present their experimental research data through the EBRAINS Research Infrastructure. As part of this effort, we investigated how to integrate and represent rat and mouse data sets in three-dimensional (3D) brain atlases. The workflow was established using 186 mouse brain data sets and 94 rat brain data sets available from the EBRAINS Knowledge Graph by 11 May 2023. An overview of all data set titles and type of location documentation is provided in Supplementary Table 1. The data sets included data files in various formats, structured metadata, and a data descriptor including summary, materials and methods, usage notes and explanation of data records. Several data sets were also associated with journal publications containing additional images and/or textual information about the anatomical location of the data. In some cases, we were in contact with data providers (custodians of the data shared through EBRAINS) directly and received additional information. These 280 data sets were contributed by 480 different researchers and acquired using 25 different experimental methods. The anatomical locations of observations or measurements in these data sets were documented using images ($n = 116$), semantic descriptions only ($n = 123$), or by specification of coordinates for points of interest (POIs; $n = 41$).

1 https://portal.brain-map.org/

2 https://www.ebrains.eu/

3 https://search.kg.ebrains.eu/

2.1 Establishing the Locare workflow

The Locare workflow takes any information that can be used to define the anatomical location of a sample (e.g., a section or a tissue block) or objects within a sample (e.g., a labeled cell or an electrode) of data as input, independent of methods, data formats, software used for visualization and analysis, and solutions used for sharing the data. This is below referred to as location documentation. Three main categories of location documentation input are distinguished: images, information about POIs, and semantic descriptions. The workflow includes four steps: (1) choosing a target atlas (a 3D brain atlas) and collecting relevant location documentation (Figure 1A); (2) assessing the location documentation (Figure 1B); (3) translating location documentation to spatial metadata in target atlas (Figure 1C); and (4) defining the geometric object representing the location of the data (Figure 1D). A geometric object is a simplified representation of the anatomical location from which the data were derived. If the exact location that the data were derived from cannot be defined, the location can be represented by a geometric object (a mesh) corresponding to an atlas region. The target atlas constitutes the common framework for spatial alignment of data from different sources, enabling meaningful comparisons and integrations.

To exemplify how the output of the workflow can be formatted in a standardized, machine-readable way, we created a collection of JavaScript Object Notation (JSON)[4] schemas to store the Locare workflow output. The JSON format is widely used due to its suitability for storing semi-structured information, language independence and human readability. Since there are several open standards related to neuroscientific data and geometric representations (such as GeoJSON, NeuroJSON, and openMINDS), we assessed these for inspiration. GeoJSON[5,6] is a format for encoding a variety of geographical data structures but is lacking fields to specify the anatomical context for neuroscience data. NeuroJSON[7] is a JSON-based neuroimaging exchange format. The NeuroJSON JMesh specification can efficiently represent 3D graphical objects, such as shape primitives (spheres, boxes, cylinders, etc.), triangular surfaces or tetrahedral meshes. However, like GeoJSON, the Jmesh specification misses the option to identify the anatomical context. openMINDS (RRID:SCR_ 023173)[8] is a metadata framework with metadata models composed of schemas that structure information on data within a graph database. Although the schemas of the openMINDS SANDS (RRID:SCR_023498)[9] metadata model allow for the identification of the anatomical context (semantic and coordinate-based location and relation of data), it is not meant to hold actual (more complex) geometrical data. We chose to base our collection of schemas (LocareJSON) on the GeoJSON standard but extended it to include 3D objects and anatomical context. We defined LocareJSON schemas to the following geometrical objects: point, sphere, line string, cylinder, polygon, polyhedron, and atlas mesh. All LocareJSON schemas define target atlas space through a reference to relevant openMINDS schemas. The Locare atlas mesh schema also defines the relevant atlas mesh through use of openMINDS. For a detailed description of the LocareJSON schemas, see the LocareJSON Github repository (v1.1.1).[10]

2.2 Demonstrating the workflow through use-cases

We demonstrate the Locare workflow in a selection of use-cases including heterogeneous data from rat and mouse brains representing each input (location documentation) and output type (geometric objects; Figure 2; Supplementary Table 2). The output resulting from these use-cases were shared in the LocareJSON repository, and as data sets on EBRAINS (Blixhavn et al., 2023a,b,c,d,e,f; Reiten et al., 2023a,b,c). Below, we describe the key tools and processes used to create the use-cases.

We used the Waxholm Space atlas of the Sprague Dawley rat brain (WHS rat brain atlas; RRID: SCR_017124; Papp et al., 2014; Kjonigsen et al., 2015; Osen et al., 2019; Kleven et al., 2023a)[11] and the Allen mouse brain atlas Common Coordinate Framework (AMBA CCF) version 3 (RRID: SCR_020999; Wang et al., 2020) as our target atlases. For spatial registration, we used the QuickNII (RRID: SCR_016854; Puchades et al., 2019)[12] and VisuAlign (RRID: SCR_017978)[13] tools, which come in versions bundled with each of the target atlases.

For extraction of coordinates for a single or a few points of interest, we used the QuickNII mouse-hover function. For more extensive efforts involving numerous points of interest, we used the manual annotation function in the LocaliZoom tool (RRID:SCR_023481),[14] or the QUINT workflow (Yates et al., 2019; Gurdon et al., 2023)[15] utilizing QuickNII for registering histological brain section images to the reference atlas followed by tools for extracting (ilastik, RRID:SCR_015246), quantifying, and sorting features according to atlas regions (Groeneboom et al., 2020; RRID: SCR_017183).[16]

To facilitate translation across different atlas terminologies and coordinate systems, we used a set of published data sets containing metadata defining the spatial registration of the rat brain atlas plates of Paxinos and Watson (2018) to the WHS rat brain atlas and the mouse brain atlas plates of Franklin and Paxinos (2007) to the AMBA CCF v3 (Bjerke et al., 2019a,b). These data sets were used to relate stereotaxic landmarks to 3D atlas coordinates, as well as for comparing atlas regions between atlases, as shown in Bjerke et al. (2020a). Since the atlases by Franklin and Paxinos (2007) and Paxinos and Watson (2018) are copyrighted, the data sets do not contain images from these atlases. However, the registration metadata for these data sets can

4 https://www.json.org/json-en.html
5 https://geojson.org/
6 https://doi.org/10.17487/RFC7946
7 https://neurojson.org/
8 https://github.com/HumanBrainProject/openMINDS
9 https://github.com/HumanBrainProject/openMINDS_SANDS
10 https://github.com/Neural-Systems-at-UIO/LocareJSON/tree/v1.1.1
11 http://www.nitrc.org/projects/whs-sd-atlas/
12 https://quicknii.readthedocs.io
13 https://visualign.readthedocs.io
14 https://localizoom.readthedocs.io/en/latest/
15 https://quint-workflow.readthedocs.io
16 https://nutil.readthedocs.io

