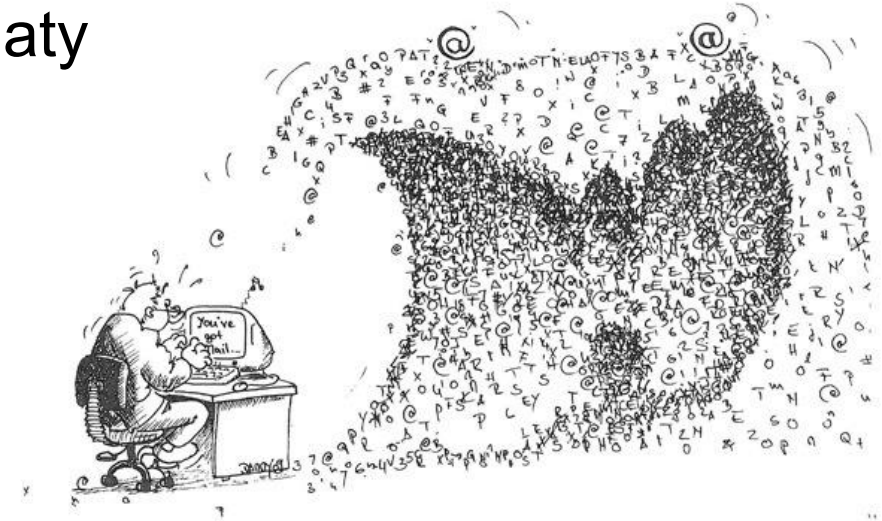
FAIR data / Open data pro uživatele CF MAFIL

Praktické představení nástrojů a postupů pro práci s daty

Tomáš Slavíček



Osnova

- Data workflow v CF MAFIL (MR)
- Datové formáty
- BIDS
- Příprava datasetu/‘curation’
- Anonymizace
- Nástroje pro práci s daty / ukázka na vzorovém datasetu

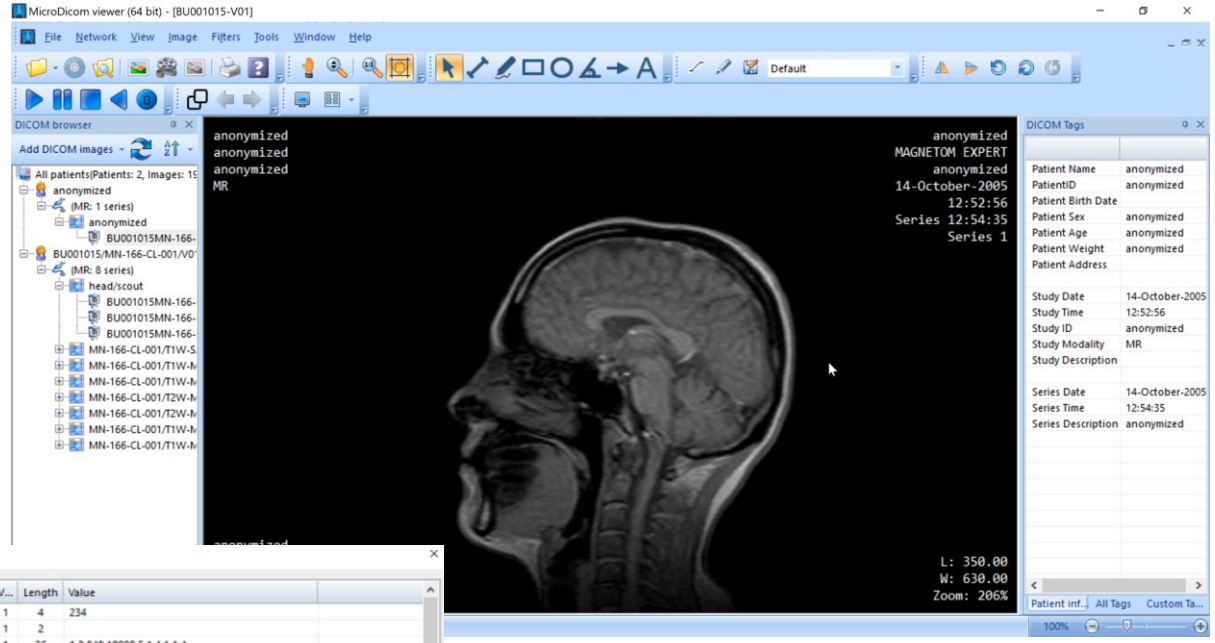
Data workflow (MR)

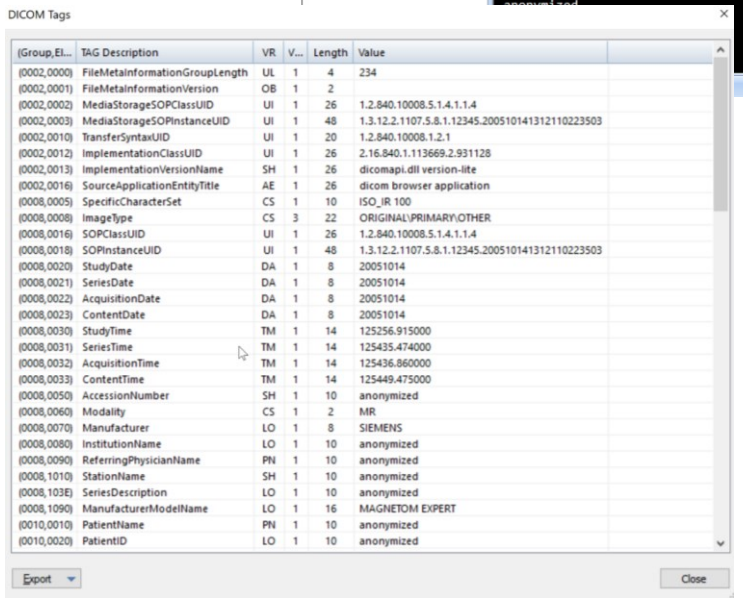
MUNI | CEITEC

DICOM

- Standard pro obrazová data v medicíně
- Z MR konzole (Syngo) do PACS
- Metadata (tzv. TAGy)
  - Parametry akvizice (TR, TE, FOV, aj.)
  - Technické parametry (použité cívky, aj.)
  - Údaje o subjektu (citlivá)

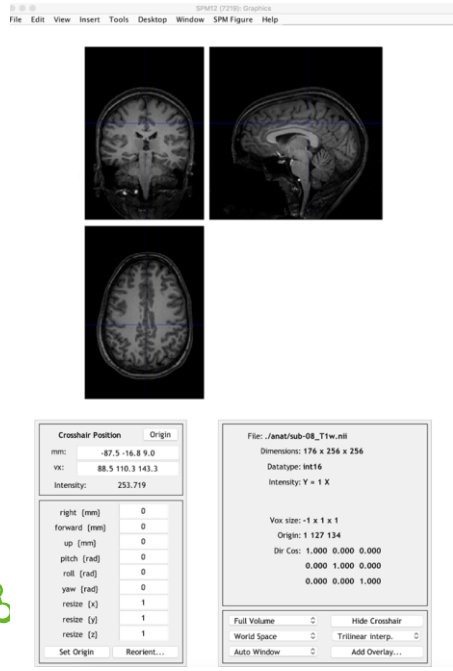
- Nevhodné pro neurovědní aplikace
  - Objem dat – pomalé
  - Ukládání (řezy, mozaika)
  - Redundantní metadata





NIfTI-1 (*.nii or *.nii.gz)

- Neuroimaging Informatics Technology Initiative (2003), https://nifti.nimh.nih.gov/
  - Nástupce ANALYZE 7.5 formátu (*.img + *.hdr)
- Pro většinu aplikací ≈ sekvencí (anat, func, diff, atd.), pro MRS – v přípravě
- Široká podpora nástrojů – SPM, FSL, FreeSurfer, AFNI, BrainVoyager, etc.
- Hlavička (348 bytes)
  - Intent type (statistika nebo obrazová data)
  - Prostorové transformace (VX to „real world“)
  - Datový typ (bits per voxel) – typicky UINT16
  - Velikosti voxelů, škálování (slope+intercept), aj.
- Datová matice
  - 3D pro anatomické snímky
  - 4D pro funkční/difuzní/jiné



