

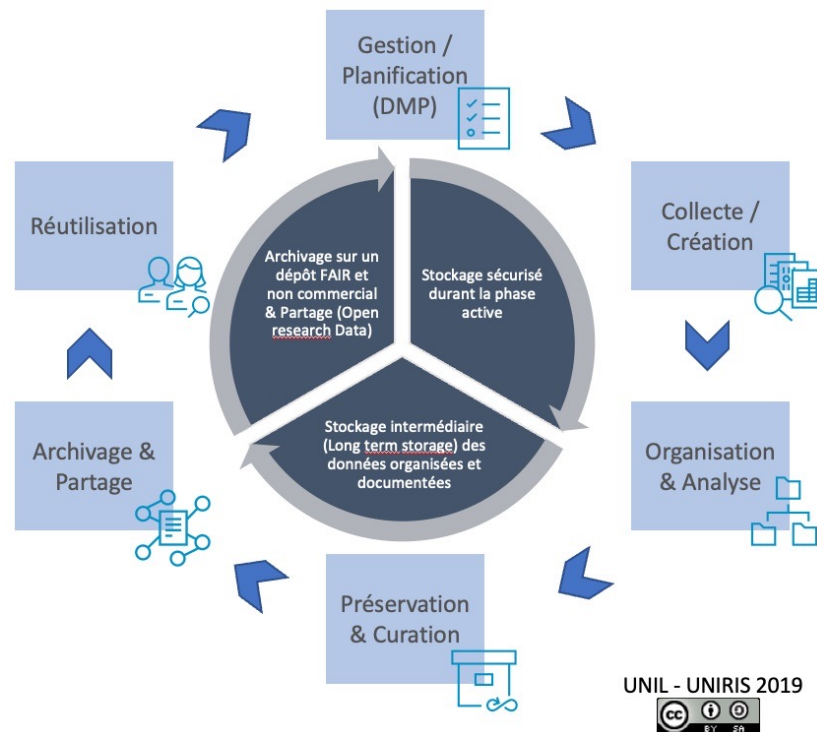
The management of neuroscience data

Olivier Coulon

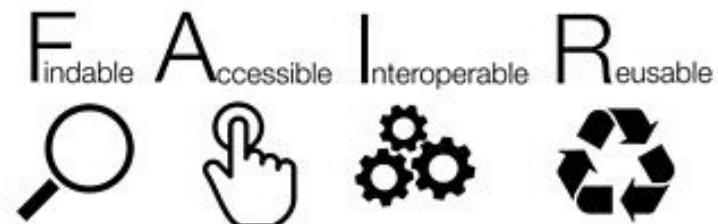
Institut de Neurosciences de la Timone

<http://int.univ-amu.fr>

Cycle de vie des données de recherche



The FAIR formalism



Findable

The first step in (re)using data is to find them. [Metadata](#) and data should be easy to find for both humans and computers. [Machine-readable](#) metadata are essential for automatic [discovery](#) of datasets and services, so this is an essential component of the FAIRification process.

Accessible

Once the user finds the required data, they need to know how they can be accessed, possibly including [authentication](#) and [authorisation](#).

Interoperable

The data usually need to be integrated with other data. In addition, the data need to interoperate with applications or workflows for [analysis](#), [storage](#), and [processing](#).

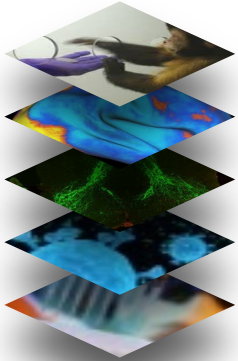
Reusable

The ultimate goal of FAIR is to optimise the reuse of data. To achieve this, metadata and data should be well-described so that they can be replicated and/or combined in different settings.

The principles refer to three types of entities: **data** (or any digital object), **metadata** (information about that digital object), and **infrastructure**. For instance, principle F4 defines that both metadata and data are registered or indexed in a searchable resource (the infrastructure component).