10.3389/fninf.2024.1284107

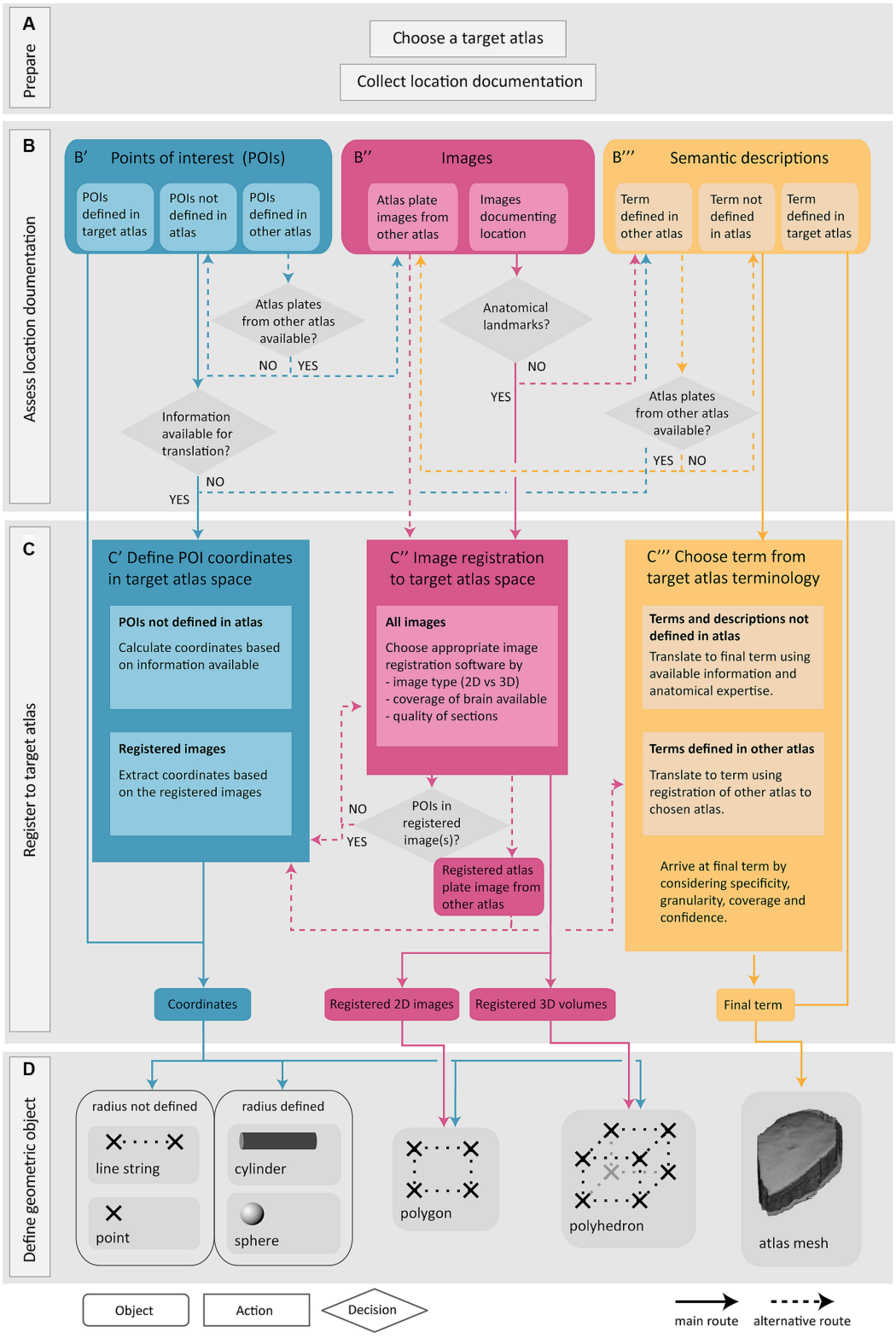


FIGURE 1
Overview of the Locare workflow. Location documentation is collected **(A)**, assessed **(B)**, and registered to a target atlas **(C)** followed by the creation of geometric objects representing the data of which the location documentation was derived **(D)**. **(A)** Preparatory steps involve choosing a target atlas in which the geometric objects should be represented and collecting of relevant location documentation. **(B)** The location documentation available, defined as points of interest (POI; **B'**), images **(B'')** or semantic descriptions **(B''')**, determines which route of the workflow is used. **(B',C')** Point route: POIs may be defined in the target atlas, in another atlas, or not in an atlas. POIs defined in target atlas are directly used to create geometric objects. POIs not defined in the target atlas must be translated to coordinates of the target atlas **(C')** (see text for details). If no information is available for translation of POIs to target atlas, the inputs are directed to semantic translation (**C'''**, blue arrow). **(B'',C'')** Image route: Images may document the

(Continued)

FIGURE 1 (Continued)
location of specific data or can also be atlas plate images used to translate points of interest or semantic descriptions to a geometric object or mesh in the target atlas. Image registration is performed if possible **(C″)**, or alternatively the workflow can be directed to the semantic route (**B‴**, pink arrow). Images registered to the target atlas containing POIs may be used for coordinate extraction (**C′**, pink arrow). Atlas plate images from other atlas registered to the target atlas is used for extraction of coordinates for POIs (**C′**, pink arrow) or for translation of semantic term (**C‴**, pink arrow). **(B″,C″)** Semantic route: Semantic descriptions may be defined in the target atlas, another atlas, or not defined in an atlas. Terms defined in target atlas are directly used as the final term. Terms defined in other atlas are translated based on the spatial registration of atlas plates from the other atlas to the target atlas (**B″**, yellow arrow). Terms not defined in any atlas are translated to the most closely corresponding term in the target atlas **(C‴)**. **(D)** The output of the workflow routes is one or several geometric objects or atlas meshes.