MUNI

Konverze DICOM do NIfTI-1

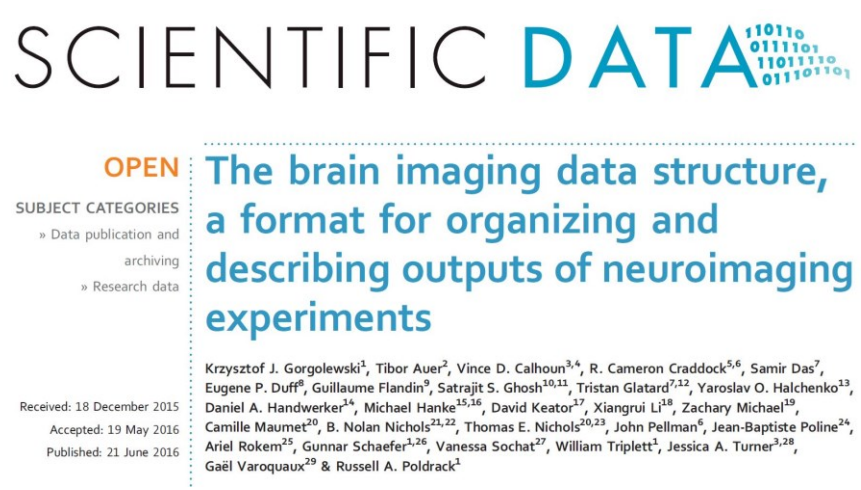
- Redukce metadat ➔ ’sidecar’ soubor (malá velikost hlavičky NIfTI)
- Parametry difuzních sekvencí ➔ *.bvec + *.bval soubory
- Nestandardní DICOM implementace – Siemens CSA header
  - Pokročilé parametry akvizice (typ „prescan normalize“, MB acceleration factor, atd.)
- Příklady kovertorů
  - https://github.com/rordenlab/dcm2nix (Linux and Windows)
  - https://github.com/xiangruili/dicm2nii (MATLAB)

MUNI | CEITEC

Organizace dat

MUNI | CEITEC

BIDS standard


SCIENTIFIC DATA

OPEN

SUBJECT CATEGORIES
» Data publication and archiving
» Research data

The brain imaging data structure, a format for organizing and describing outputs of neuroimaging experiments

Krzysztof J. Gorgolewski[1], Tibor Auer[2], Vince D. Calhoun[3,4], R. Cameron Craddock[5,6], Samir Das[7], Eugene P. Duff[8], Guillaume Flandin[9], Satrajit S. Ghosh[10,11], Tristan Glatard[7,12], Yaroslav O. Halchenko[13], Daniel A. Handwerker[14], Michael Hanke[15,16], David Keator[17], Xiangrui Li[18], Zachary Michael[19], Camille Maumet[20], B. Nolan Nichols[21,22], Thomas E. Nichols[20,23], John Pellman[6], Jean-Baptiste Poline[24], Ariel Rokem[25], Gunnar Schaefer[1,26], Vanessa Sochat[27], William Triplett[1], Jessica A. Turner[3,28], Gaël Varoquaux[29] & Russell A. Poldrack[1]

Received: 18 December 2015
Accepted: 19 May 2016
Published: 21 June 2016

- Brain Imaging Data Structure
  - https://bids.neuroimaging.io
  - K.Gorgolewski, 2015
  - Pokrývá většinu neurozobrazovacích modalit (MRI, PET, EEG, MEG, …)
  - Vychází z „best practice“, která se a posteriori stala „standardem“

- Je to **datová skladba**; nikoliv nový formát dat (NIfTI, TSV, JSON)

- Je o tom jak:
  - Utřídíme data do složek
  - Pojmenujeme soubory
  - Dokumentujeme metadata

MUNI | CEITEC

BIDS standard

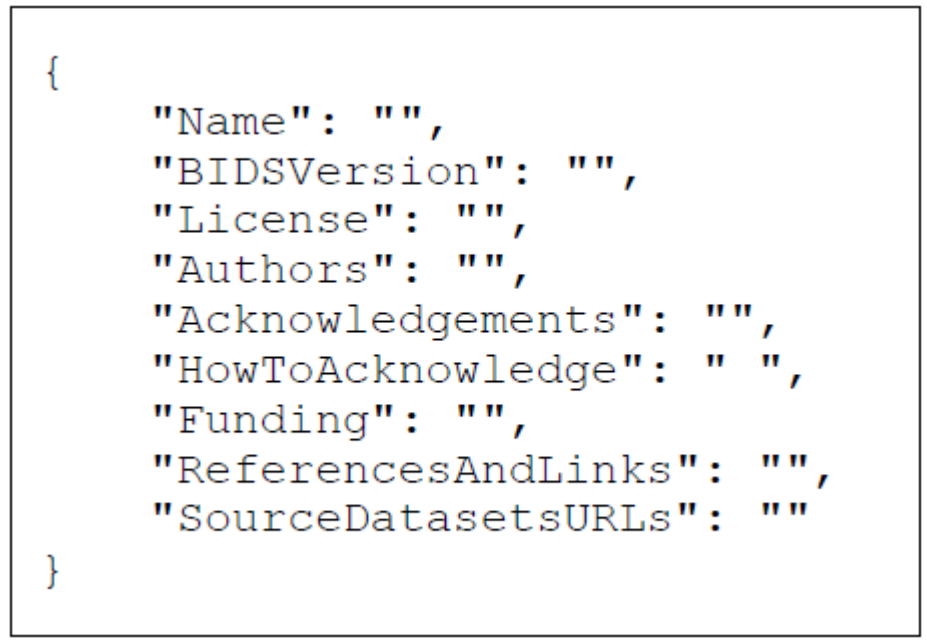
- Snadno čitelný pro člověka a srozumitelný pro počítač
  - Usnadňuje sdílení + archivaci (noví členové týmu, spolupráce, repozitáře)
  - Snižuje výskyt chyb (polo/automatizace zpracování dat)
  - Snadné zpracování – podpora pipeline (MRIQC, fMRIPrep, HCP Pipelines, aj.)
- Založený na **klíčových slovech** a **slovnících**
  - Dataset, subject, session, data type, task, event, run, <index>, <label>, sufix, etc.
  - Povinná, doporučená a doplňková data a metadata
- Princip dědičnosti (metadata)
  - Shora-dolů
  - Maximálně jeden na dané úrovni
  - Výjimky

```
dset/
├── dataset_description.json
├── participants.tsv
└── sub-01
    ├── anat
    │   ├── sub-01_T1w.json
    │   └── sub-01_T1w.nii.gz
    ├── dwi
    │   ├── sub-01_dwi.bval
    │   ├── sub-01_dwi.bvec
    │   ├── sub-01_dwi.json
    │   └── sub-01_dwi.nii.gz
    ├── fmap
    │   ├── sub-01_acq-dwi_dir-AP_epi.json
    │   ├── sub-01_acq-dwi_dir-AP_epi.nii.gz
    │   ├── sub-01_acq-dwi_dir-PA_epi.json
    │   ├── sub-01_acq-dwi_dir-PA_epi.nii.gz
    │   ├── sub-01_acq-func_dir-AP_epi.json
    │   └── sub-01_acq-func_dir-AP_epi.nii.gz
```