Managing data at institute-level: an example

Platforms



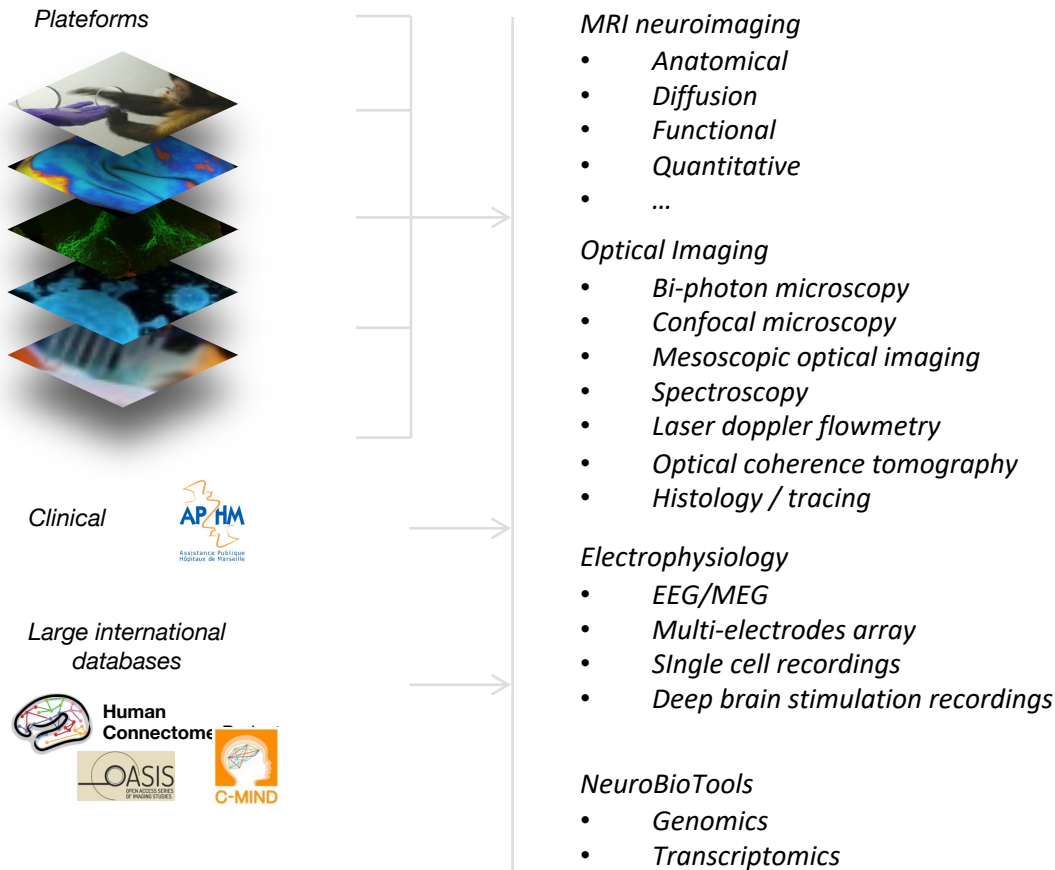
Clinical



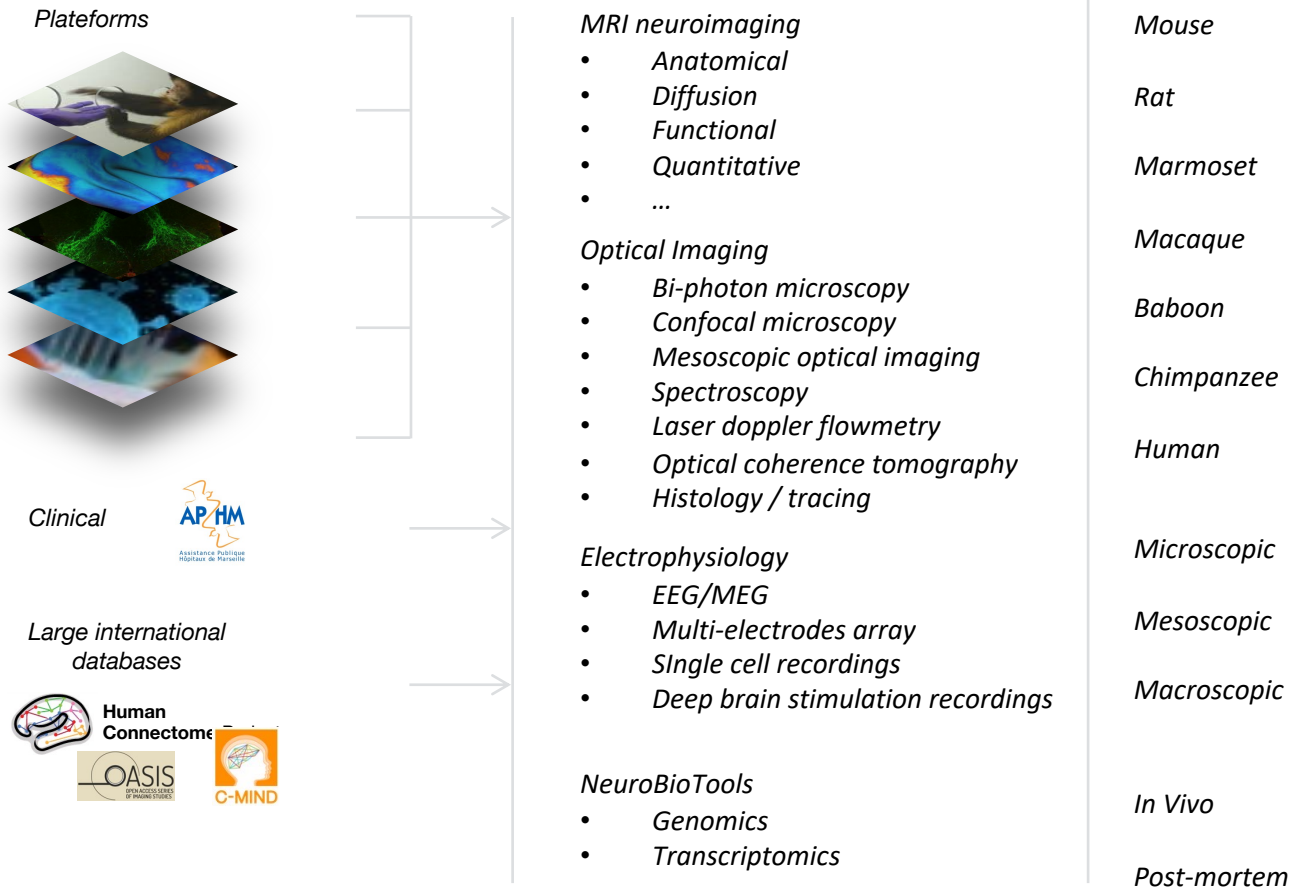
Large international databases



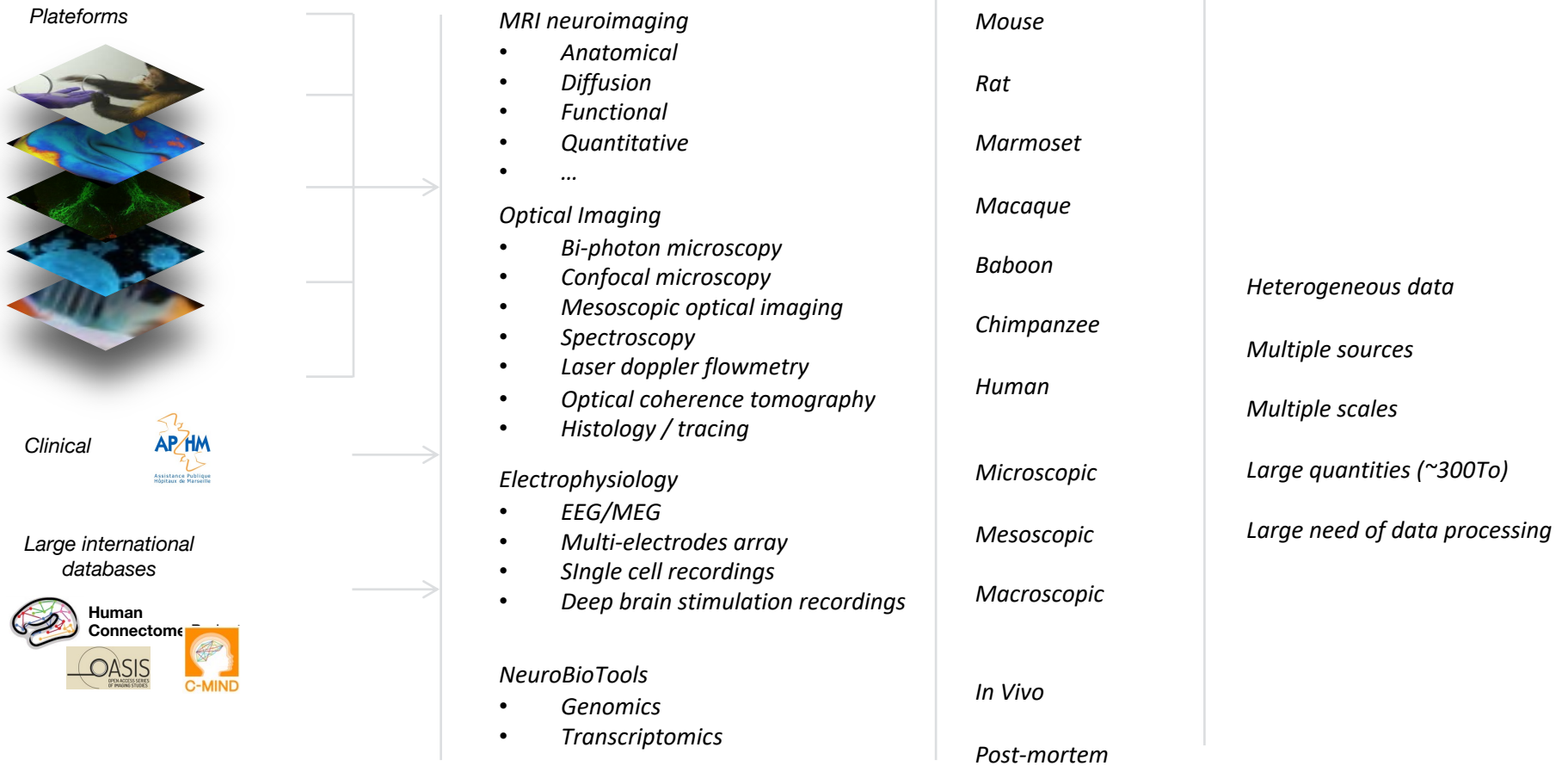
Managing data at institute-level: an example



Managing data at institute-level: an example



Managing data at institute-level: an example



Findable: where's my data ?

Findable: where's my data ?

« On a portable hard drive. My PhD student has got it. I'll email him »

Non secure and unreliable storage. No backup.

Major risk: Complete data loss

Other risks: loss of associated data and impossibility to reprocess.

Findable: where's my data ?

« On a portable hard drive. My PhD student has got it. I'll email him »

Non secure and unreliable storage. No backup.

Major risk: Complete data loss

Other risks: loss of associated data and impossibility to reprocess.

« On a workstation in the experimental room. From time to time I make a copy of the hard drive. »

Non secure storage. Random backup.

Risk: data loss

Other risks: loss of associated data and impossibility to reprocess.

Findable: where's my data ?

« On a portable hard drive. My PhD student has got it. I'll email him »

Non secure and unreliable storage. No backup.

Major risk: Complete data loss

Other risks: loss of associated data and impossibility to reprocess.

« On a workstation in the experimental room. From time to time I make a copy of the hard drive. »

Non secure storage. Random backup.

Risk: data loss

Other risks: loss of associated data and impossibility to reprocess.

« On a (professional level) storage server »

Secure storage, guaranteed backup

Can we find the data, can we proceed to new analyses ?

Rationalizing data management. Goals and motivations

To eliminate all possibility of data loss

Rationalizing data management. Goals and motivations

To eliminate all possibility of data loss

To offer an easy and reliable access to all data using specific queries



databasing, indexation

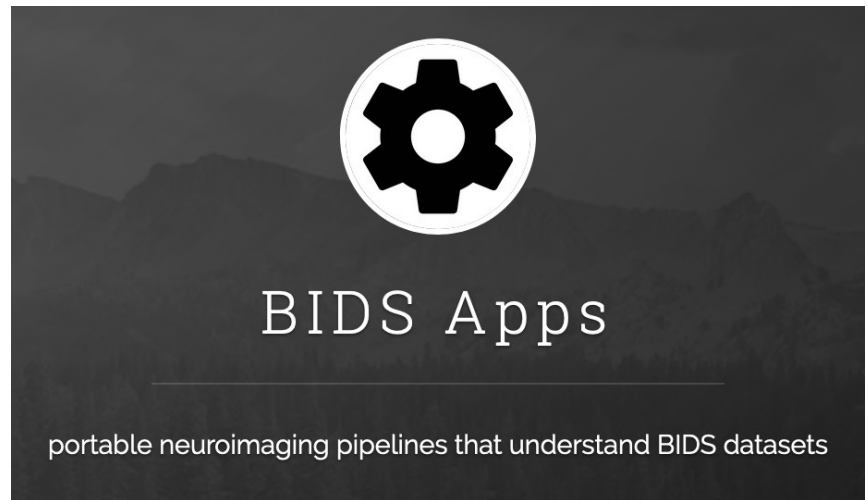
Rationalizing data management. Goals and motivations

To eliminate all possibility of data loss

To offer an easy and reliable access to all data using specific queries

To ease or automate data processing

Formatage / standardisation du stockage



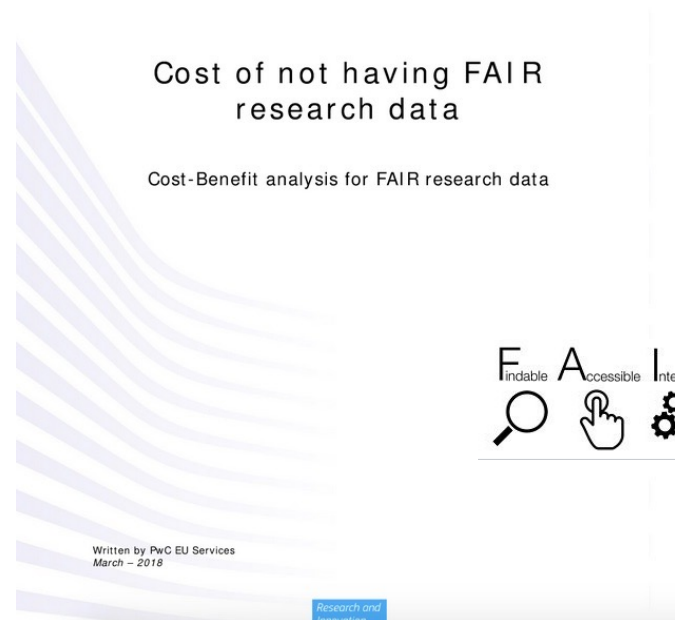
Rationalizing data management. Goals and motivations