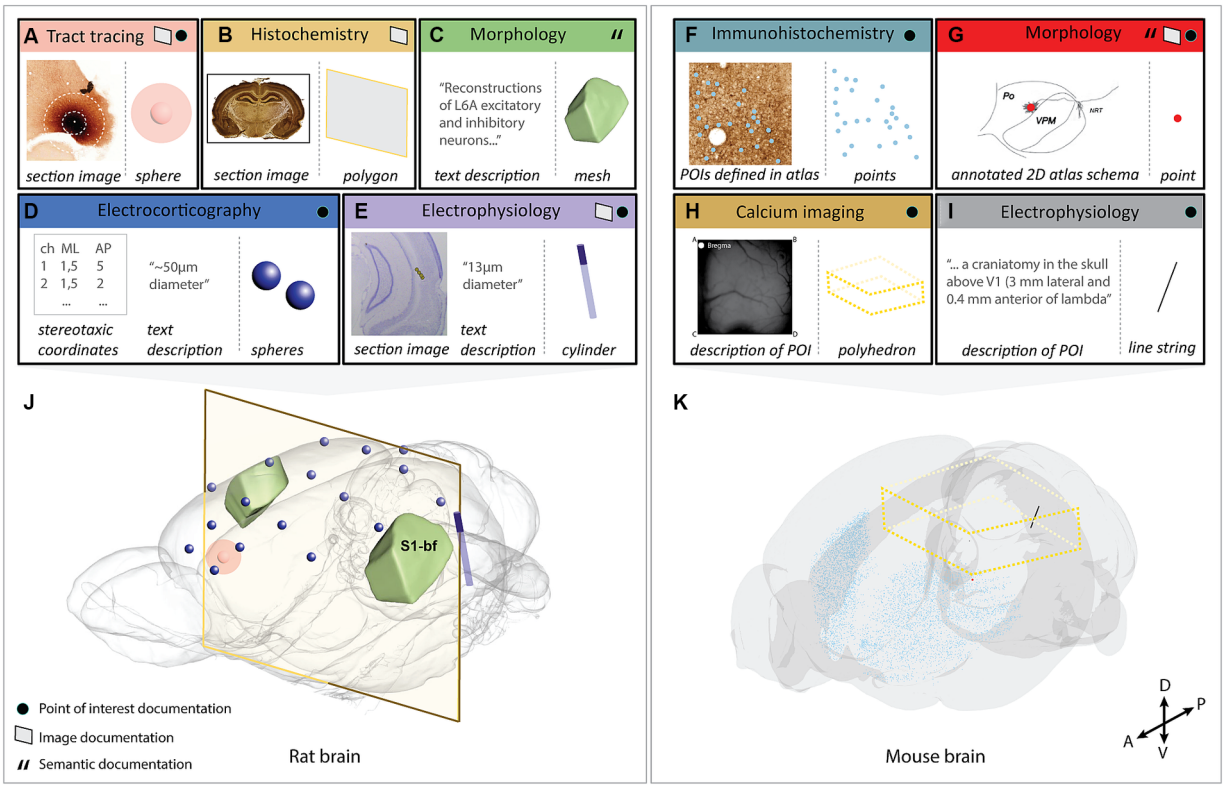


FIGURE 2
Visualization of the selected use-cases demonstrating the use of the Locare workflow. Use-cases **(A–I)** represented by an input (location documentation) and output (geometric object representation), where the outputs are co-visualized in the respective target atlases **(J)** [Waxholm Space atlas of the Sprague Dawley rat brain or **(K)**; Allen mouse brain atlas Common Coordinate Framework version 3]. **(A)** Image from an anterograde tract tracing experiment showing the injection site placed in the medial orbital area (Kondo et al., 2022). Two spheres represent the position and size of the injection site core and shell, respectively. **(B)** Image from a histochemistry experiment (Blixhavn et al., 2022). A polygon represents the location of the section image. **(C)** Text description from a neuronal reconstruction study (Feldmeyer et al., 2020). An atlas mesh represents the location of the reconstructions. **(D)** Stereotaxic coordinates and radius measurement from electrocorticography experiments (Arena and Storm, 2018; Arena et al., 2019a,b, 2020) using 17 epidural electrodes. A sphere represents the position and extent of each electrode. **(E)** Image from an electrophysiology experiment (Fiorilli et al., 2022) where the electrode track is annotated. A cylinder represents the location of the electrode. **(F)** Image from an immunohistochemistry experiment (Bjerke et al., 2020b) with extracted parvalbumin positive cells annotated. Points represent the extracted cells. **(G)** 2D atlas illustration showing the location of a neuronal reconstruction (García-Amado et al., 2020). A point represents the neuronal soma. **(H)** Descriptions of the field of view used in a calcium imaging experiment (Conti et al., 2019; Resta et al., 2021). A polyhedron represents the field of view. **(I)** A text description of the POI used in an electrophysiology experiment (Schnabel et al., 2020). A line string represents the location of the electrode. **(J)** All use-cases containing data from the rat brain co-visualized in the Waxholm Space atlas of the Sprague Dawley rat brain version 4 (RRID: SCR_017124; Papp et al., 2014; Kleven et al., 2023a; http://www.nitrc.org/projects/whs-sd-atlas/). The coordinates of the objects are opened using MeshView (RRID: SCR_017222) the atlas mesh is opened using Scalable Brain Atlas Composer (Bakker et al., 2015), and objects are overlaid. **(K)** All use-cases containing data from the mouse brain co-visualized in the Allen mouse brain atlas Common Coordinate Framework version 3 (RRID: SCR_020999; Wang et al., 2020). The coordinates of the objects are opened using MeshView and objects are overlaid. S1-bf; primary somatosensory cortex, barrel field.

be opened and inspected with locally stored .png images using QuickNII, either to inspect the correspondence of delineations across atlases or to extract WHS rat brain atlas or AMBA CCF v3 coordinates.

To translate spatial metadata from established tools to our example schemas, we created Python scripts for extraction and formatting of (1) QuickNII .json files and (2) Nutil .json coordinate files. The output from QuickNII consists of vectors indicating the position of the 2D image in a 3D atlas (the vector components o, u, v represent the top left corner, and the horizontal and vertical edges of the image, respectively). Coordinates for all four corners can therefore be calculated by

addition of vectors. We created scripts[17] to transform the coordinate output from QuickNII .json files into the LocareJSON schema for polygons. In the Nutil tool, utilized in the QUINT workflow, users can choose whether output coordinates should be given per pixel of an image segmentation, or per centroid of each segmented object. We created scripts[18] to transform centroid coordinate output from the Nutil tool into the LocareJSON schema for points.

3 Results

We here present the Locare workflow and a collection of JSON schemas (LocareJSON) for representing the location of data as geometric objects in 3D atlases. First, we outline the generic steps of the workflow, followed by a description of three different routes for use of the workflow based on the type of location documentation available. Second, we describe the LocareJSON schemas for storing the geometric objects. Lastly, we demonstrate the workflow through nine use-cases representing five different experimental approaches and all the geometrical object types defined by the LocareJSON schemas. Figure 2 gives an overview of the input (location documentation) and output (geometric object representation) for each use-case and visualizes their outputs in their respective 3D target atlases. A summary of details for each use-case is found in Supplementary Table 2.

3.1 The Locare workflow

The Locare workflow consists of four steps (Figure 1). The first step (Figure 1A) is to select a target atlas and collect available location documentation, serving as the workflow input. The second step is to assess the available documentation (Figure 1B). The Locare workflow separates location documentation into three main categories: images showing anatomical features, specification of points of interest (POIs), and semantic descriptions. The third step of the workflow (Figure 1C) involves a registration and/or translation process to define coordinates or terms in the target atlas representing the anatomical location of the data set of interest. The fourth and last step (Figure 1D) is to define a geometric object using the appropriate LocareJSON schema. The image and point routes through the workflow yield representations of data location in form of geometric objects, such as points, spheres, line strings, cylinders, polygons, or polyhedrons. The semantic route results in atlas mesh polyhedrons representing an atlas term, which can be used to indicate that data resided somewhere within, or intersecting a given region. The link between the geometric object(s) defined in the Locare workflow and the data set containing the data described in the location documentation is defined in the LocareJSON schema (see section 3.2). Below, we describe the different routes of the workflow in more detail.

3.1.1 The workflow route for points of interest

POIs in a data set can be specified with a broad range of location documentation but are often specified as 2D or 3D points in a coordinate space or image. The POI route through the workflow translates POIs to coordinates in the target atlas and allows users to define geometric objects based on combinations of atlas coordinates. Of the 280 data sets evaluated (Supplementary Table 1), 41 provided documentation of their study target location as POIs.

The Locare workflow distinguishes between three different types of POI documentation (Figure 1B′). First, points may be given as coordinates defined in the target atlas, e.g., coordinates representing features extracted from images, as given for parvalbumin neurons in the data provided by Bjerke et al. (2020b). These coordinates can be used directly to create geometric objects in the target atlas (Figure 1D). Second, points may be specified as coordinates defined in other atlases than the target atlas, for example using coordinates from stereotaxic book atlases (e.g., for the position of implanted electrodes, as provided in use-cases shown in Figures 2D,I). If images from the atlas used to define the POIs are available (Figure 1B′, blue arrow), these can be spatially registered as described in the image route (Figure 1C″, see also section 3.1.2) to enable the translation of the POIs to coordinates in the target atlas. Thirdly, POIs may also represent information about the location of recording sites, images, or other spatial information that can be translated to the target atlas via anatomical landmarks (Figures 2G–I).

When coordinates are defined in the target atlas, they can be used to create all types of geometric objects supported by the LocareJSON schemas. For example, points can be used to represent cell soma positions (Figures 2F,G), a line string could represent the location of an electrode track (Figure 2I), or a polygon could represent the location of a camera field-of-view (the latter may also be extended to a polyhedron to represent the imaging depth captured by the camera; Figure 2H). If the radius for the POI is known, the point object could be replaced by a sphere, or a line string by a cylinder. For example, the location of an electrode track may be represented by a cylinder (Figure 2E), and the location of an injection site core and shell can be represented by a set of spheres with the same centroid point (Figure 2A).

3.1.2 The workflow route for image location documentation

Location documentation in the form of images varies greatly. Images may be magnified microscopy images focusing on specific structures or cover entire brain sections. Image series may contain only a few sections or cover the whole brain (see use-cases shown in Figures 2A,B,F). Image documentation may also be illustrations based on microscopy images, visualizations of reconstructions, or annotations made on atlas plates, as exemplified in Figure 2G. The main process of the image route is to register the images to the target atlas so that coordinate information can be extracted and used to create geometric objects. Of the 280 data sets evaluated in the work with defining this workflow (Supplementary Table 1), 116 provided documentation of their study target location through images.

Images are suitable for spatial registration if they contain specific anatomical features that allow identification of positions in the brain. Thus, in the second step of the workflow route for images (Figure 1C″), the images are examined to see if they meet this criterium. 2D images to be registered should ideally cover whole brain sections, or at least

17 https://github.com/Neural-Systems-at-UIO/LocareJSON/tree/v1.1.1/scripts/quicknii_to_locarePolygons

18 https://github.com/Neural-Systems-at-UIO/LocareJSON/tree/v1.1.1/scripts/centroids_to_locarePoints

include unique landmarks (Bjerke et al., 2018a) that can be used to determine the angle of sectioning. 3D volumes may cover the whole brain or be partial volumes. Partial 3D volumes to be registered should preferably contain a combination of external and internal anatomical landmarks to allow identification of corresponding locations in an atlas. A range of image registration software are available (Klein et al., 2010; Niedworok et al., 2016; Fürth et al., 2018; Puchades et al., 2019; Tappan et al., 2019), suitable for different types of data and purposes. Further discussions about the choice and application of such tools are provided in reviews by Tyson and Margrie (2022) and Kleven et al. (2023b). Whether or not suitable anatomical landmarks are available for determining the specific anatomical location of a sample should be considered case by case. If the images lack anatomical landmarks, the available information is considered using the semantic route of the workflow.

When registration is performed, the spatial registration output can be used to define geometric objects in the appropriate LocareJSON schema. For 2D images, polygons are used, representing the full plane of the image through defining its four corners (Figure 1D, see also Figures 2A,B). For 3D images, polyhedrons are used, representing the volume through defining the object's eight corners. For images containing POIs (e.g., annotations of electrode tracks, see Figure 2E), the image route would be used primarily as a mean to define coordinates corresponding to these points. In these cases, it might not be relevant to define geometric objects for the images themselves; instead, the extracted points are taken through the last two steps of the points route (Figures 1C′,D).