CEITEC

Metadata jako textové soubory (*.tsv, *.json)

```
dset/
├── dataset_description.json
├── participants.tsv
└── sub-01
    ├── anat
    │   ├── sub-01_T1w.json
    │   └── sub-01_T1w.nii.gz
    ├── dwi
    │   ├── sub-01_dwi.bval
    │   ├── sub-01_dwi.bvec
    │   ├── sub-01_dwi.json
    │   └── sub-01_dwi.nii.gz
    ├── fmap
    │   ├── sub-01_acq-dwi_dir-AP_epi.json
    │   ├── sub-01_acq-dwi_dir-AP_epi.nii.gz
    │   ├── sub-01_acq-dwi_dir-PA_epi.json
    │   ├── sub-01_acq-dwi_dir-PA_epi.nii.gz
    │   ├── sub-01_acq-func_dir-AP_epi.json
    │   ├── sub-01_acq-func_dir-AP_epi.nii.gz
    │   ├── sub-01_acq-func_dir-PA_epi.json
    │   └── sub-01_acq-func_dir-PA_epi.nii.gz
    └── func
        ├── sub-01_task-nback_run-01_bold.json
        ├── sub-01_task-nback_run-01_bold.nii.gz
        ├── sub-01_task-nback_run-01_events.tsv
        ├── sub-01_task-nback_run-01_sbref.json
        └── sub-01_task-nback_run-01_sbref.nii.gz
```

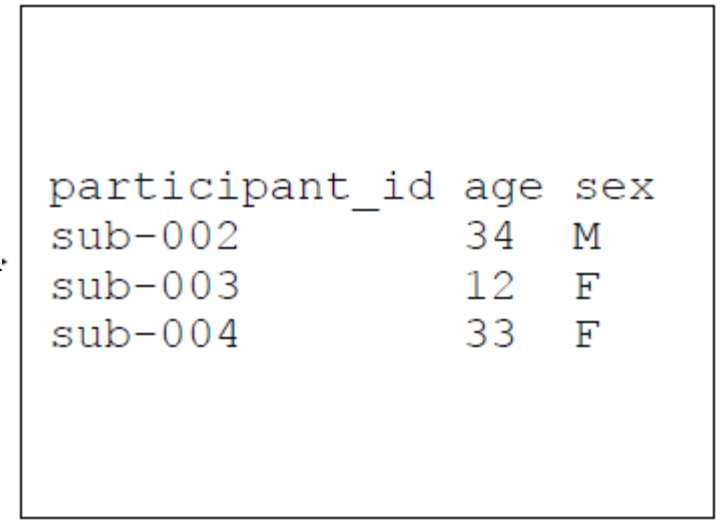


```
{
    "Name": "",
    "BIDSVersion": "",
    "License": "",
    "Authors": "",
    "Acknowledgements": "",
    "HowToAcknowledge": " ",
    "Funding": "",
    "ReferencesAndLinks": "",
    "SourceDatasetsURLs": ""
}
```


MUNI | CEITEC

Metadata jako textové soubory (*.tsv, *.json)

```
dset/
├── dataset_description.json
├── participants.tsv
└── sub-01
    ├── anat
    │   ├── sub-01_T1w.json
    │   └── sub-01_T1w.nii.gz
    ├── dwi
    │   ├── sub-01_dwi.bval
    │   ├── sub-01_dwi.bvec
    │   ├── sub-01_dwi.json
    │   └── sub-01_dwi.nii.gz
    ├── fmap
    │   ├── sub-01_acq-dwi_dir-AP_epi.json
    │   ├── sub-01_acq-dwi_dir-AP_epi.nii.gz
    │   ├── sub-01_acq-dwi_dir-PA_epi.json
    │   ├── sub-01_acq-dwi_dir-PA_epi.nii.gz
    │   ├── sub-01_acq-func_dir-AP_epi.json
    │   ├── sub-01_acq-func_dir-AP_epi.nii.gz
    │   ├── sub-01_acq-func_dir-PA_epi.json
    │   └── sub-01_acq-func_dir-PA_epi.nii.gz
    └── func
        ├── sub-01_task-nback_run-01_bold.json
        ├── sub-01_task-nback_run-01_bold.nii.gz
        ├── sub-01_task-nback_run-01_events.tsv
        ├── sub-01_task-nback_run-01_sbref.json
        └── sub-01_task-nback_run-01_sbref.nii.gz
```

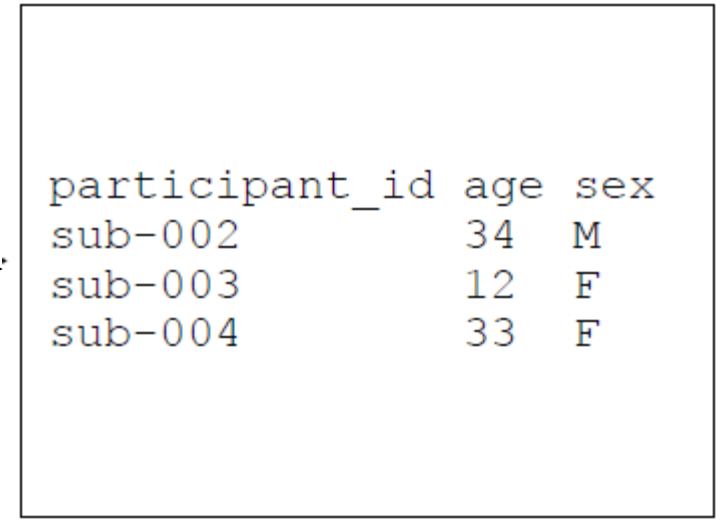


participant_id	age	sex
sub-002	34	M
sub-003	12	F
sub-004	33	F

MUNI | CEITEC

Metadata jako textové soubory (*.tsv, *.json)

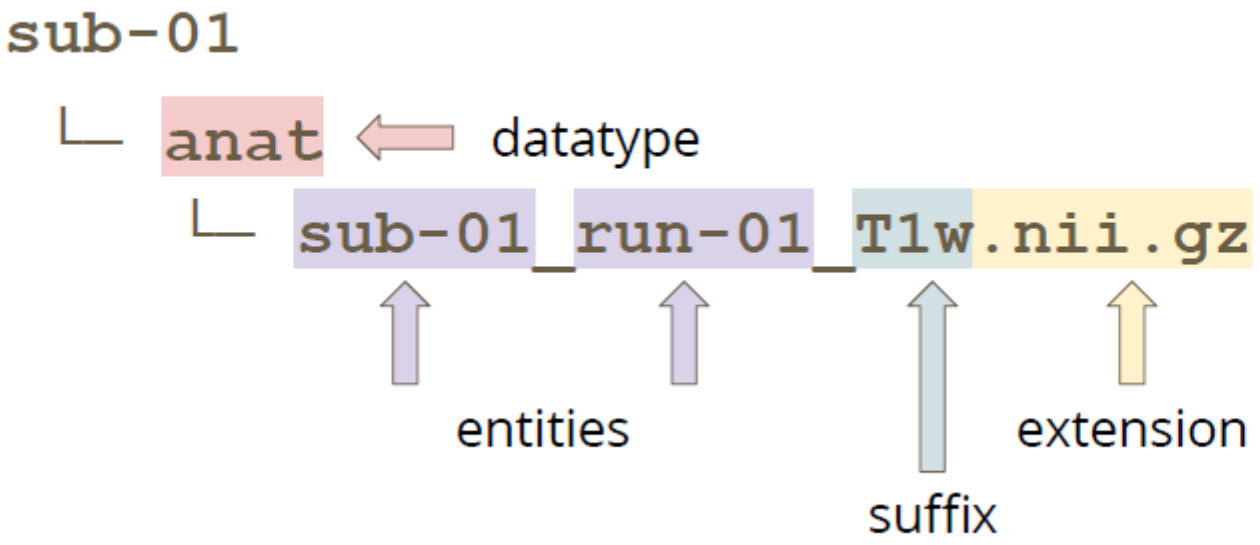
```
dset/
├── dataset_description.json
├── participants.tsv
└── sub-01
    ├── anat
    │   ├── sub-01_T1w.json
    │   └── sub-01_T1w.nii.gz
    ├── dwi
    │   ├── sub-01_dwi.bval
    │   ├── sub-01_dwi.bvec
    │   ├── sub-01_dwi.json
    │   └── sub-01_dwi.nii.gz
    ├── fmap
    │   ├── sub-01_acq-dwi_dir-AP_epi.json
    │   ├── sub-01_acq-dwi_dir-AP_epi.nii.gz
    │   ├── sub-01_acq-dwi_dir-PA_epi.json
    │   ├── sub-01_acq-dwi_dir-PA_epi.nii.gz
    │   ├── sub-01_acq-func_dir-AP_epi.json
    │   ├── sub-01_acq-func_dir-AP_epi.nii.gz
    │   ├── sub-01_acq-func_dir-PA_epi.json
    │   └── sub-01_acq-func_dir-PA_epi.nii.gz
    └── func
        ├── sub-01_task-nback_run-01_bold.json
        ├── sub-01_task-nback_run-01_bold.nii.gz
        ├── sub-01_task-nback_run-01_events.tsv
        ├── sub-01_task-nback_run-01_sbref.json
        └── sub-01_task-nback_run-01_sbref.nii.gz
```