To eliminate all possibility of data loss

To offer an easy and reliable access to all data using specific queries

To ease or automate data processing

Reduce costs



Likely cost of not having FAIR research data

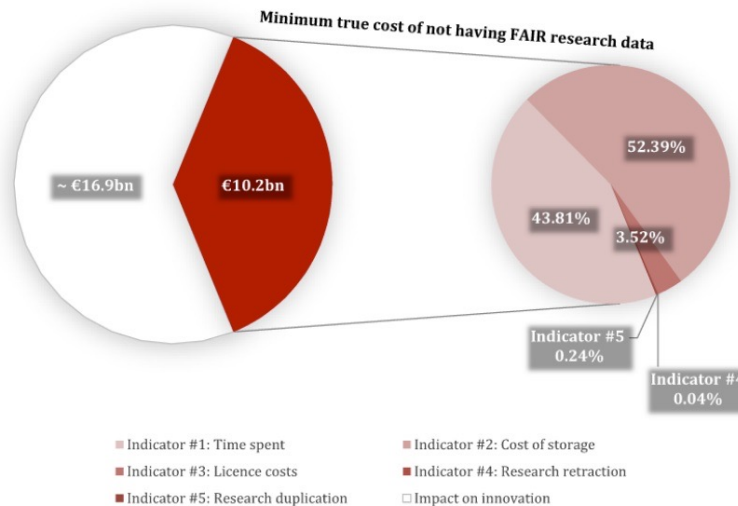


Figure 5: Cost breakdown

« We estimate the annual cost of not having FAIR data to a minimum of €10.2bn per year. The actual cost is likely to be much higher due to unquantifiable elements such as the value of improved research quality and other indirect positive spill-over effects of FAIR research data. »

Rationalizing data management. Goals and motivations

To eliminate all possibility of data loss

To offer an easy and reliable access to all data using specific queries

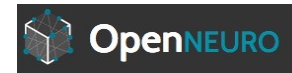
To ease or automate data processing

Reduce costs

To facilitate data sharing between researchers, and/or journals requiring an access to experimental data

Universal formatting of data

eLife39497



uploaded on 09/03/2018 - 11 days ago
last modified 09/03/2018 - 10 days ago
authored by Jiefeng Jiang, Anthony D. Wagner, Tobias Egner

★ 0 🐾 1

Files: 2836, Size: 7.99GB, Subjects: 22, Session: 1

Available Tasks : Main

Available Modalities : T1w, bold

🟢 Published 🧑 Uploader is Following

AUTHORS

Jiefeng Jiang, Anthony D. Wagner, Tobias Egner

BIDS Validation

✔ Valid

Dataset File Tree

eLife39497

- dataset_description.json
 - 📄 DOWNLOAD
 - 👁 VIEW
- participants.tsv
 - 📄 DOWNLOAD
 - 👁 VIEW
- task-Main_bold.json
 - 📄 DOWNLOAD
 - 👁 VIEW
- sourcedata

Rationalizing data management. Goals and motivations

To eliminate all possibility of data loss

To offer an easy and reliable access to all data using specific queries

To ease or automate data processing

Reduce costs

To facilitate data sharing between researchers, and/or journals requiring an access to experimental data

To propose a Data Management Plan to researchers



Qu'est-ce que le pilote de libre accès aux données de recherche Open Research Data ?

Il s'agit d'une opération pilote tendant à rendre accessible au plus grand nombre d'utilisateurs les données de recherche générées dans des projets financés dans le cadre du programme Horizon 2020.

Les bénéficiaires qui y sont tenus doivent rendre accessibles gratuitement les données de recherche issues des projets financés.

Le Work Programme définit les domaines dans lesquels le pilote est applicable.

De quelles données s'agit-il ?

- données et métadonnées nécessaires à la validation des publications : obligatoire ;
- autres données et métadonnées que le bénéficiaire a choisi de diffuser en accès ouvert : spécifiées dans le plan de gestion des données ou **DMP - "Data Management Plan"**.

Si certaines données ne pourront être rendues accessibles, cela devra être justifié dans le DMP (risque de compromettre le projet, raisons éthiques, réglementation relative aux données personnelles, propriété intellectuelle, sécurité...).

Où ?

Dans une base de données de recherche - "**research data repository**" - permettant de

Qu'est-ce que le DMP - Data Management Plan ?

- le DMP est un livrable du projet attendu dans les 6 premiers mois de la vie du projet (des améliorations du DMP peuvent également faire l'objet de livrables subséquents) ;
- le DMP est obligatoire dans les projets inscrits au pilote Open Research Data ;
- le DMP décrit comment les données de recherche collectées ou générées seront gérées pendant et après le projet (méthodologie, standards...), quelles données seront partagées ou diffusées en Open Data, mais aussi comment les données seront conservées ;
- le DMP n'est pas contenu dans la proposition de projet soumise et ne fait pas partie de l'évaluation.
En revanche, dans les actions de recherche et d'innovation (RIA) et les actions d'innovation (IA) le "**template proposal**" inclut une section management des données de recherches, évaluée sous le critère "impact".

Rationalizing data management. Goals and motivations

To eliminate all possibility of data loss

To offer an easy and reliable access to all data using specific queries

To ease or automate data processing

Reduce costs

To facilitate data sharing between researchers, and/or journals requiring an access to experimental data

To propose a Data Management Plan to researchers

To promote and facilitate reproducible and open science



Ministère de l'Enseignement supérieur, de la Recherche et de l'Innovation

Liberté • Égalité • Fraternité
RÉPUBLIQUE FRANÇAISE

MINISTÈRE STRATÉGIE ENSEIGNEMENT SUPÉRIEUR RECHERCHE INNOVATION

[Accueil](#) > [Recherche](#)

RECHERCHE

Le Plan national pour la science ouverte : les résultats de la recherche scientifique ouverts à tous, sans entrave, sans délai, sans paiement

Le Plan national pour la science ouverte annoncé par Frédérique Vidal, le 4 juillet 2018, rend obligatoire l'accès ouvert pour les publications et pour les données issues de recherches financées sur projets. Il met en place un Comité pour la science ouverte et soutient des initiatives majeures de structuration du paysage concernant les publications et les données. Enfin, il est doté d'un volet formation et d'un volet international qui sont essentiels à la mobilisation des communautés scientifiques et à l'influence de la France dans ce paysage en cours de constitution.

Actualité - 1ère publication : 4.07.2018 - Mise à jour : 12.07.2018

Rationalizing data management. Goals and motivations

To eliminate all possibility of data loss

To offer an easy and reliable access to all data using specific queries

To ease or automate data processing

Reduce costs

To facilitate data sharing between researchers, and/or journals requiring an access to experimental data

To propose a Data Management Plan to researchers

To promote and facilitate reproducible and open science



La Plateforme canadienne de neuroscience ouverte (PCNO)

La Plateforme canadienne de neuroscience ouverte (PCNO) a pour mission de mettre en place une plateforme nationale pour le libre échange de données issues de la recherche en neurosciences. Cette plateforme réunit bon nombre des meilleurs chercheurs en neurosciences cliniques et fondamentales, informaticiens et experts en politique scientifique du pays pour former un réseau interactif de collaboration pour la recherche sur le cerveau, l'enseignement interdisciplinaire, les partenariats internationaux, les applications cliniques et la publication ouverte.

La plateforme fournira une interface unifiée à la communauté scientifique et propulsera la recherche canadienne en neurosciences par le partage de données et de méthodes, la création de bases de données à grande échelle, le développement de normes de partage, la facilitation de stratégies d'analyses avancées, la dissémination ouverte de données et de méthodes en neurosciences à la collectivité mondiale et la mise en place de programmes de formation pour la prochaine génération de chercheurs en neurosciences computationnelles. La PCNO vise à éliminer les barrières techniques entravant la science ouverte et à améliorer l'accessibilité et la réutilisabilité de la recherche en neurosciences pour accélérer le rythme auquel les découvertes sont faites.

Rationalizing data management. Goals and motivations

To eliminate all possibility of data loss

To offer an easy and reliable access to all data using specific queries

To ease or automate data processing

Reduce costs

To facilitate data sharing between researchers, and/or journals requiring an access to experimental data

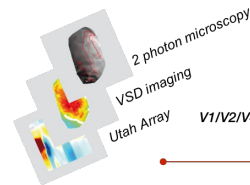
To propose a Data Management Plan to researchers

To promote and facilitate reproducible and open science

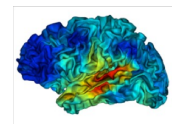
To facilitate scientific projects using heterogeneous multi-modal data, or to facilitate machine learning

Dynamics of cortical maps for decision, action and perception
(Teams: CoMCo, NeOpto, InViBe, BanCO)

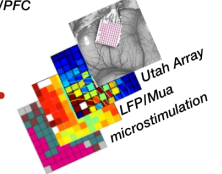
Visual maps for motion computation



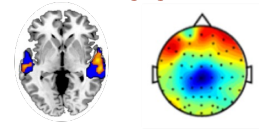
Auditory maps for voice processing



Motor maps for hand/eye movements



Human and monkey behaviours and brain imaging



The 3 pillars of good data management

Storage

Must guarantee security and regular data backup

All data must be stored as automatically as possible on storage servers



No loss

The 3 pillars of good data management

Storage

Must guarantee security and regular data backup

All data must be stored as automatically as possible on storage servers



No loss

Indexing

Ensures that the data is traceable, and possibly accessible according to specific queries based on descriptive metadata

This indexation is usually performed via a database engine.