3.1.3 The workflow route for semantic location documentation

Semantic location documentation can be any term or description of an anatomical location. This includes a range of documentation types that do not meet the criteria for use in the other routes but still are useful to determine the data location. For example, images that do not contain sufficient anatomical landmarks for spatial registration may be useful for morphological observations of cells of tissue that can be used to determine the anatomical location of data. Semantic location documentation may also include functional characteristics of cells or tissue recorded which could help confirm the location of electrode tracks. The most common form of semantic location documentation, however, is one or more anatomical terms, with or without reference to a brain atlas. Of the 280 data sets evaluated in the work with defining this workflow (Supplementary Table 1), 123 provided documentation of their study target location through semantic descriptions only.

With the semantic route, a brain region term in the target atlas is chosen to represent the location of the data. In the second step of the semantic route (Figure 1B‴), we distinguish between terms defined in the target atlas, terms defined in another atlas, and terms not defined in an atlas. In the third step (Figure 1C‴), data are semantically registered to the target atlas by choosing a final term from the target atlas terminology to represent the data. The approach depends on which type of term was provided. For terms that are already associated to the target atlas, we generally use the term directly as the final term. For terms from other atlases, the registration to the target atlas involves a translation between terminologies, a process depending on defining the correspondence of the region in the other atlas with region(s) in the target atlas. If images of atlas plates from the other atlas are available (Figure 1B‴, yellow arrows), they can be spatially registered as described in the image route (Figure 1C″) and atlas plates can be overlayed with custom atlas overlays from the target atlas. This facilitates translation of terms from the other atlas to the target as described in our previous papers (Bjerke et al., 2020a; Kleven et al., 2023b). If alternative spelling or terms differing from the atlas nomenclature are used, further consideration about underlying definitions and correspondence to the atlas nomenclature is needed. For example, the term "striatum" can be ambiguous, since it may refer to the caudate-putamen (or caudoputamen) alone or the caudate-putamen combined with the nucleus accumbens. Use of parent terms, such as the "substantia nigra" to describe smaller subsets of a region can also introduce ambiguity. In all such cases it is necessary to evaluate available documentation and seek the most precise definition possible.

There are several considerations underpinning the choice of a final term when the initial term comes from another atlas or is not defined in an atlas. This process relies primarily on interpretation of the initial term and documentation by a researcher employing knowledge of neuroanatomy and neuroanatomical atlases, nomenclatures, and conventions. The documentation is evaluated in the choice of final terms, with essential considerations being the specificity, granularity, coverage, and confidence (defined in Figure 3). For example, if a term from another atlas is used, but there is no closely corresponding term in the target atlas, a fine-grained term might be substituted with a coarser term. This would decrease the granularity, but increase the confidence, in the final term. The final term will be chosen from the target atlas terminology, with a corresponding atlas mesh associated to the data set (Figure 1D).

3.2 The Locare workflow output: LocareJSON

To exemplify how the geometric object representing the anatomical location of a data element can be formalized in a machine-readable format, we created a collection of JavaScript Object Notation (JSON) schemas, collectively referred to as LocareJSON schemas. These schemas are based on GeoJSON elements and are hosted in the LocareJSON GitHub repository. These LocareJSON schemas provide suitable starting points for researchers who wish to create JSON files storing information about spatial location in the brain. Below we describe the structure and content of the LocareJSON schemas. Each schema consists of a general part (the locareCollection schema) and a part specific to the object it describes (individual object schemas).

The locareCollection schema include the following required properties: versioning of the schema (version), reference to the 3D target atlas (targetAtlas) and one or several persistent links to the original sources for the data (sourcePublication). The targetAtlas is referenced through a link to an openMINDS_SANDS (see text footnote 10) instance (commonCoordinateSpaceVersion). Details about the dimension, resolution, orientation, and origin of target atlas is essential to enable representation of geometric objects in any atlas space, e.g., in an online tool or viewer. The locareCollection schema has two optional properties: related publications (relatedPublications), and online resources (linkedURI, Uniform Resource Identifier). The linkedURI should be used to state an online resource primarily if it links to relevant data already embedded in a tool or viewer (e.g., as for

10.3389/fninf.2024.1284107

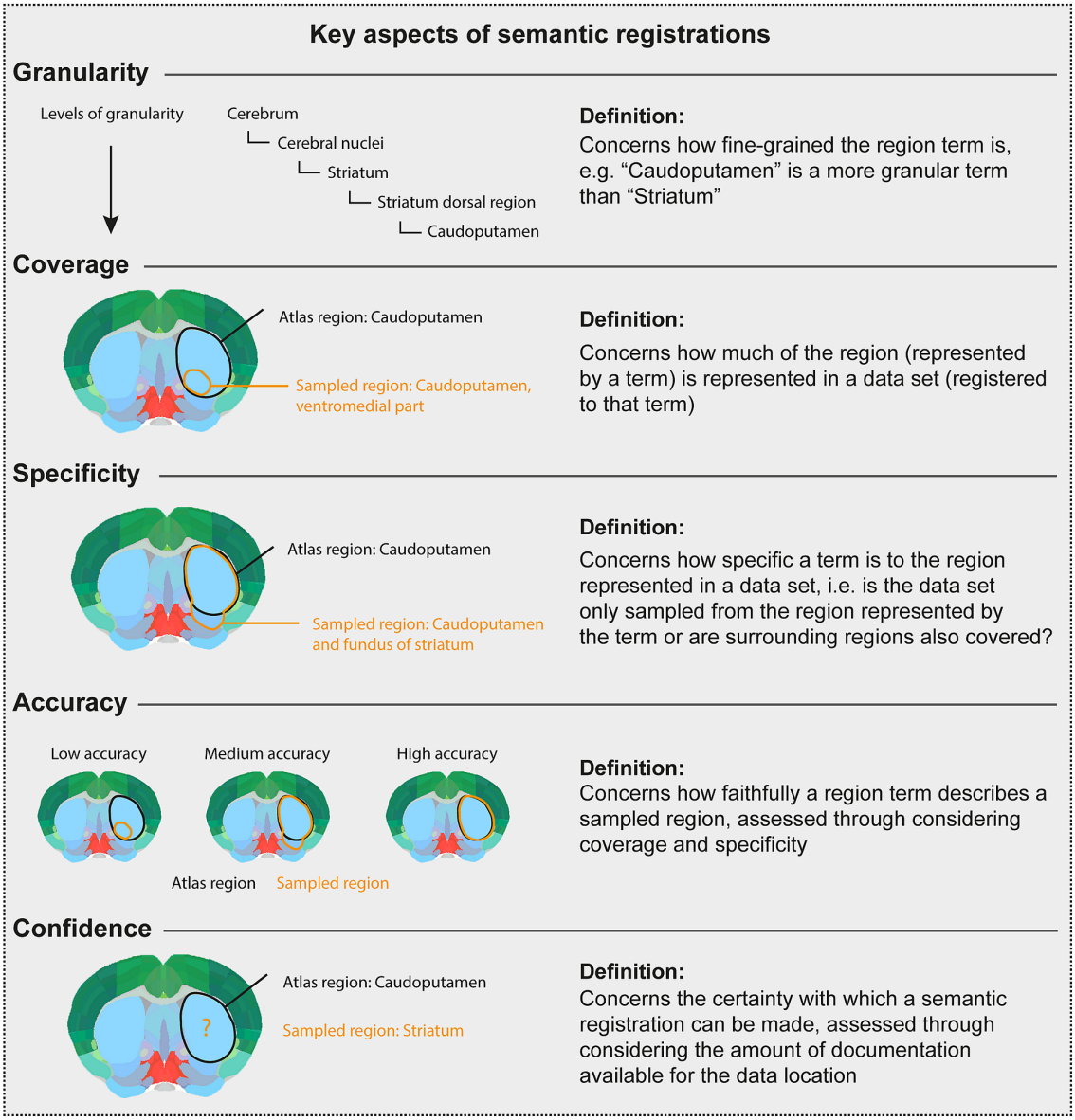


FIGURE 3
Key aspects of semantic registrations. The Figure [modified from Bjerke (2021)] gives the definition of key considerations when using the semantic route to represent data described using terms from other atlases than the target atlas, or using terms not defined in an atlas. In the atlas plates, black text and lines illustrate a term and an area, respectively, corresponding to a target atlas region. Orange text and lines illustrate a term and an area, respectively, corresponding to a sampled region reported for a data set. Thus, the orange text and lines illustrate the region term and area that should be registered to a target atlas term. The Figure defines and illustrates concepts of granularity, coverage, specificity, accuracy, and confidence.

brain section images embedded in the LocaliZoom viewer on EBRAINS, Figure 2A).