participant_id	age	sex
sub-002	34	M
sub-003	12	F
sub-004	33	F

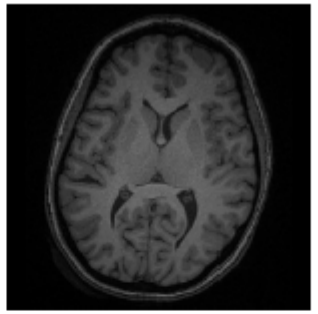
MUNI | CEITEC

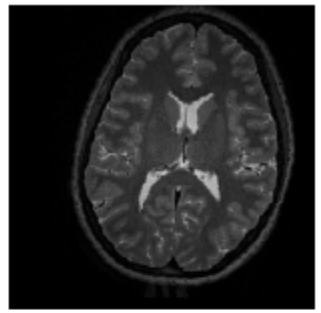
Princip pojmenování souborů



MUNI | CEITEC

Anatomické snímky





```
sub-01
└─ anat
   ├─ sub-01_T1w.nii.gz
   ├─ sub-01_T1w.json
   ├─ sub-01_mod-T1w_defacemask.nii.gz
   ├─ sub-01_T2w.nii.gz
   ├─ sub-01_T2w.json
   └─ sub-01_mod-T2w_defacemask.nii.gz
```


MUNI | CEITEC

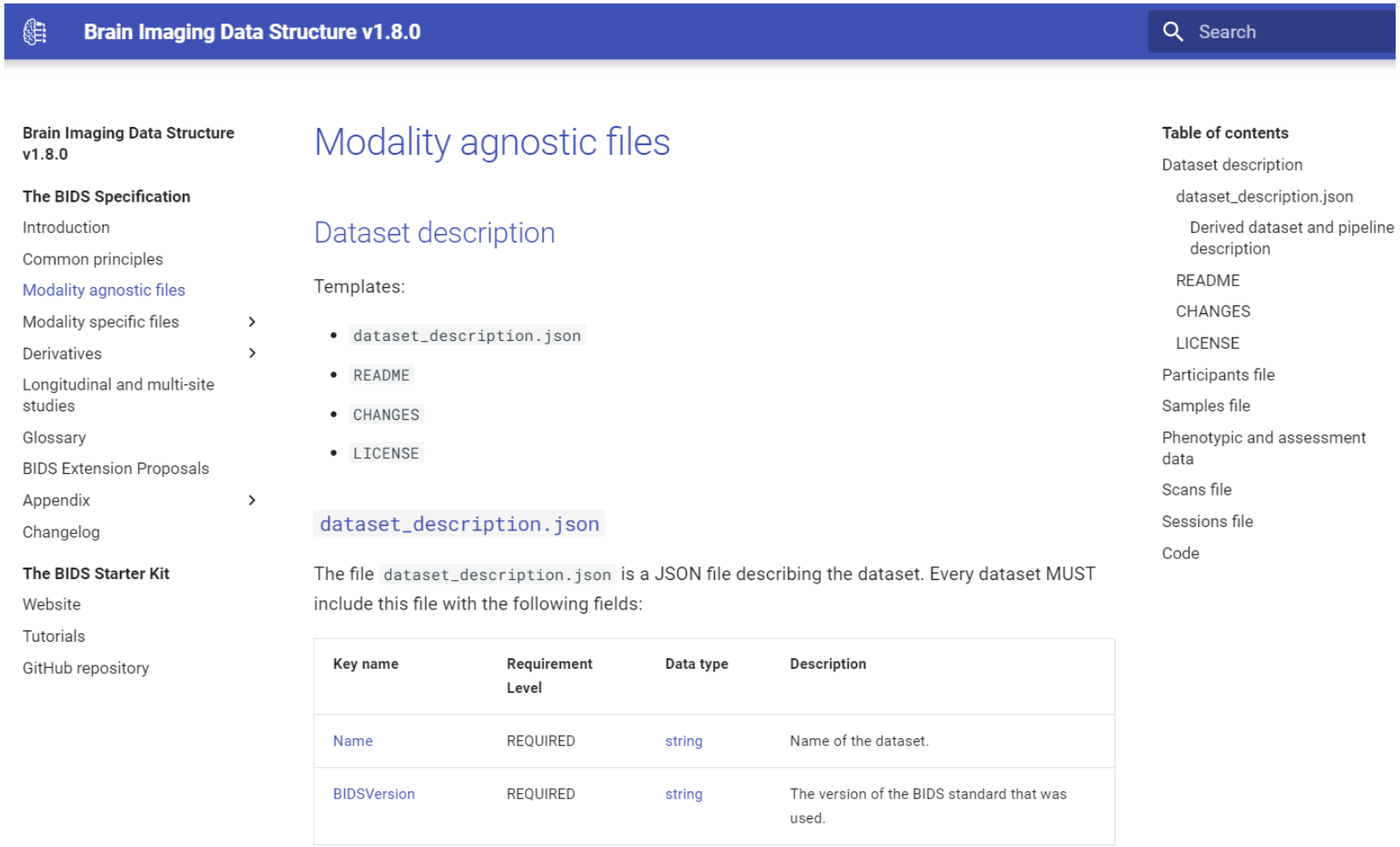
Anatomické snímky

```
sub-01
└─ anat
   ├─ sub-01_T1w.nii.gz
   ├─ sub-01_T1w.json
   ├─ sub-01_mod-T1w_defacemask
   ├─ sub-01_T2w.nii.gz
   ├─ sub-01_T2w.json
   └─ sub-01_mod-T2w_defacemask.nii.gz
```


```
{
    "EchoTime": 0.06052,
    "EffectiveEchoSpacing": 0.000245314,
    "FlipAngle": 77,
    "MagneticFieldStrength": 3,
    "Manufacturer": "Siemens",
    "ManufacturersModelName": "Prisma",
    "MultibandAccelerationFactor": 3,
    "ParallelReductionFactorInPlane": 2,
    "PhaseEncodingDirection": "j-",
    "RepetitionTime": 1.5,
    "SliceThickness": 2.5,
    "SpacingBetweenSlices": 2.5,
    "TaskName": "strangerthings",
    "TotalReadoutTime": 0.0208517
}
```


MUNI | CEITEC

https://bids-specification.readthedocs.io/en/stable/



Brain Imaging Data Structure v1.8.0

Search

Brain Imaging Data Structure v1.8.0

The BIDS Specification

Introduction

Common principles

Modality agnostic files

Modality specific files

Derivatives

Longitudinal and multi-site studies

Glossary

BIDS Extension Proposals

Appendix

Changelog

The BIDS Starter Kit

Website

Tutorials

GitHub repository

Modality agnostic files

Dataset description

Templates:

- `dataset_description.json`
- `README`
- `CHANGES`
- `LICENSE`

`dataset_description.json`

The file `dataset_description.json` is a JSON file describing the dataset. Every dataset MUST include this file with the following fields:

Key name	Requirement Level	Data type	Description
Name	REQUIRED	string	Name of the dataset.
BIDSVersion	REQUIRED	string	The version of the BIDS standard that was used.