Access

The 3 pillars of good data management

Storage

Must guarantee security and regular data backup

All data must be stored as automatically as possible on storage servers



No loss

Indexing

Ensures that the data is traceable, and possibly accessible according to specific queries based on descriptive metadata

This indexation is usually performed via a database engine.



Access

Structuring

Standardised nomenclature defining storage and organization of data and associated metadata.

Ensures that data can be exchanged and analysed autonomously



Sharing, reproducing, reusing

The 3 pillars of good data management

Storage

Must guarantee security and regular data backup

All data must be stored as automatically as possible on storage servers

No loss

Indexing

Ensures that the data is traceable, and possibly accessible according to specific queries based on descriptive metadata

This indexation is usually performed via a database engine.

Access

Structuring

Standardised nomenclature defining storage and organization of data and associated metadata.

Ensures that data can be exchanged and analysed autonomously

Sharing, reproducing, reusing

Automatic processing



Standard data structure for MRI data

Open & community based project

Growing ecosystem: Validation, Database integration, BIDS Apps

Extensions for related modalities

MEG

EEG

I EEG

PET

Physiological data (respiration, cardiac activity, ...)

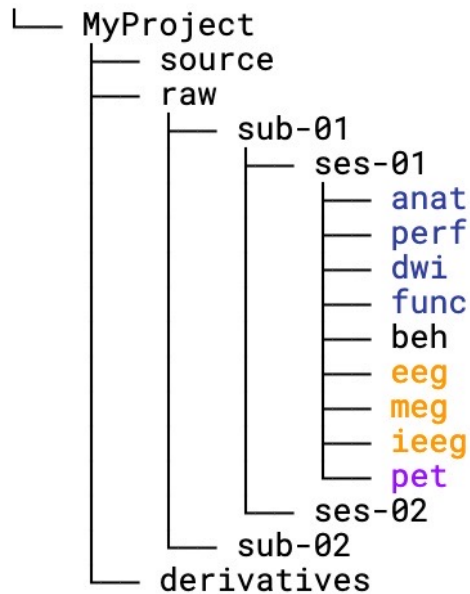
Behavioural data

Microscopy

NIRS

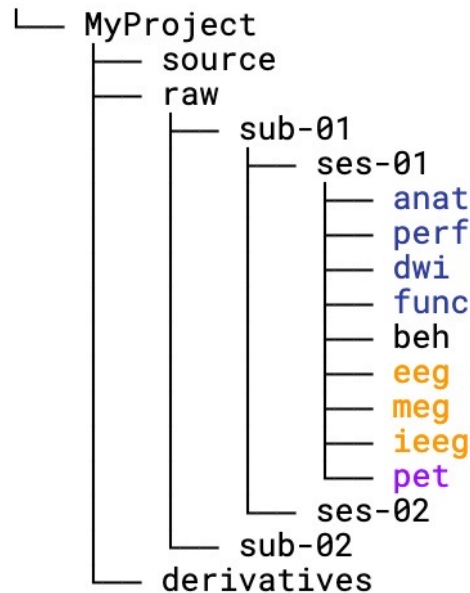
- BIDS is:
 - organizing data in a folder
 - naming files
 - documenting metadata
 - facilitating re-use by your future self and others

- BIDS is:
 - organizing data in a folder
 - naming files
 - documenting metadata
 - facilitating re-use by your future self and others
- BIDS defines a data structure:



- BIDS is:
 - organizing data in a folder
 - naming files
 - documenting metadata
 - facilitating re-use by your future self and others

- BIDS defines a data structure:

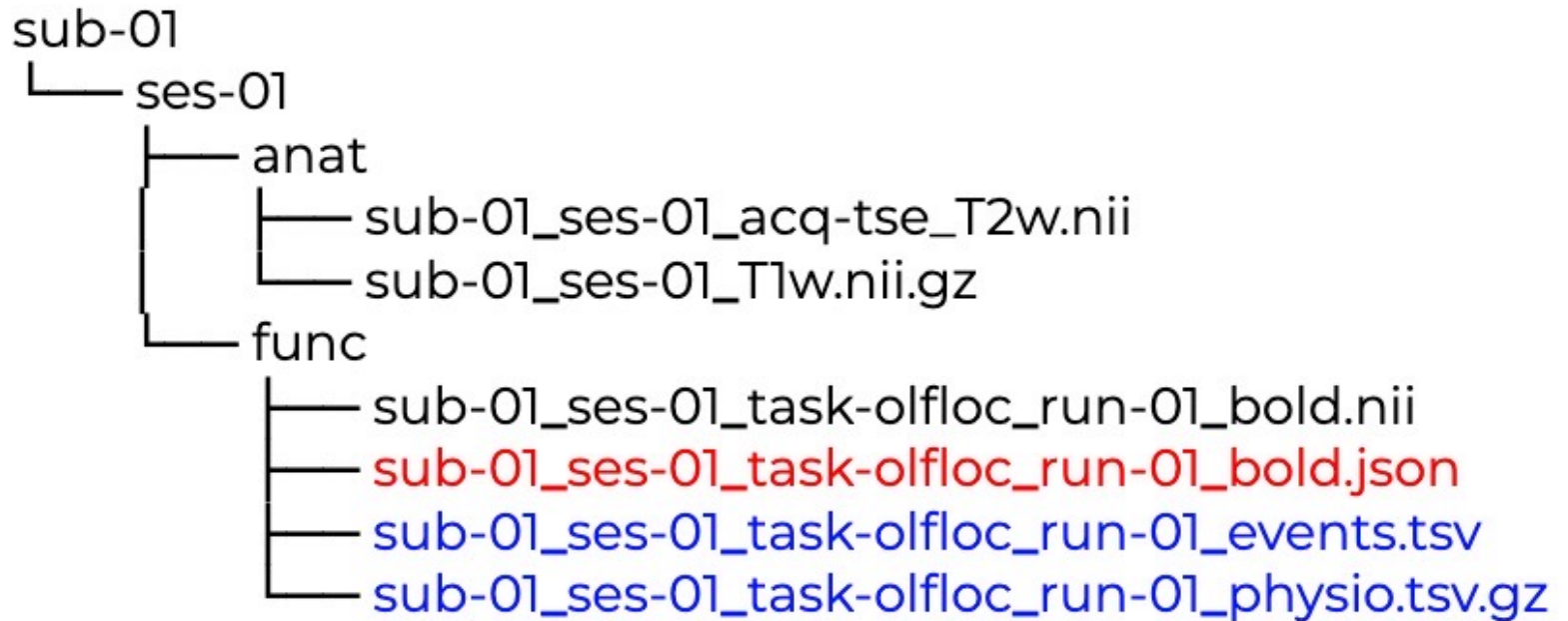


- BIDS defines a file name nomenclature :



BIDS filenames:

- **suffix** preceded by an **underscore**
- **entity-label** pairs separated by **underscores**
- **Entities, labels, suffixes** can only contain letters and / or numbers.
- For a given **suffix**, some **entities** are required and some others are [optional].
- **Entity-label** pairs have a specific order in which they must appear in filename.



- **JSON files:** JavaScript Object Notation
 - for attribute-values pairs
- **TSV files :** Tabulation Separated Values
 - for spreadsheet data

The BIDS ecosystem

BIDS is accompanied by a large ecosystem of tools and resources:

The BIDS ecosystem

BIDS is accompanied by a large ecosystem of tools and resources:

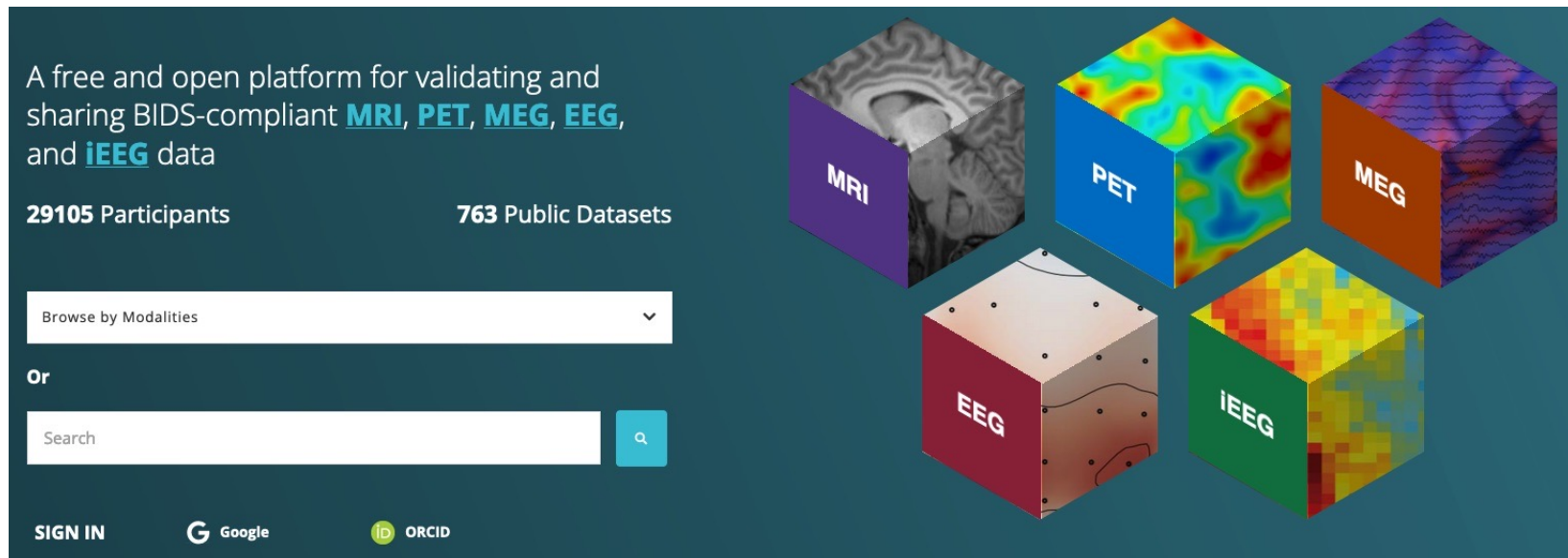
BIDS validator to automatically check if a dataset follows the specifications.

The BIDS ecosystem

BIDS is accompanied by a large ecosystem of tools and resources:

BIDS validator to automatically check if a dataset follows the specifications.

OpenNeuro: an international database that can host BIDS formatted datasets.





A free and open platform for validating and sharing BIDS-compliant [MRI](#), [PET](#), [MEG](#), [EEG](#), and [iEEG](#) data

29105 Participants **763** Public Datasets

Browse by Modalities

Or

Search

SIGN IN  Google  ORCID

The image also features five 3D cubes representing different data modalities: MRI (purple), PET (blue), MEG (orange), EEG (red), and iEEG (green). Each cube is decorated with a representative image of its respective modality.

The BIDS ecosystem

BIDS is accompanied by a large ecosystem of tools and resources:

BIDS validator to automatically check if a dataset follows the specifications.

OpenNeuro: an international database that can host BIDS formatted datasets.

Converters to convert all sorts of data to the BIDS structure.

The BIDS ecosystem

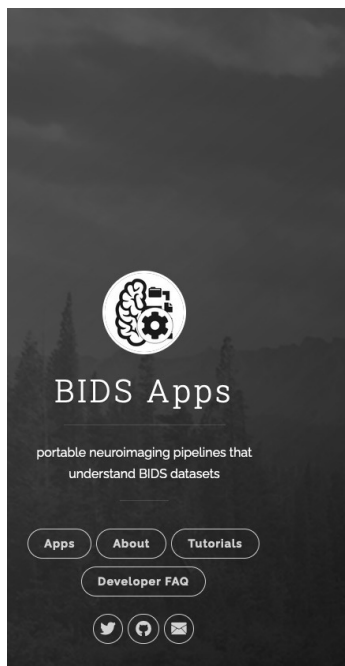
BIDS is accompanied by a large ecosystem of tools and resources:

BIDS validator to automatically check if a dataset follows the specifications.

OpenNeuro: an international database that can host BIDS formatted datasets.

Converters to convert all sorts of data to the BIDS structure.

BIDS apps



Available BIDS Apps

bids-apps/aa	version v0.2.0	open iss
bids-apps/afni_proc	version v0.0.2	open iss
bids-apps/antsCorticalThickness	version v2.2.0-1	open iss
bids-apps/baracus	version v1.1.4	open iss
bids-apps/brainiak-srm	version initial	open iss
bids-apps/BROCCOLI	version v1.0.1	open iss
bids-apps/CPAC	version v1.0.1a_22	open iss
bids-apps/DPARSF	version v4.3.12	open iss
bids-apps/example	version v0.0.7	open iss
bids-apps/FibreDensityAndCrosssection	version v0.0.1	open iss
bids-apps/freesurfer	version v6.0.1-6.1	open iss
bids-apps/HCPpipelines	version v4.3.0-3	open iss
bids-apps/hyperalignment	version v0.0.5	open iss
bids-apps/MAGeTbrain	version v0.3.1	open iss
bids-apps/mindboggle	version v0.0.4-1	open iss
bids-apps/MRtrix3_connectome	version v0.5.3	open iss
bids-apps/ndmg	version v0.1.0	open iss
bids-apps/niak	version v1.0	open iss

Full BIDSification at INT

Some data types are easily transformed into BIDs and specifications have already been defined, for instance MRI, MEG, EEG, ...

But: some data have no defined BIDS structure. We must define it ourselves

INT has started a project for handling all data in a BIDS format.

We started to define new BIDS format for :

- Animal electrophysiology: mono or multi-electrodes recordings in various experimental conditions and setups.
- Eye-tracking

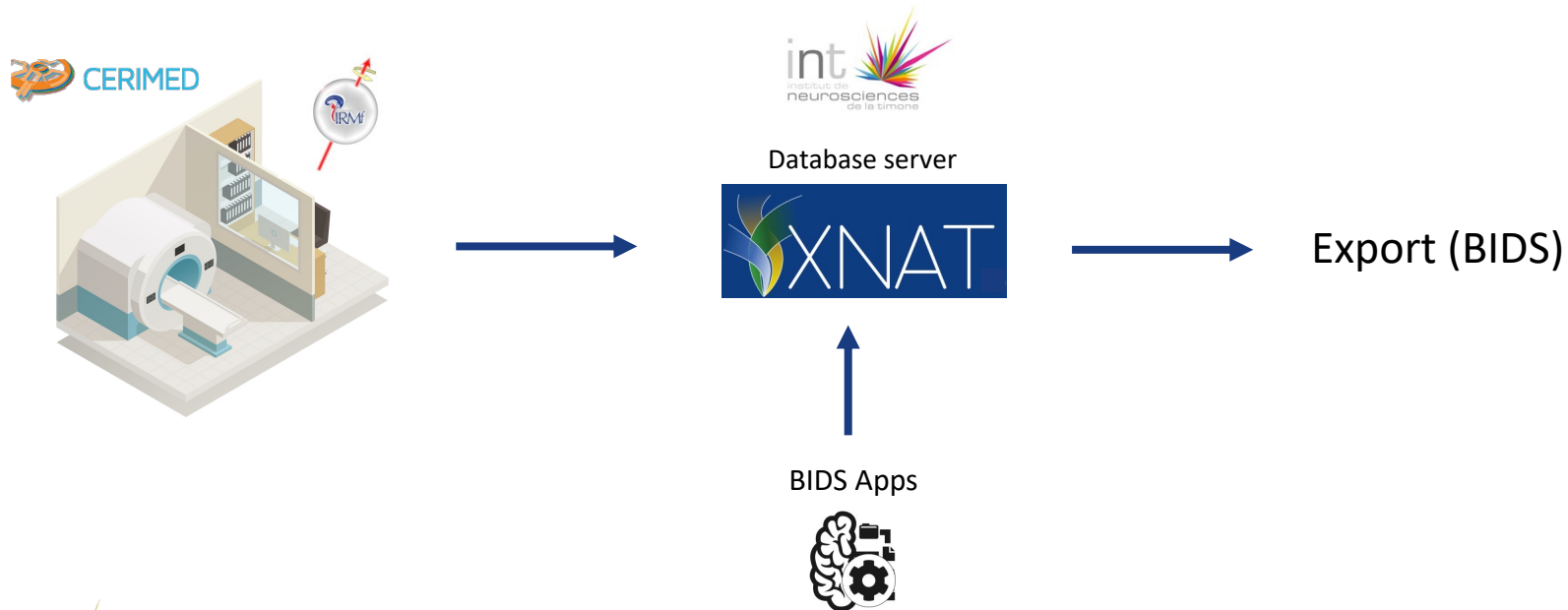
These two proposals have been accepted as official BIDS extension proposals (BEP020 and BEP032)

An easy case: Magnetic Resonance Imaging

The neuroimaging community was ahead in terms of open-science, reproducibility, and proper data management.

BIDS came from that community and MRI was the first data type to be handled.

Specific databasing tools are also available (Xnat)



2 challenges for animal electrophysiology:

- data structuration
- Metadata standardization

Specificities:

- A wide variety of experimental setups (hardware and experimental conditions), so we cannot stream directly from acquisition to storage.
- Metadata are not acquired automatically.
- Diversity of meta data: what metadata to track and how to standardize them into the BIDS structure ?

We developed

- A specific BIDS format (BIDS extension proposal 032)
- A standardized metadata collection process.

Animal ePhys – metadata collection

Requirements:

- User friendly interface
- Minimal overhead (resources, learning time)
- Usage of common terminologies
- Standardization across projects

Animal ePhys – metadata collection

Requirements:

- User friendly interface
- Minimal overhead (resources, learning time)
- Usage of common terminologies
- Standardization across projects

Use of existing resources at INT:



- Browser-based tool for (meta)data capture via surveys
- Python API: PyCap (<https://pycap.readthedocs.io>)
- Registration / Export of collected data
- Text based (csv, json) representation of surveys and collected data

Animal ePhys – metadata collection

Requirements:

- User friendly interface
- Minimal overhead (resources, learning time)
- Usage of common terminologies
- Standardization across projects

Use of existing resources at INT:



- Browser-based tool for (meta)data capture via surveys
- Python API: PyCap (<https://pycap.readthedocs.io>)
- Registration / Export of collected data
- Text based (csv, json) representation of surveys and collected data

Veillez remplir le questionnaire ci-dessous.