The objects supported by LocareJSON (point, sphere, line string, cylinder, polygon, polyhedron, and atlas mesh) are defined in individual schemas. Point representations consist of coordinate triplets, with each triplet defining a specific point in a 3D atlas. Sphere representations build upon point representations and consists of coordinate triplets defining the sphere centroid, with information about radius to create a sphere measured from the centroid. Line string representations consist of two or more coordinate triplets, as a minimum defining the start and end point of a segment. Cylinder representations build upon line string representations with additional information about radius to create a cylinder around the length of the line string. Polygon representations consist of coordinate triplets defining corners of a delimited 2D plane. Polyhedron representations consist of coordinate triplets defining corners of a 3D object (vertices), including information about how vertices create polygons (faces) that can be used to represent 3D objects. Atlas meshes, a unique form of polyhedron, contain the name of a specific term from a 3D atlas, provided by a link to openMINDS_SANDS.

One or several objects can be defined within a locareCollection schema. The schemas for geometric objects include the following required properties: "type," stating the geometric object type, and "coordinates," a coordinate list formatted based on the type. The schema for atlas mesh includes the "parcellationEntityVersion," stating the brain region's URI. Each object also includes a set of properties

pointing to the original data the schema represents. These properties include: the name of the data ("name," required), clearly directing to a subject, file, or group of files; a description of the data ("description," required), e.g., "position of cell body"; and a direct link to the data source for the geometric object ("linkedURI"; optional), e.g., the LocaliZoom viewer link for the individual brain section image used to create spheres shown in Figure 2A.

3.3 The Locare workflow use-cases

To demonstrate the workflow we applied it to represent the location of data from rats and mice acquired by different methods, including electrophysiology (2 data sets), electrocorticography (1 data set), (immuno-)histochemistry (2 data sets), axonal tract tracing (1 data set), neuronal morphology (2 data sets) and calcium imaging (1 data set), all shown in Figure 2. Technical information about the use-cases is provided in Supplementary Table 2. The rat- and mouse brain data sets were co-visualized in the Waxholm Space atlas of the Sprague Dawley rat brain (Papp et al., 2014; Kjonigsen et al., 2015; Osen et al., 2019; Kleven et al., 2023a) or the Allen mouse brain atlas Common Coordinate Framework (Wang et al., 2020), respectively. For each use-case, we utilized a separate route in the Locare workflow, based on the type of location documentation available, resulting in a LocareJSON schema of which the type depended on the object chosen to represent the data (point, line string, sphere, cylinder, polygon, polyhedron, or atlas mesh). Each use-case is available as a LocareJSON file in the LocareJSON repository and as data sets on EBRAINS, where links to their source data sets and detailed methodological descriptions are also provided.

Figure 2 illustrates how different types of neuroscience data can be represented as geometric objects (Figures 2A–I) that can be co-visualized in an atlas space (Figures 2J,K). The geometric data created as examples are available as derived data sets via EBRAINS (Blixhavn et al., 2023a,b,c,d,e; Reiten et al., 2023a,b,c). The derived data sets are listed in Supplementary Table 2, providing links to LocareJSON files for each use case, as well as to the landing page for each derived data set shown in Figure 2. From the landing page, a data descriptor document is provided, explaining how the geometric data were specified following the Locare workflow, and how the LocareJSON file is organized. These resources provide detailed descriptions of the geometric location data, with suggestions of how they can be visualized. The data coordinates provided can, e.g., be co-visualized in an atlas viewer, such as the MeshView tool, available from EBRAINS.[19,20] This tool visualizes brain structures from WHS rat brain atlas and the AMBA CCF mouse brain atlas as geometric meshes and includes a feature for importing point coordinates, such as those provided with our data sets, as shown in Figure 2.

The use-cases demonstrate that the object representation that best represent the data is highly dependent on how the data are made available, and the nature and extent of associated documentation provided with it.

19 https://www.ebrains.eu/tools/meshview

20 https://meshview-for-brain-atlases.readthedocs.io/en/latest/index_.html

4 Discussion

The Locare workflow specifies different ways in which highly variable documentation describing the anatomical location of neuroscience data can be used to create representations of the data as geometric objects in a reference atlas space. The collection of LocareJSON schemas exemplify how such objects can be structured in a machine-readable way. The workflow was established and validated using 280 rat and mouse brain data sets generated using highly different methodologies (Supplementary Table 1). These data sets, shared on the EBRAINS Knowledge Graph between 2018 and 2023, allowed us to categorize the location documentation into three main categories. The geometric object data created for the nine examples used to demonstrate the Locare workflow (Figure 2) are shared as derived data sets on EBRAINS with links to their source data sets (Supplementary Table 2). In our use-cases, coordinates were specified using tools provided via the EBRAINS Research Infrastructure, but numerous other tools for generating 3-D geometric objects and coordinates (see Tyson et al., 2022; Fuglstad et al., 2023) may also be suitable as a starting point to create Locare JSON files. Below, we consider the potential impact, advantages, and limitations of the Locare workflow, including the geometric representations it delivers, and discuss possibilities for utilizing such geometric representations for visualization and spatial queries.

The FAIR guiding principle for data management and stewardship emphasize machine-readability and use of persistent identifiers to optimize reuse of scientific data (Wilkinson et al., 2016). Web-based open data infrastructures, structured metadata, and copyright licenses make data findable, accessible, and re-usable, while use of standardized file formats ensure interoperability of data files with different tools and among similar types of data (Pagano et al., 2013). In the context of the FAIR principles, the Locare workflow allows creation of machine-readable files representing the anatomical location and relevance of different data that otherwise would be difficult to find, access, and compare. By defining geometric objects using atlas-based coordinates, the data representations are spatially integrated and interoperable, in the sense that they can be co-visualized using viewer tools and utilized in various computational processes, including spatial search.

Our use-cases (Figure 2; Supplementary Table 2) show that the usefulness of location documentation depends more on the amount and level of detail of the documentation provided, than the method used to obtain the data. This highlights the need for good reporting practices. It is well established that the amount and consistency of metadata provided with research data varies considerably (see Bjerke et al., 2018a, 2020a), which in turn also contributes to the known problems with low replicability and reproducibility of studies (Goodman et al., 2016; Stupple et al., 2019). The different routes through the Locare workflow accommodates the variability of location documentation typically provided with experimental data sets, thus guiding researchers to define the most specific geometric representations possible with the documentation available for their data sets. In this way, data generated using the same methodology may be represented by different geometric objects when the available metadata differ. The location of a neuronal reconstruction can be defined as a singular point in an atlas (Figure 2G), or only as a mesh representing an entire anatomical subregion when less specific location documentation was provided (Figure 2C). Similarly, a series of histological images registered to an atlas may also be represented in different ways; as polygons representing the locations of sections in

atlas space (use-case B), or as a population of points representing specific cellular features extracted from section images (Figure 2F). Improved routines for recording and sharing location documentation for neuroscience data will enable more precise spatial representation of data (Bjerke et al., 2018a; Tyson and Margrie, 2022; Kleven et al., 2023a).

The most detailed and accurate spatial representations of data are achieved by spatial registration of images showing anatomical features. A range of image registration tools are available (Puchades et al., 2019; Tappan et al., 2019; Carey et al., 2023; for review, see Tyson and Margrie, 2022), tailored for different types of 2D or 3D image data, and compatible with different brain atlases. Both manual and automated methods exist for different applications. Scripts are available for converting the output from the spatial registration tool QuickNII to LocareJSON polygon schema (see Figures 2A,B). Similar scripts can readily be adapted to different tools. Once images are spatially registered to an atlas, they can be used to specify points or volumes of interest, such as labeled objects (Figures 2F,G), electrode recording sites (Figure 2C), or tracer injection sites (Figure 2A).