Table of contents

Dataset description

dataset_description.json

Derived dataset and pipeline description

README

CHANGES

LICENSE

Participants file

Samples file

Phenotypic and assessment data

Scans file

Sessions file

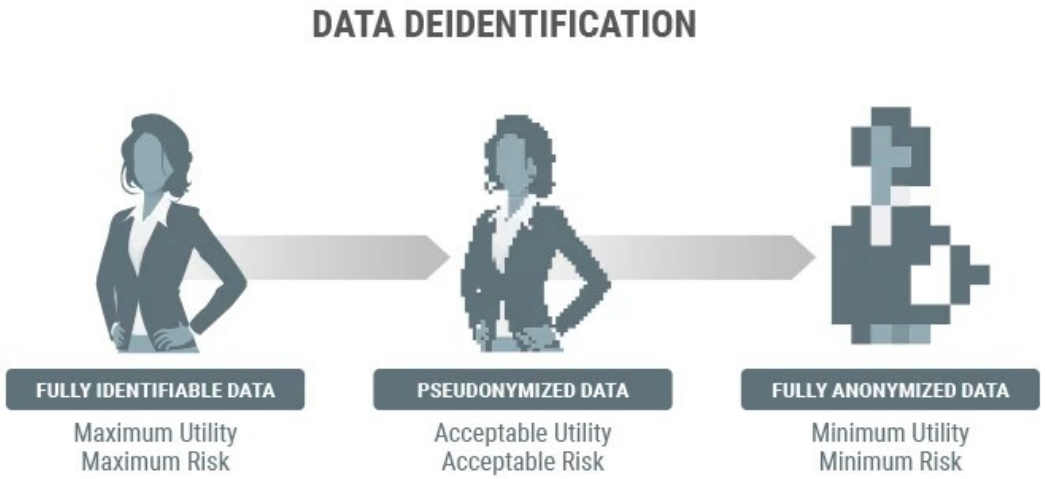
Code

MUNI | CEITEC

Příprava datasetu (‘curation’)

- Předání dat MAFIL ➔ výzkumník
  - Stažení + konverze + **pseudonymizace ➔ preBIDS**
  - Pseudonymizace použitím VisitID (nutná znalost pro „dešifrování“)
    - Párování s jinými modalitami v rámci MAFIL – stimulační logy, EEG, ExG, aj.
  - Upload na MAFIL servery do složky projektu nebo zaslání přes Filesender
- Kompletace + čištění
  - Doplnění dalších vstupů (zdraví vs nemocní, krevní rozbory, psychologické testy, aj.)
  - Konverze formátů (BIDS kompatibilní)
  - Kontrola kvality – pohybové parametry, SNR
  - BIDS validátor (online i offline verze)
- Příprava na publikování v repozitáři
  - Detailněji viz praktická ukázka

Anonymizace

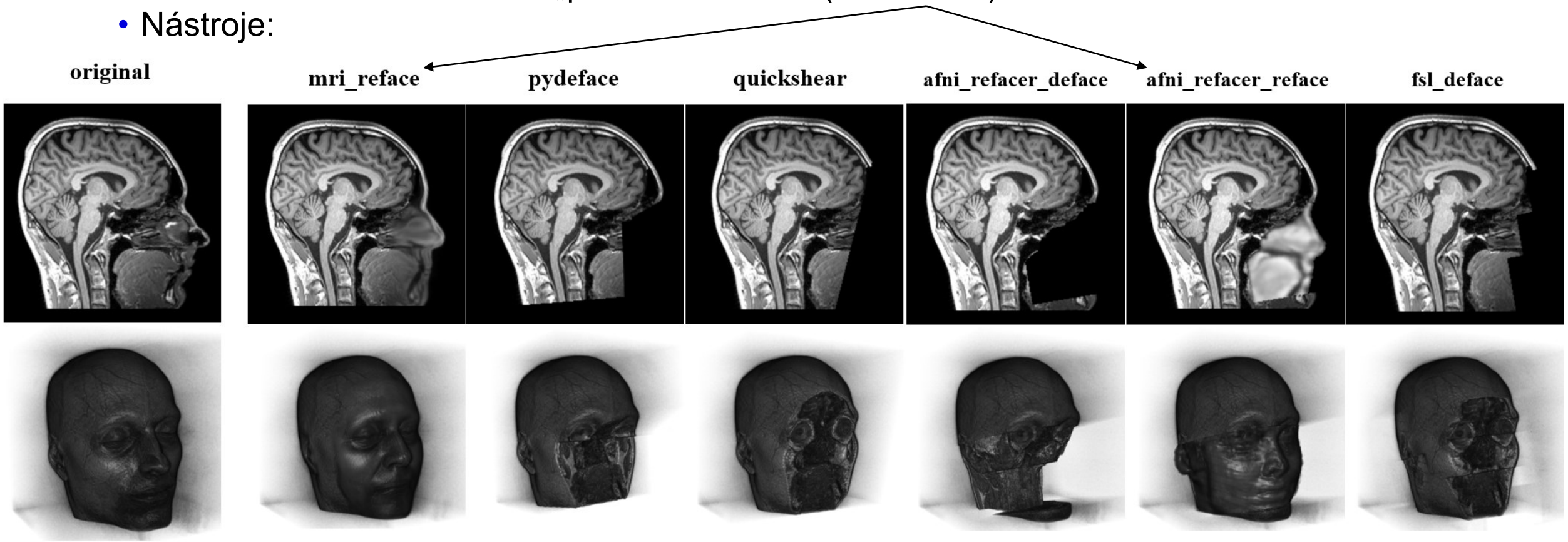


- Cíl – zachování soukromí účastníků
- Nařízení GDPR + informovaný souhlas
  - … zda je osoba identifikovatelná s vynaložením přiměřeného úsilí
- Pseudonymizace
  - Smazání/náhrada osobních údajů kódy,
  - Existuje **převodní tabulka**, lze doplnit data později
- Anonymizace
  - Nevratný jednosměrný proces, převodní tabulka neexistuje, nelze doplnit data
  - Absolutní anonymizace není (!) – rovnováha mezi zachováním soukromí a využitelností dat

MUNI | CEITEC

Deface

- Anatomické snímky s vysokým rozlišením – 1 × 1 × 1mm (T1, T2, FLAIR)
- Poškození dat ➔ nevhodné pro morfometrii, problematická normalizace
- Odstranění nebo nahrazení „průměrnou“ tváří (tzv. reface)
- Nástroje:



Ukázka na vzorovém datasetu (screenshoty)

- Kompletace, čištění
- Kontrola kvality
- Deface
- Anonymizace
- Doplnění informací o datasetu
- Validace
- Přidělení DOI
- Upload do repozitáře
- Ověření záznamu DOI

Kompletace, čištění

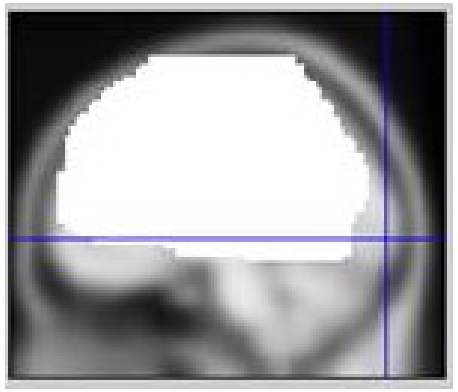
- BOSAPS3 – vliv MB faktoru a velikosti FA na kvalitu dat

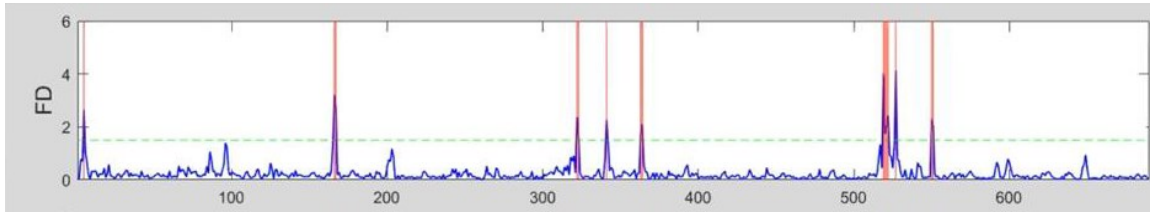
- 50 subjektů, 1 session
- MR data
  - Anat – T1
  - Fmap
  - Func – 7× bold (multiecho + multiband)
- Vizuální stimulace (blokový design)