Merci !

General	
image	
Ethical Protocol Identifier <small>* must provide value</small>	APAFIS_13894_2018030217116218_v4 ▾
User <small>* must provide value</small>	<input type="text"/>
Session date <small>* must provide value</small>	30-11-2022 Today D-M-Y
Experiment Name <small>* must provide value</small>	<input type="text"/>
Subject GUID <small>* must provide value</small>	<input type="text"/>

Incomplete session	<input type="radio"/> yes	reset
Animal Behaviour <small>* must provide value</small>	<input type="checkbox"/> Very motivated <input type="checkbox"/> Working <input type="checkbox"/> Thirsty <input type="checkbox"/> Sleepy <input type="checkbox"/> Unmotivated <input type="checkbox"/> Agitated	
Data recorded after last trial?	<input type="radio"/> yes	reset
Fluid (reward)	<input type="text"/>	in ml
Fluid (additional)	<input type="text"/>	in ml
Other reward (additional)	<input type="checkbox"/> Fruit (fresh) <input type="checkbox"/> Fruit (dry) <input type="checkbox"/> Seeds <input type="checkbox"/> Treats <input type="checkbox"/> Insects	

Animal ePhys – metadata collection

Requirements:

- User friendly interface
- Minimal overhead (resources, learning time)
- Usage of common terminologies
- Standardization across projects

Use of existing resources at INT:



- Browser-based tool for (meta)data capture via surveys
- Python API: PyCap (<https://pycap.readthedocs.io>)
- Registration / Export of collected data
- Text based (csv, json) representation of surveys and collected data

Survey definition using DigLabTools

<https://github.com/INT-NIT/DigLabTools>



Veuillez remplir le questionnaire ci-dessous.

Merci !

General	
image	
Ethical Protocol Identifier <small>* must provide value</small>	APAFIS_13894_2018030217116218_v4
User <small>* must provide value</small>	<input type="text"/>
Session date <small>* must provide value</small>	30-11-2022 Today D-M-Y
Experiment Name <small>* must provide value</small>	<input type="text"/>
Subject GUID <small>* must provide value</small>	<input type="text"/>

Incomplete session	<input type="radio"/> yes	reset
Animal Behaviour <small>* must provide value</small>	<input type="checkbox"/> Very motivated <input type="checkbox"/> Working <input type="checkbox"/> Thirsty <input type="checkbox"/> Sleepy <input type="checkbox"/> Unmotivated <input type="checkbox"/> Agitated	
Data recorded after last trial?	<input type="radio"/> yes	reset
Fluid (reward)	<input type="text"/>	in ml
Fluid (additional)	<input type="text"/>	in ml
Other reward (additional)	<input type="checkbox"/> Fruit (fresh) <input type="checkbox"/> Fruit (dry) <input type="checkbox"/> Seeds <input type="checkbox"/> Treats <input type="checkbox"/> Insects	



BIDS Extension Proposal 032

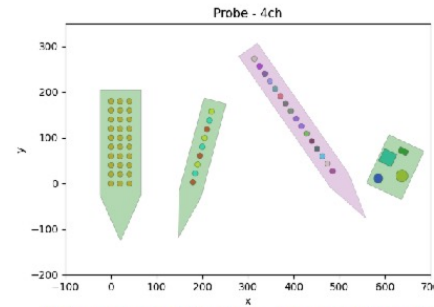
<http://bids.neuroimaging.io/bep032>

Content:

- Open file format (nix/nwb)
- Probe and wiring description
- Metadata

an example...

```
my_dataset/
├─ dataset_description.json
├─ participants.json
├─ participants.tsv
├─ tasks.json
├─ tasks.tsv
├─ sub-i/
│  └─ sub-i_sessions.json
│     └─ sub-i_sessions.tsv
│        └─ ses-140703/
│           └─ ephys/
│              ├── sub-i_ses-140703_task-r2g_run-001_channels.tsv
│              ├── sub-i_ses-140703_task-r2g_run-001_contacts.tsv
│              ├── sub-i_ses-140703_task-r2g_run-001_ephys.json
│              ├── sub-i_ses-140703_task-r2g_run-001_ephys.nix
│              └─ sub-i_ses-140703_task-r2g_run-001_probes.tsv
└─ sub-l/
   ├── sub-l_sessions.json
   ├── sub-l_sessions.tsv
   └─ ses-101210/
      └─ ephys/
         ├── sub-l_ses-101210_task-r2g_run-001_channels.tsv
         ├── sub-l_ses-101210_task-r2g_run-001_contacts.tsv
         ├── sub-l_ses-101210_task-r2g_run-001_ephys.json
         ├── sub-l_ses-101210_task-r2g_run-001_ephys.nix
         └─ sub-l_ses-101210_task-r2g_run-001_probes.tsv
```



<https://probeinterface.readthedocs.io>

Naming of files and directories :

- follows the generic rules of BIDS
- intuitive hierarchy (*project/animal/session/modality*)
- redundancy of information in file and directory names
- added specific infos for electrophysiology

Supported data file format (INCF standards) :

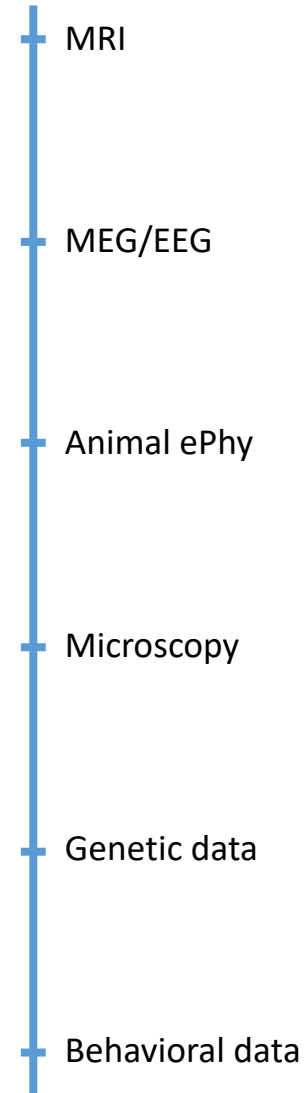
- NIX
- NWB



Supported metadata file formats (as in generic BIDS) :