The location of POIs, derived from text descriptions or extracted from atlas-registered images, can result in any geometric object representation. When coordinates for POIs have been extracted, an important consideration is therefore which geometric object would best represent it. There might be several alternatives, as, e.g., in the case of electrode tracks. A point can be used to represent the end or the entry point of the electrode (although the end point is usually most relevant as this is where recordings are made), and a line string may represent both the end and entry point, which would be appropriate when there are recording sites along the track (see Figure 2I, where a linear electrode array with 16 recording sites along the electrode was used). If the radius of the object (e.g., the electrode) is known, points and line strings may alternatively be replaced by spheres and cylinders, by introducing the radius of the object. Determining a radius should be the preferred practice as it benefits both visualization and spatial query purposes. In many cases, however, information about the radius is missing. Whether a best approximation is the better choice must be evaluated on a case-by-case basis.

The Locare workflow defines how the location of disparate neuroscience data can be represented as geometric objects in an atlas space. The workflow was developed using rat and mouse data sets with associated atlases, tools, and resources shared via the EBRAINS Research Infrastructure. The concept of data integration through geometric representations is generic and system independent, and the Locare workflow is therefore in principle applicable for other species for which an open access 3D brain atlas is available, such as, e.g., the zebrafish larvae (Kunst et al., 2019), macaque (Balan et al., 2024), or human brain (Amunts et al., 2020).

With the Locare workflow, we propose a streamlined approach to specify, organize, and store information about anatomical positions in the brain, yielding machine-readable files suitable for search engines, viewers, and other tools. The focus is to represent the location of data in a simplified and standardized format, rather than aiming to integrate the actual data files. We believe this will ensure the relevance of the workflow even when facing new methods, tools, and file formats. Standardized representation of data as geometric objects in 3D coordinate space can be utilized in spatial queries of neuroscience databases. Spatial queries will likely make it easier for researchers to find and reuse relevant data compared to free-text searches, and possibly open for more analytic approaches for re-use of shared data (Cao et al., 2023).

We envision that the Locare workflow can guide researchers describing anatomical locations in their data, and provide a starting point for defining new standards for current and future platforms, thus making neuroscience data more findable, accessible, interoperable and reusable, in accordance with the principles set forward by Wilkinson et al. (2016). Future work will include extension of the concept and workflow to human and non-human primate data and implementation into software for querying and accessing the location and distribution of neuroscience data through atlases.

Data availability statement

The original contributions presented in the study are included in the article/Supplementary material, further inquiries can be directed to the corresponding author.

Author contributions

CB: Conceptualization, Data curation, Investigation, Methodology, Validation, Visualization, Writing – original draft. IR: Conceptualization, Data curation, Investigation, Methodology, Validation, Visualization, Writing – review & editing. HK: Data curation, Investigation, Methodology, Validation, Writing – review & editing. MØ: Data curation, Investigation, Methodology, Validation, Writing – review & editing. SY: Investigation, Methodology, Validation, Writing – review & editing. US: Data curation, Investigation, Methodology, Validation, Writing – review & editing. MP: Data curation, Investigation, Methodology, Validation, Writing – review & editing. OS: Data curation, Investigation, Methodology, Project administration, Software, Writing – review & editing. JB: Conceptualization, Funding acquisition, Investigation, Project administration, Resources, Supervision, Validation, Writing – review & editing. IB: Conceptualization, Data curation, Methodology, Supervision, Validation, Visualization, Writing – review & editing. TL: Project administration, Resources, Supervision, Visualization, Writing – review & editing, Conceptualization, Funding acquisition, Methodology.

Funding

The author(s) declare financial support was received for the research, authorship, and/or publication of this article. This work was funded by the European Union's Horizon 2020 Framework Programme for Research and Innovation under the Specific Grant Agreement No. 785907 (Human Brain Project SGA2), Specific Grant Agreement No. 945539 (Human Brain Project SGA3), the Specific Grant Agreement No. 101147319 (EBRAINS 2.0) and the Research Council of Norway under Grant Agreement No. 269774 (INCF Norwegian Node, to JB and TL).

Acknowledgments

The present work builds on our earlier contributions to curation of neuroscience data and development of tools and workflows in the Human Brain Project and EBRAINS Research Infrastructure with

contributions from many researchers. We thank Xiao Gui, Timo Dicksheid, Lyuba Zehl, Harry Carey, Rembrandt Bakker, and Tom Gillespie for valuable discussion and contributions during the different stages of developing the Locare workflow and LocareJSON format and Øystein Hagen Blixhavn for valuable technical input and assistance.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

The author(s) declared that they were an editorial board member of Frontiers, at the time of submission. This had no impact on the peer review process and the final decision.

Publisher's note

Supplementary material

The Supplementary material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fninf.2024.1284107/full#supplementary-material

References

Abrams, M. B., Bjaalie, J. G., Das, S., Egan, G. F., Ghosh, S. S., Goscinski, W. J., et al. (2022). A standards Organization for Open and FAIR neuroscience: the international Neuroinformatics coordinating facility. *Neuroinformatics* 20, 25–36. doi: 10.1007/s12021-020-09509-0

Amunts, K., Knoll, A., Lippert, T., Pennartz, C., Ryvlin, P., Destexhe, A., et al. (2019). The human brain project—synergy between neuroscience, computing, informatics, and brain-inspired technologies. *PLoS Biol.* 17:e3000344. doi: 10.1371/journal.pbio.3000344

Amunts, K., Mohlberg, H., Bludau, S., and Zilles, K. (2020). Julich-brain: a 3D probabilistic atlas of the human brain's cytoarchitecture. *Science* 369, 988–992. doi: 10.1126/science.abb4588

Arena, A., Nilsen, A., Thon, S., and Storm, J. (2020). Test of consciousness metrics in rodents. [Data set]. EBRAINS. doi: 10.25493/8CQN-Y8S

Arena, A., and Storm, J. (2018). Large scale multi-channel EEG in rats. [Data set]. EBRAINS. doi: 10.25493/4SPM-V00

Arena, A., Thon, S., and Storm, J. (2019a). Mechanistic analysis of ERP in rodents. [Data set]. EBRAINS. doi: 10.25493/5ZJY-PHB

Arena, A., Thon, S., and Storm, J. (2019b). PCI-like measure in rodents. [Data set]. EBRAINS. doi: 10.25493/S0DM-BK5

Ascoli, G., Maraver, P., Nanda, S., Polavaram, S., and Armañanzas, R. (2017). Win–win data sharing in neuroscience. *Nat. Methods* 14, 112–116. doi: 10.1038/nmeth.4152

Bakker, R., Tiesinga, P., and Kötter, R. (2015). The scalable brain atlas: instant web-based access to public brain atlases and related content *Neuroinform* 13, 353–366. doi: 10.1007/s12021-014-9258-x

Balan, P., Zhu, Q., Li, X., Niu, M., Rapan, L., Funck, T., et al. (2024). MEBRAINS 1.0: a new population-based macaque atlas. Imaging. *Neuroscience*. doi: 10.1162/imag_a_00077

Bassett, D., and Sporns, O. (2017). Network neuroscience. *Nat. Neurosci.* 20, 353–364. doi: 10.1038/nn.4502

Bjerke, I. E. (2021). Quantifying cellular parameters across the murine brain: New practices for integrating and analysing neuroscience data using 3D brain atlases. Available at: https://www.duo.uio.no/handle/10852/83758?show=full

Bjerke, I., Øvsthus, M., Andersson, K., Blixhavn, C., Kleven, H., Yates, S., et al. (2018a). Navigating the murine brain: toward best practices for determining and documenting neuroanatomical locations in experimental studies. *Front. Neuroanat.* 12, 1–15. doi: 10.3389/fnana.2018.00082

Bjerke, I., Øvsthus, M., Papp, E., Yates, S., Silvestri, L., Fiorilli, J., et al. (2018b). Data integration through brain atlasing: human brain project tools and strategies. *Eur. Psychiatry* 50, 70–76. doi: 10.1016/j.eurpsy.2018.02.004

Bjerke, I., Puchades, M., Bjaalie, J., and Leergaard, T. (2020a). Database of literature derived cellular measurements from the murine basal ganglia. *Sci. Data* 7:211. doi: 10.1038/s41597-020-0550-3

Bjerke, I., Schlegel, U., Puchades, M., Bjaalie, J., and Leergaard, T. (2019a). Franklin & Paxinos' "the mouse brain in stereotaxic coordinates" (3rd edition) spatially registered to the Allen mouse common coordinate framework. [Data set]. EBRAINS. doi: 10.25493/WFCZ-FSN