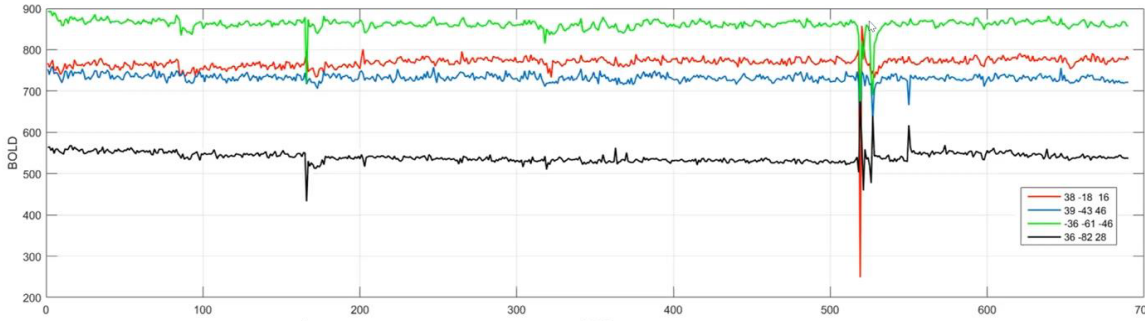
- Čištění
  - opakovaná měření (run-02)
  - sjednocení protokolu (2017 + 2021), nová část navíc part-phase u bold

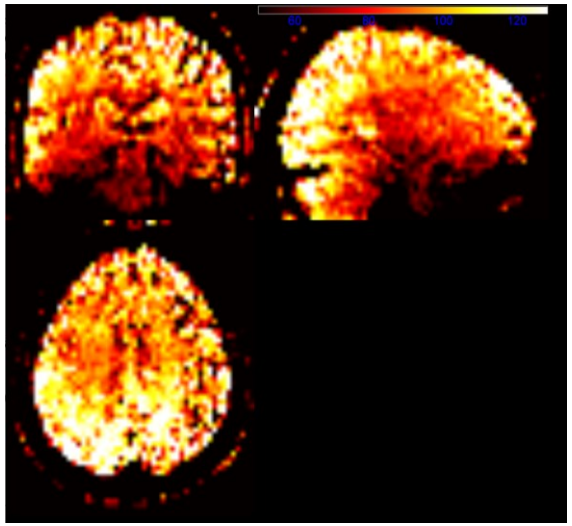
Kontrola kvality

- Pokrytí dat
  - Překryv validních dat napříč celým souborem
- Pohybové artefakty
  - FWD, DVARS, …
- Kvalita dat
  - SNR, tSNR, ...







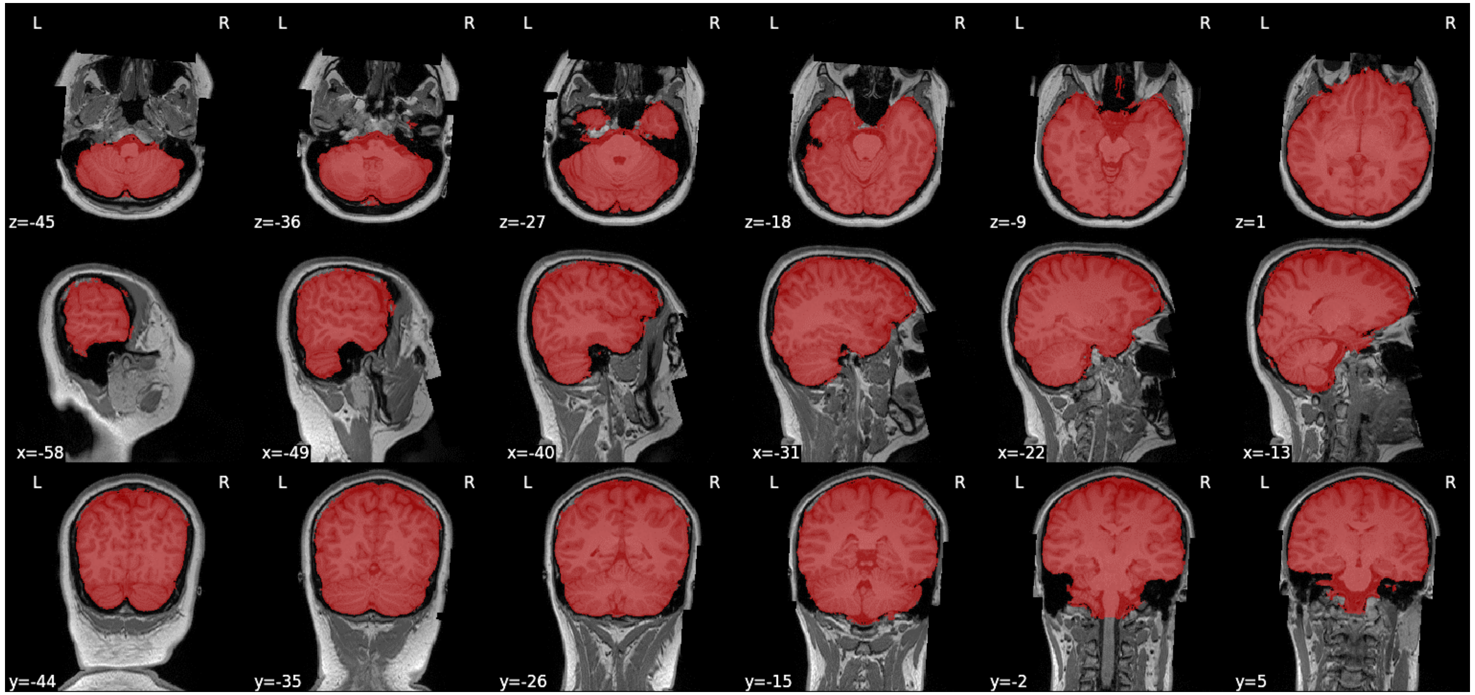


MUNI | CEITEC

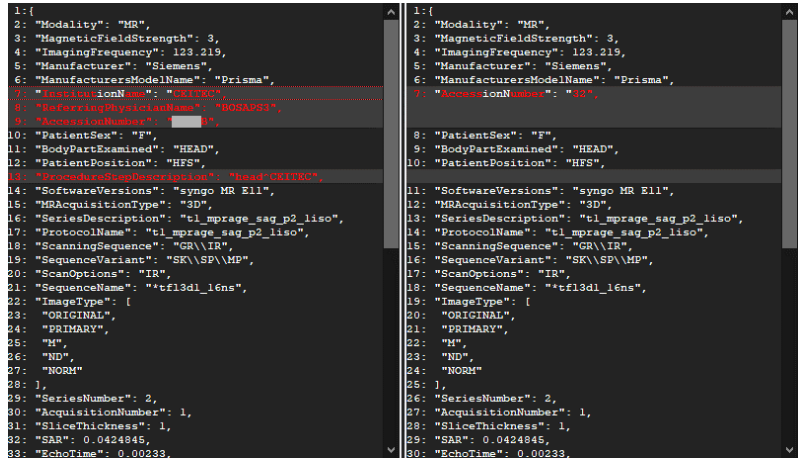
Deface

- FSL_deface
  - BIAS corrected
- Ověření integrity
  - Pomocí ’inbrain’ masky
  - Metriky podobnosti
  - Manuální kontrola