- tsv
- json

A larger picture



A larger picture



- MRI
- MEG/EEG
- Animal ePhy
- Microscopy
- Genetic data
- Behavioral data

A larger picture



2017: one of the first french labs to deliver full BIDS data

MRI

MEG/EEG

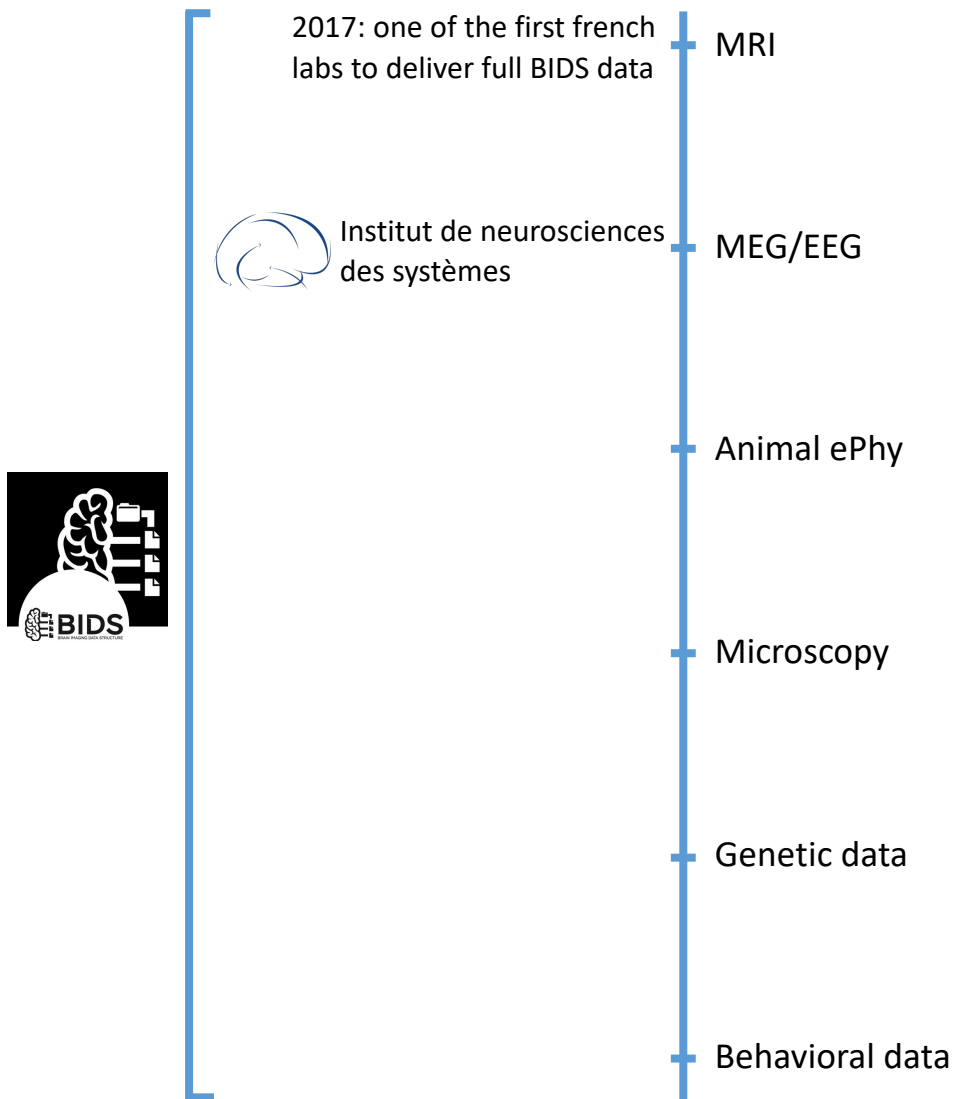
Animal ePhy

Microscopy

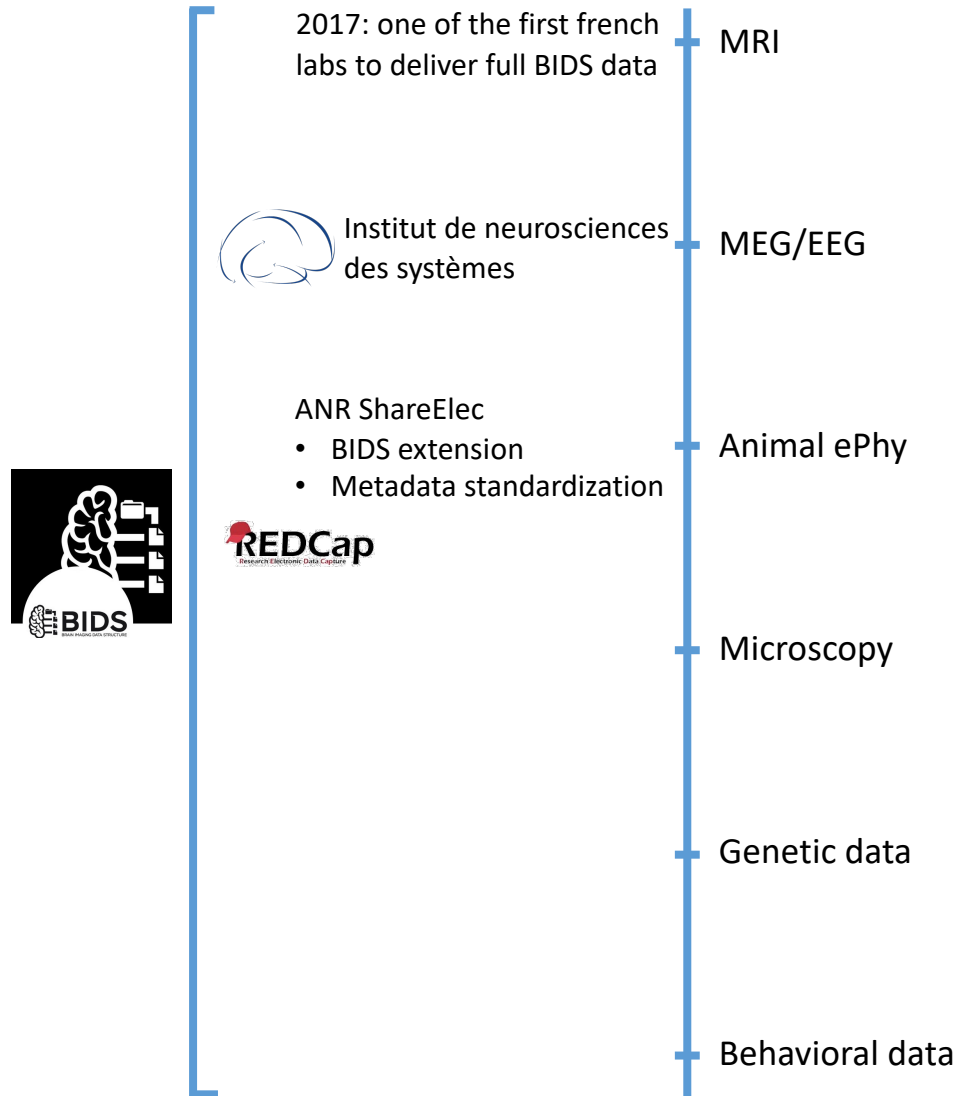
Genetic data

Behavioral data

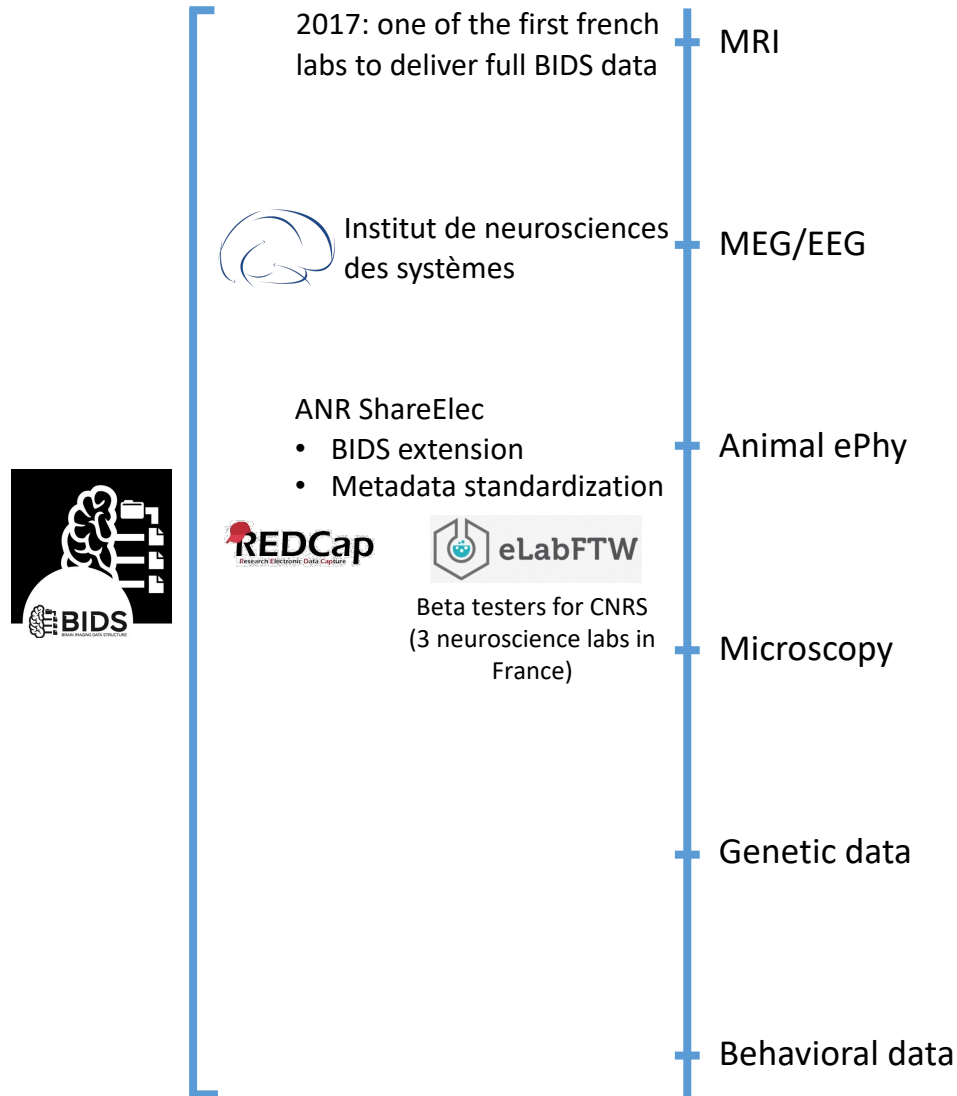
A larger picture



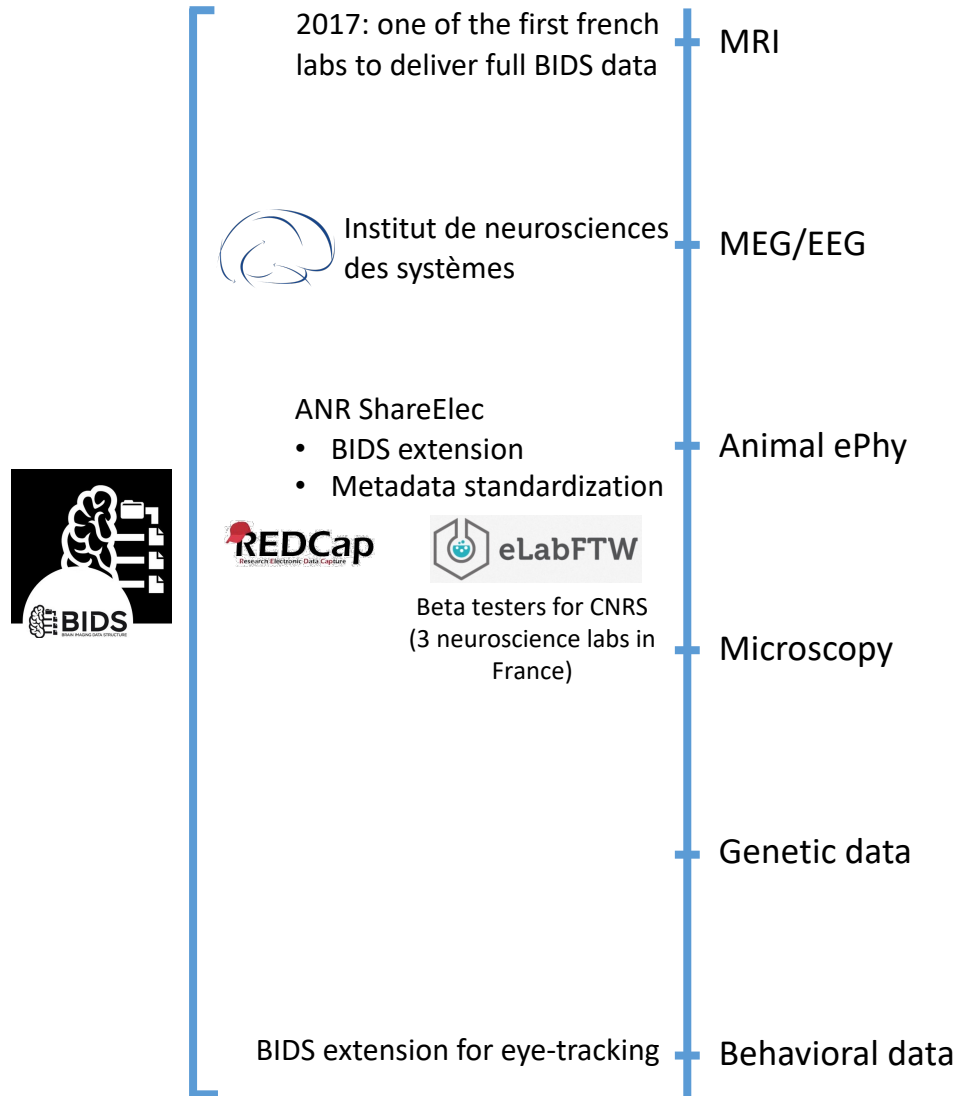
A larger picture



A larger picture

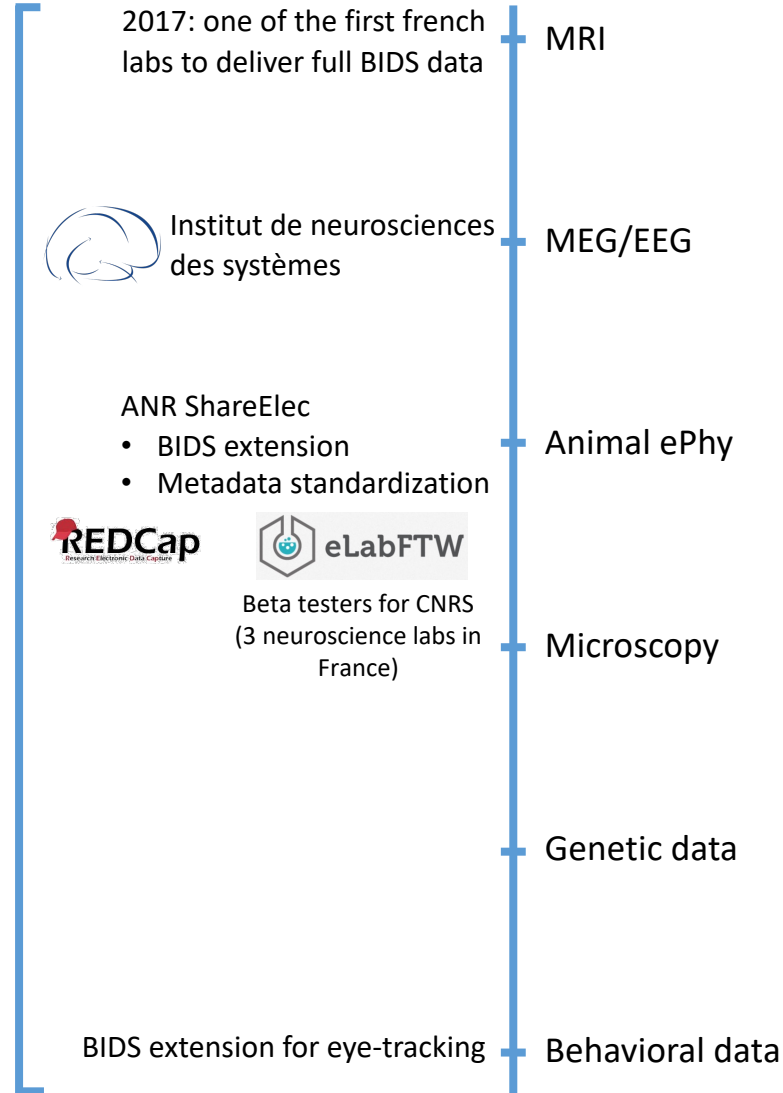


A larger picture



A larger picture

IRM
Xnat +BIDS apps
automatic processing

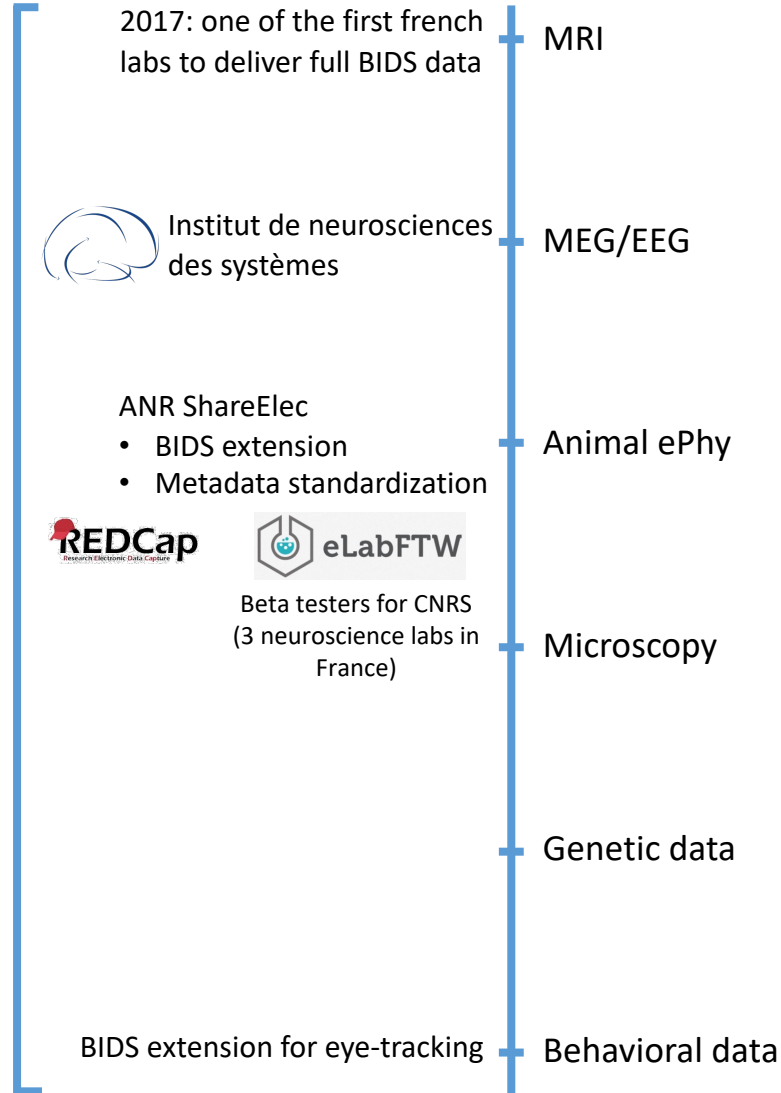


A larger picture

IRM
Xnat +BIDS apps
automatic processing



Animal ePhy
BIDS + Spike Interface



A larger picture

IRM
Xnat +BIDS apps
automatic processing



Animal ePhy
BIDS + Spike Interface



2017: one of the first french labs to deliver full BIDS data



Institut de neurosciences des systèmes

ANR ShareElec

- BIDS extension
- Metadata standardization



Beta testers for CNRS
(3 neuroscience labs in France)

MRI

MEG/EEG

Animal ePhy

Microscopy

Genetic data

BIDS extension for eye-tracking

Behavioral data



CEDRE
INT/Inmed
/IBDM

A larger picture

IRM
Xnat +BIDS apps
automatic processing



Animal ePhy
BIDS + Spike Interface



2017: one of the first french labs to deliver full BIDS data



Institut de neurosciences des systèmes

ANR ShareElec

- BIDS extension
- Metadata standardization



Beta testers for CNRS
(3 neuroscience labs in France)

MRI



A unique Xnat server (INT/CRMBM) at datacenter ?

MEG/EEG

Animal ePhy

Microscopy



CEDRE
INT/Inmed
/IBDM

Genetic data

BIDS extension for eye-tracking

Behavioral data

Thank you !

INT

Institute of Neuroscience and Medicine
Jülic, germany



ANR ShareElec

Sylvain Takerkart

Julia Sprenger

David Meunier

Dipankar bachar

Frédéric Barthélémy

Sonja Grün

Michael Denker

Junji Ito