Bjerke, I., Schlegel, U., Puchades, M., Bjaalie, J., and Leergaard, T. (2019b). Paxinos & Watson's "the rat brain in stereotaxic coordinates" (7th edition) spatially registered to the Waxholm space atlas of the rat brain. [data set]. EBRAINS. doi: 10.25493/APWV-37H

Bjerke, I., Yates, S., Carey, H., Bjaalie, J., and Leergaard, T. (2023). Scaling up cell-counting efforts in neuroscience through semi-automated methods. *iScience* 26:107562. doi: 10.1016/j.isci.2023.107562

Bjerke, I., Yates, S., Puchades, M., Bjaalie, J., and Leergaard, T. (2020b). Brain-wide quantitative data on parvalbumin positive neurons in the mouse. [Data set]. EBRAINS. doi: 10.25493/BT8X-FN9

Blixhavn, C., Bjerke, I., Reiten, I., and Leergaard, T. (2023a). 3D atlas locations of rat brain section images showing nissl bodies, zincergic terminal fields and metal-containing glia. [Data set] EBRAINS. doi: 10.25493/QFFN-H67

Blixhavn, C., Haug, F., Kleven, H., Puchades, M., Bjaalie, J., and Leergaard, T. (2022). Multiplane microscopic atlas of rat brain zincergic terminal fields and metal-containing glia stained with Timm's sulphide silver method (v1) [data set] EBRAINS. doi: 10.25493/T686-7BX

Blixhavn, C., Ovsthus, M., Reiten, I., and Leergaard, T. (2023b). 3D atlas location of the field of view of a calcium imaging recording in mouse after stroke (v1). [Data set] EBRAINS. doi: 10.25493/TS1A-K28

Blixhavn, C., Ovsthus, M., Reiten, I., and Leergaard, T. (2023c). 3D atlas locations of mouse thalamocortical projection neuron reconstructions. [Data set] EBRAINS. doi: 10.25493/KGCK-773

Blixhavn, C., Reiten, I., and Leergaard, T. (2023d). 3D atlas location of electrode recordings in mice during presentation of figure-ground stimuli. [Data set] EBRAINS. doi: 10.25493/H52W-QF4

Blixhavn, C., Reiten, I., Ovsthus, M., Bjerke, I., Puchades, M., and Leergaard, T. (2023e). 3D atlas locations of epidural electrode EEG recordings in rats. [Data set] EBRAINS. doi: 10.25493/AK1G-WQQ

Blixhavn, C., Reiten, I., Yates, S., and Leergaard, T. (2023f). 3D atlas locations of parvalbumin positive neurons in the adult mouse brain. [Data set] EBRAINS. doi: 10.25493/Q52E-ESE

Cao, R., Ling, Y., Meng, J., Jiang, A., Luo, R., He, Q., et al. (2023). SMDB: a spatial multimodal data browser. *Nucleic Acids Res.* 51, W553–W559. doi: 10.1093/nar/gkad413

Carey, H., Pegios, M., Martin, L., Saleeba, C., Turner, A., Everett, N., et al. (2023). DeepSlice: rapid fully automatic registration of mouse brain imaging to a volumetric atlas. *Nat. Commun.* 14:5884. doi: 10.1038/s41467-023-41645-4

Clarkson, M. D. (2016). Representation of anatomy in online atlases and databases: a survey and collection of patterns for interface design. *BMC Dev. Biol.* 16:18. doi: 10.1186/s12861-016-0116-y

Conti, E., Pavone, F., and Allegra Mascaro, A. (2019). Fluorescence cortical recording of mouse activity after stroke. [Data set] EBRAINS. doi: 10.25493/Z9J0-ZZQ

Feldmeyer, D., Qi, G., and Yang, D. (2020). Morphological data of cortical layer 6 neurons and synaptically coupled neuronal pairs. [Data set] EBRAINS. doi: 10.25493/YMV3-45H

Ferguson, A., Nielson, J., Cragin, M., Bandrowski, A., and Martone, M. (2014). Big data from small data: data-sharing in the 'long tail' of neuroscience. *Nat. Neurosci.* 17, 1442–1447. doi: 10.1038/nn.3838

Fiorilli, J., Marchesi, P., Ruikes, T., Buckton, R., Quintero, M., Reitan, I., et al. (2023). Neural correlates of object identity and reward outcome in the corticohippocampal hierarchy: double dissociation between perirhinal and secondary visual cortex. bioRxiv [Preprint], bioRxiv: 2023.05.24.542117

Fiorilli, J., Ruikes, T., Huis, G., and Pennartz, C. (2022). Sensory, perirhinal and hippocampal tetrode recordings during visual, tactile and visuotactile discrimination task in the freely moving rat (v1). [Data set] EBRAINS. doi: 10.25493/AM91-2D

Franklin, K., and Paxinos, G. (2007). *The mouse brain in stereotaxic coordinates. 3rd* Edn. San Diego: Elsevier Academic Press.

Fuglstad, J. G., Saldanha, P., Paglia, J., and Whitlock, J. R. (2023). Histological E-data registration in rodent brain spaces. *Elife* 12:e83496. doi: 10.7554/eLife.83496

Fürth, D., Vaissière, T., Tzortzi, O., Xuan, Y., Märtin, A., Lazaridis, I., et al. (2018). An interactive framework for whole-brain maps at cellular resolution. *Nat. Neurosci.* 21:895. doi: 10.1038/s41593-017-0058-0

García-Amado, M., Porrero, C., Rubio, M., Evangelio, M., and Clascá, F. (2020). 3D reconstruction and measurement of individual thalamocortical projection neuron axons of somatosensory and visual thalamic nuclei. [Data set] EBRAINS. doi: 10.25493/AWS5-MZG

Goodman, S., Fanelli, D., and Ioannidis, J. (2016). What does research reproducibility mean? *Sci. Transl. Med.* 8:341ps12. doi: 10.1126/scitranslmed.aaf5027

Groeneboom, N., Yates, S., Puchades, M., and Bjaalie, J. (2020). Nutil: a pre- and post-processing toolbox for histological rodent brain section images. *Front. Neuroinform.* 14, 1–9. doi: 10.3389/fninf.2020.00037

Gurdon, B., Yates, S., Csucs, G., Groeneboom, N. E., Hadad, N., Telpoukhovskaia, M., et al. (2023). Detecting the effect of genetic diversity on brain composition in an Alzheimer's disease mouse model. *bioRxiv* [preprint].

Jorgenson, L. A., Newsome, W. T., Anderson, D. J., Bargmann, C. I., Brown, E. N., Deisseroth, K., et al. (2015). The BRAIN initiative: developing technology to catalyse neuroscience discovery. *Philos. Trans. R. Soc. B* 370:20140164. doi: 10.1098/rstb.2014.0164

Kim, Y., Yang, G., Pradhan, K., Venkataraju, K., Bota, M., Garcie del Molino, L., et al. (2017). Brain-wide maps reveal stereotyped cell-type-based cortical architecture and subcortical sexual resource brain-wide maps reveal stereotyped cell-type-based cortical architecture and subcortical sexual dimorphism. *Cell* 171, 456–469. doi: 10.1016/j.cell.2017.09.020

Kjonigsen, L., Lillehaug, S., Bjaalie, J., Witter, M., and Leergaard, T. (2015). Waxholm space atlas of the rat brain hippocampal region: three-dimensional delineations based on magnetic resonance and diffusion tensor imaging. *Neuroimage* 108, 441–449. doi: 10.1016/j.neuroimage.2014.12.080

Klein, S., Staring, M., Murphy, K., Viergever, M., and Pluim, J. (2010). Elastix: a toolbox for intensity-based medical image registration. *IEEE Trans. Med. Imaging* 29, 196–205. doi: 10.1109/TMI.2009.2035616

Kleven, H., Bjerke, I., Clascá, F., Groenewegen, H., Bjaalie, J., and Leergaard, T. (2023a). Waxholm space atlas of the rat brain: a 3D atlas supporting data analysis and integration. *Nat. Methods* 20, 1822–1829. doi: 10.1038/s41592-023-02034-3

Kleven, H., Reiten, I., Blixhavn, C., Schlegel, U., Øvsthus, M., Papp, E., et al. (2023b). A neuroscientist's guide to using murine brain atlases for efficient analysis and transparent reporting. *Front. Neuroinform.* 17, 1–8. doi: 10.3389/fninf.2023.1154080