sub-	B	fsl_deface	bias_corrected	0	0	2	1
sub-	B	fsl_deface	bias_corrected	0	0	2	1
sub-	B	fsl_deface	bias_corrected	0,00011	0,01881	1,99998	0,99999
sub-	B	fsl_deface	bias_corrected	0	0	2	1
sub-	B	fsl_deface	bias_corrected	0	0	2	1
sub-	B	fsl_deface	bias_corrected	0	0	2	1
sub-	B	fsl_deface	bias_corrected	0	0	2	1
sub-	B	fsl_deface	bias_corrected	0	0	2	1
sub-	B	fsl_deface	bias_corrected	0	0	2	1
sub-	B	fsl_deface	bias_corrected	0	0	2	1
sub-	B	fsl_deface	bias_corrected	0	0	2	1
sub-	B	fsl_deface	bias_corrected	0	0	2	1
sub-	B	fsl_deface	bias_corrected	0	0	2	1
sub-	B	fsl_deface	bias_corrected	0	0	2	1
sub-	B	fsl_deface	bias_corrected	0	0	2	1
sub-	B	fsl_deface	bias_corrected	0	0	2	1
sub-	B	fsl_deface	bias_corrected	0	0	2	1
sub-	B	fsl_deface	bias_corrected	0	0	2	1
sub-	B	fsl_deface	bias_corrected	0	0	2	1
sub-	B	fsl_deface	bias_corrected	0	0	2	1
sub-	B	fsl_deface	bias_corrected	0	0	2	1
sub-	B	fsl_deface	bias_corrected	0	0	2	1
sub-	B	fsl_deface	bias_corrected	0	0	2	1
sub-	B	fsl_deface	bias_corrected	0	0	2	1
sub-	B	fsl_deface	bias_corrected	0	0	2	1
sub-	B	fsl_deface	bias_corrected	0	0	2	1
sub-	B	fsl_deface	bias_corrected	0	0	2	1
sub-	B	fsl_deface	bias_corrected	1,6E-05	0,00319	2	1
sub-	B	fsl_deface	bias_corrected	0	0	2	1



Anonymizace (vlastní MATLAB skript)



- Převodní tabulka pro VisitID
  - Náhodné přiřazení čísel (1234A ➔ 08)
  - Na konci procesu bude smazána
- Odstranění/nahrazení citlivých údajů
  - Instituce, akviziční protokol (CEITEC, adresa, head^ceitec)
  - Projekt = ReferringPhysician (BOSAPS3)
  - VisitID = AccessionNumber (1234A)
  - Osobní údaje (jméno, příjmení, datum narození, rodné číslo) – zde již pseudonymizováno
  - Pozor na jednoznačné identifikátory (StudyUID, SeriesUID, časová značka)
- Aktualizace metadat (participants.tsv, sub-*.json, *.nii hlavička)
- Přejmenování názvu souborů a složek

MUNI | CEITEC

Doplnění informací o datasetu

- Dataset_description.json
- README (volný text)
- Participants.json, task-*_bold.json, aj.

dataset_description.json

```
{
 "Name": "Multi-echo simultaneous multislice fMRI dataset: Effect of acquisition parameters on fMRI data",
 "BIDSVersion": "1.8.0",
 "DatasetType": "raw",
 "Authors": [
  "Mikl Michal",
  "Kovarova Anezka",
  "Gajdos Martin",
  "Marecek Radek",
  "Novakova Marie",
  "Slavicek Tomas"
 ],
 "License": "CC0",
 "Acknowledgements": "The work was supported from the European Regional Development Fund project National infrastr
 "HowToAcknowledge": "We acknowledge the core facility MAFIL supported by the Czech-BioImaging large RI project (I
 "Funding": [
  "LM2018129 and LM2023050 funded by MEYS CR"
 ],
 "ReferencesAndLinks": [
  "https://mafil.ceitec.cz"
 ],
 "DatasetDOI": "10.58071/njhx-b011",
 "GeneratedBy": [
  {
   "Name": "mafil_dicom2preBIDS",
   "Version": "v1.07"
  }
 ]
}
```

Brain Imaging Data Structure v1.8.0

Search

Brain Imaging Data Structure v1.8.0

The BIDS Specification

Introduction
Common principles
Modality agnostic files
Modality specific files
Derivatives
Longitudinal and multi-site studies
Glossary
BIDS Extension Proposals
Appendix
Changelog

The BIDS Starter Kit

Website
Tutorials
GitHub repository

Modality agnostic files

Dataset description

Templates:

- `dataset_description.json`
- `README`
- `CHANGES`
- `LICENSE`

`dataset_description.json`

The file `dataset_description.json` is a JSON file describing the dataset. Every dataset MUST include this file with the following fields:

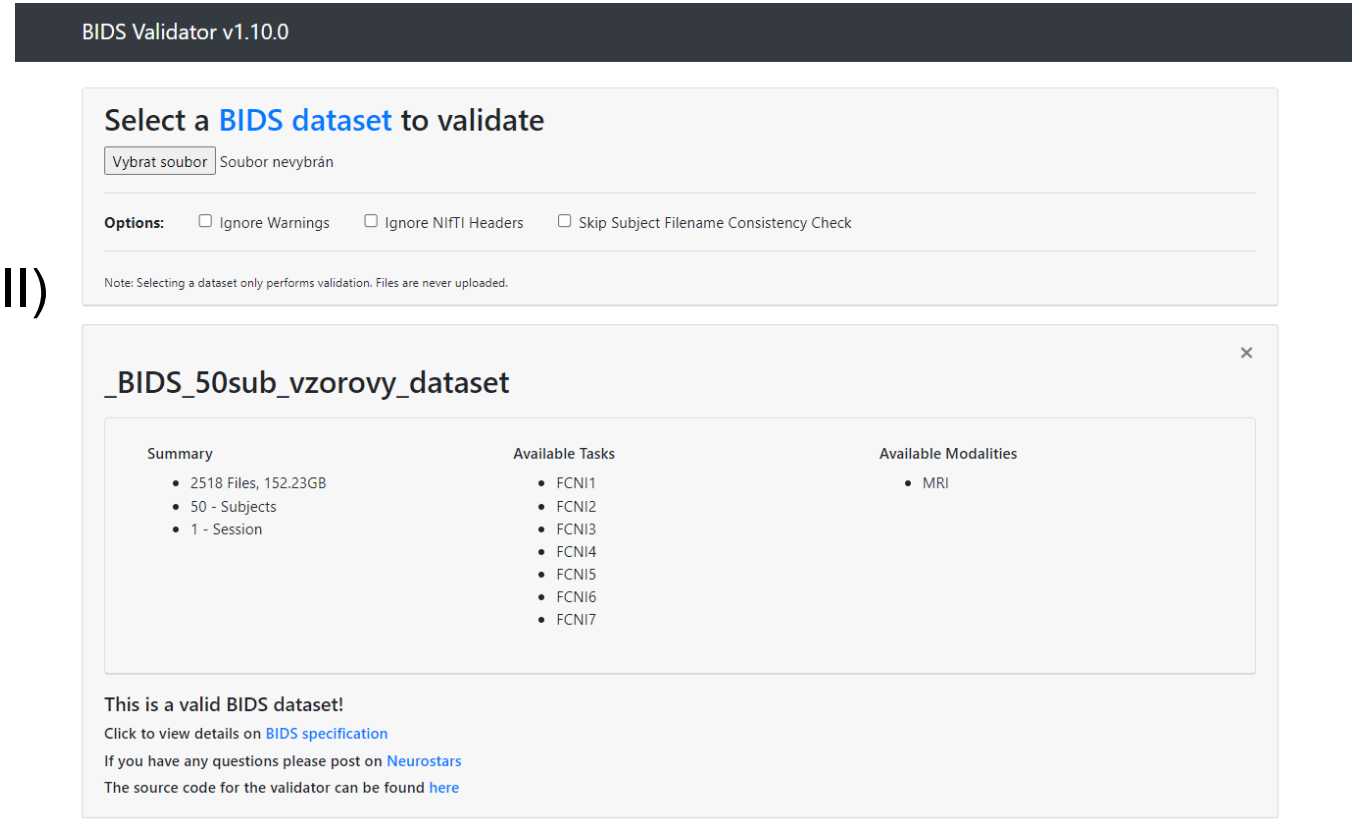
Key name	Requirement Level	Data type	Description
Name	REQUIRED	string	Name of the dataset.
BIDSVersion	REQUIRED	string	The version of the BIDS standard that was used.
HEDVersion	RECOMMENDED	string or array of strings	If HED tags are used: The version of the HED schema used to validate HED tags for study. May include a single schema or a base schema and one or more library schema.
DatasetLinks	REQUIRED if BIDS URIs are used	object of strings	Used to map a given `<dataset-name>` from a BIDS URI of the form `bids:<dataset-name>:path/within/dataset` to a local or