Kondo, H., Olsen, G., Gianatti, M., Monterotti, B., Sakshaug, T., and Witter, M. (2022). Anterogradely labeled axonal projections from the orbitofrontal cortex in rat (v1). [Data set] EBRAINS. doi: 10.25493/2MX9-3XF

Kunst, M., Laurell, E., Mokayes, N., Kramer, A., Kubo, F., Fernandes, A., et al. (2019). A cellular-resolution atlas of the larval zebrafish brain. *Neuron* 103, 21–38.e5. doi: 10.1016/j.neuron.2019.04.034

Leergaard, T., and Bjaalie, J. (2022). Atlas-based data integration for mapping the connections and architecture of the brain. *Science* 378, 488–492. doi: 10.1126/science.abq2594

Martone, M., Gupta, A., and Ellisman, M. (2004). E-neuroscience: challenges and triumphs in integrating distributed data from molecules to brains. *Nat. Neurosci.* 7, 467–472. doi: 10.1038/nn1229

Niedworok, C., Brown, A., Cardoso, M., Osten, P., Ourselin, S., Modat, M., et al. (2016). AMAP is a validated pipeline for registration and segmentation of high-resolution mouse brain data. *Nat. Commun.* 7:11879. doi: 10.1038/ncomms11879

Oh, S., Harris, J., Ng, L., Winslow, B., Cain, N., Mihalas, S., et al. (2014). A mesoscale connectome of the mouse brain. *Nature* 508, 207–214. doi: 10.1038/nature13186

Osen, K., Imad, J., Wennberg, A., Papp, E., and Leergaard, T. (2019). Waxholm space atlas of the rat brain auditory system: three-dimensional delineations based on structural and diffusion tensor magnetic resonance imaging. *Neuroimage* 199, 38–56. doi: 10.1016/j.neuroimage.2019.05.016

Pagano, P., Candela, L., and Castelli, D. (2013). Data interoperability. *Data Sci. J.* 12, GRDI19–GRDI25. doi: 10.2481/dsj.GRDI-004

Papp, E., Leergaard, T., Calabrese, E., Johnson, G., and Bjaalie, J. (2014). Waxholm space atlas of the Sprague Dawley rat brain. *Neuroimage* 97, 374–386. doi: 10.1016/j.neuroimage.2014.04.001

Paxinos, G., and Watson, C. (2018). *Paxinos and Watson's The rat brain in stereotaxic coordinates compact. 7th* Edn. San Diego: Elsevier Academic Press.

Puchades, M., Csucs, G., Ledergerber, D., Leergaard, T., and Bjaalie, J. (2019). Spatial registration of serial microscopic brain images to three-dimensional reference atlases with the QuickNII tool. *PloS One* 14:e0216796. doi: 10.1371/journal.pone.0216796

Reiten, I., Blixhavn, C., Bjerke, I., and Leergaard, T. (2023a). 3D atlas location of rat cortical neuron reconstructions. [Data set] EBRAINS. doi: 10.25493/CBTH-1G9

Reiten, I., Blixhavn, C., and Leergaard, T. (2023b). 3D atlas locations of rat brain injection sites and section images from tract tracing experiments involving the orbitofrontal cortex. [Data set] EBRAINS. doi: 10.25493/R8E4-YKU

Reiten, I., Ovsthus, M., Schlegel, U., Blixhavn, C., and Leergaard, T. (2023c). 3D atlas locations of tetrode recordings in rats performing a cross-modal memory recall task. [Data set] EBRAINS. doi: 10.25493/T5VW-SRJ

Resta, F., Allegra Mascaro, A. L., and Pavone, F. (2021). Study of slow waves (SWs) propagation through wide-field calcium imaging of the right cortical hemisphere of GCaMP6f mice (v2). [Data set] EBRAINS. doi: 10.25493/QFZK-FXS

Schnabel, U., Lorteije, J., Self, M., and Roelfsema, P. (2020). Neuronal activity elicited by figure-ground stimuli in primary visual cortex of the awake mouse. [Data set] EBRAINS. doi: 10.25493/NHHM-1S5

Sejnowski, T., Churchland, P., and Movshon, J. (2014). Putting big data to good use in neuroscience. *Nat. Neurosci.* 17, 1440–1441. doi: 10.1038/nn.3839

Stupple, A., Singerman, D., and Celi, L. (2019). The reproducibility crisis in the age of digital medicine. *NPJ Digit. Med.* 2:79. doi: 10.1038/s41746-019-0079-z

Tappan, S., Eastwood, B., O'Connor, N., Wang, Q., Ng, L., Feng, D., et al. (2019). Automatic navigation system for the mouse brain. *J. Comp. Neurol.* 527, 2200–2211. doi: 10.1002/cne.24635

Tyson, A., and Margrie, T. (2022). Mesoscale microscopy and image analysis tools for understanding the brain. *Prog. Biophys. Mol. Biol.* 168, 81–93. doi: 10.1016/j.pbiomolbio.2021.06.013

Tyson, A. L., Vélez-Fort, M., Rousseau, C. V., Cossell, L., Tsitoura, C., Lenzi, S. C., et al. (2022). Accurate determination of marker location within whole-brain microscopy images. *Sci. Rep.* 12:867. doi: 10.1038/s41598-021-04676-9

Wang, Q., Ding, S., Li, Y., Royall, J., Feng, D., Lesnar, P., et al. (2020). The Allen mouse brain common coordinate framework: a 3D reference atlas. *Cell* 181, 936–953.e20. doi: 10.1016/j.cell.2020.04.007

Wilkinson, M., Dumontier, M., Aalbersberg, I., Appleton, G., Axton, M., Baak, A., et al. (2016). The FAIR guiding principles for scientific data management and stewardship. *Sci. Data* 3, 160018–160019. doi: 10.1038/sdata.2016.18

Yates, S., Groeneboom, N., Coello, C., Lichtenthaler, S., Kuhn, P., Demuth, H., et al. (2019). QUINT: workflow for quantification and spatial analysis of features in histological images from rodent brain. *Front. Neuroinform.* 13, 1–14. doi: 10.3389/fninf.2019.00075

Zaslavsky, I., Baldock, R. A., and Boline, J. (2014). Cyberinfrastructure for the digital brain: spatial standards for integrating rodent brain atlases. *Front. Neuroinform.* 8:74. doi: 10.3389/fninf.2014.00074

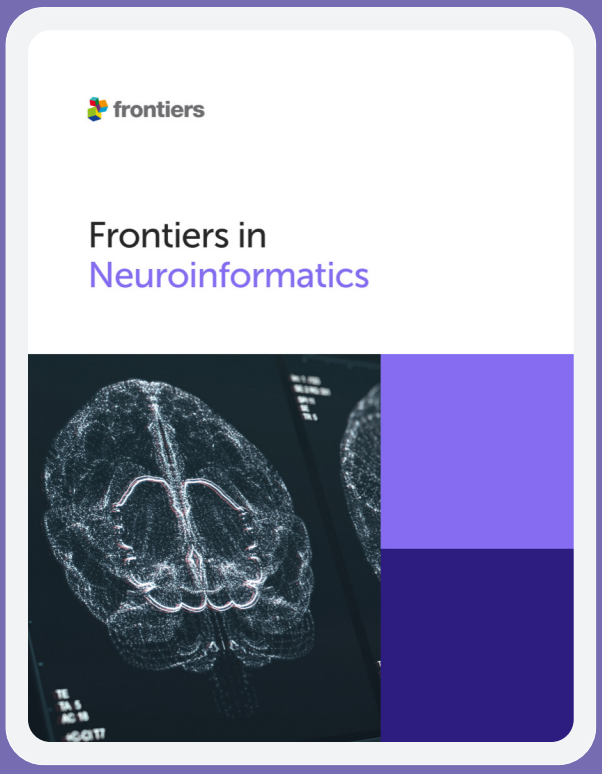
Frontiers in Neuroinformatics

Leading journal supporting neuroscience in the information age

Part of the most cited neuroscience journal series, developing computational models and analytical tools used to share, integrate and analyze experimental data about the nervous system functions.

Discover the latest Research Topics

See more →



Frontiers

Avenue du Tribunal-Fédéral 34
1005 Lausanne, Switzerland
frontiersin.org

Contact us

+41 (0)21 510 17 00
frontiersin.org/about/contact