Table of contents

Dataset description
dataset_description.json
Derived dataset and pipeline description
README
CHANGES
LICENSE
Participants file
Samples file
Phenotypic and assessment data
Scans file
Sessions file
Code

Validace

- Online verze – snadné použití
  - https://bids-standard.github.io/bids-validator/
- Kontroluje metadata i data (hlavičku NII)
- Platné názvosloví pro daná data
  - Tzv. ‘Entity Table’
  - Např. nelze multiecho pro T1w
- Uniformní struktura napříč subjekty
  - Např. sekvence navíc ➔ warning
- Včetně parametrů akvizice
  - Např. různé TR u stejné bold sekvence
- Výjimky – soubor .bidsignore



Přidělení DOI (spravuje MAFIL)

DataCite Fabrica

About Support GCNC.CCLSVN

Multimodal and Functional Imaging Laboratory / DOIs

10.58071/njhx-b011

Update DOI (Form)
Update DOI (File Upload)

Findable

Metadata Export
DataCite XML
DataCite JSON
Schema.org JSON-LD
BibTeX

DOI created
February 17, 2023 at 12:36:30 UTC

DOI registered
February 17, 2023 at 12:42:06 UTC

DOI last updated
February 17, 2023 at 13:16:06 UTC

Metadata version
3

Schema
DataCite Metadata Schema 4

URL

https://openneuro.org/datasets/ds004499

Metadata

Summary View

Multi-echo simultaneous multislice fMRI dataset: Effect of acquisition parameters on fMRI data

Dataset

Michal Mikl, Anezka Kovarova, Martin Gajdoš, Radek Mareček, Marie Nováková, Tomáš Slavíček,
Dataset published 2023 via MAFIL-CEITEC-MU

https://doi.org/10.58071/njhx-b011

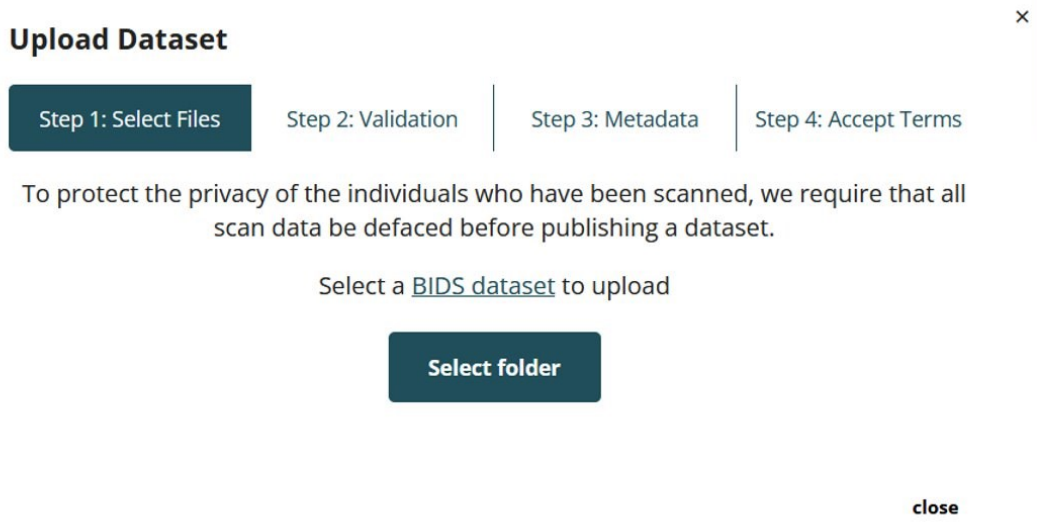
Citation

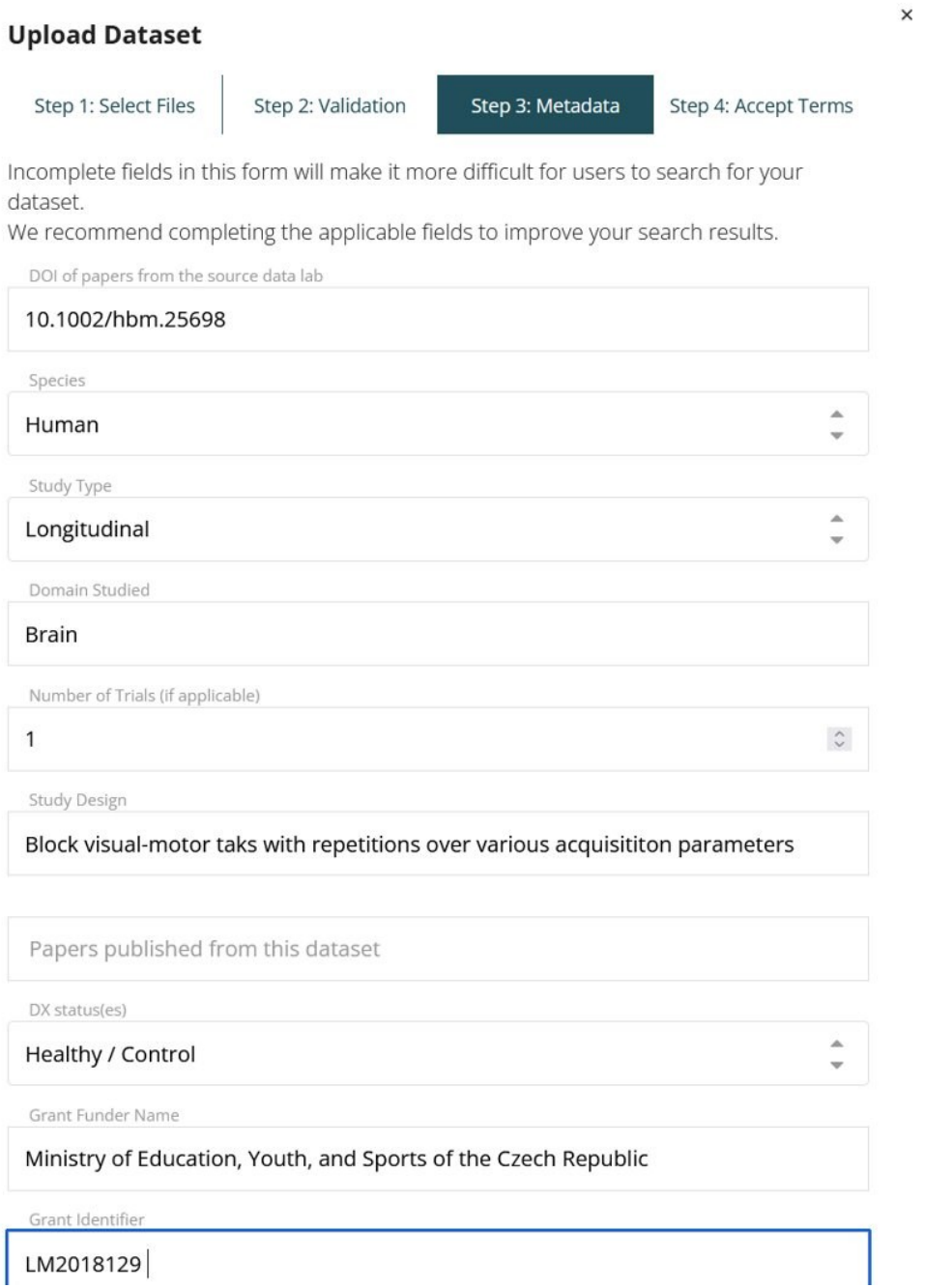
APA

Mikl, M., Kovarova, A., Gajdoš, M., Mareček, R., Nováková, M., & Slavíček, T. (2023). *Multi-echo simultaneous multislice fMRI dataset: Effect of acquisition parameters on fMRI data* [Data set]. MAFIL-CEITEC-MU. https://doi.org/10.58071/NJHX-B011

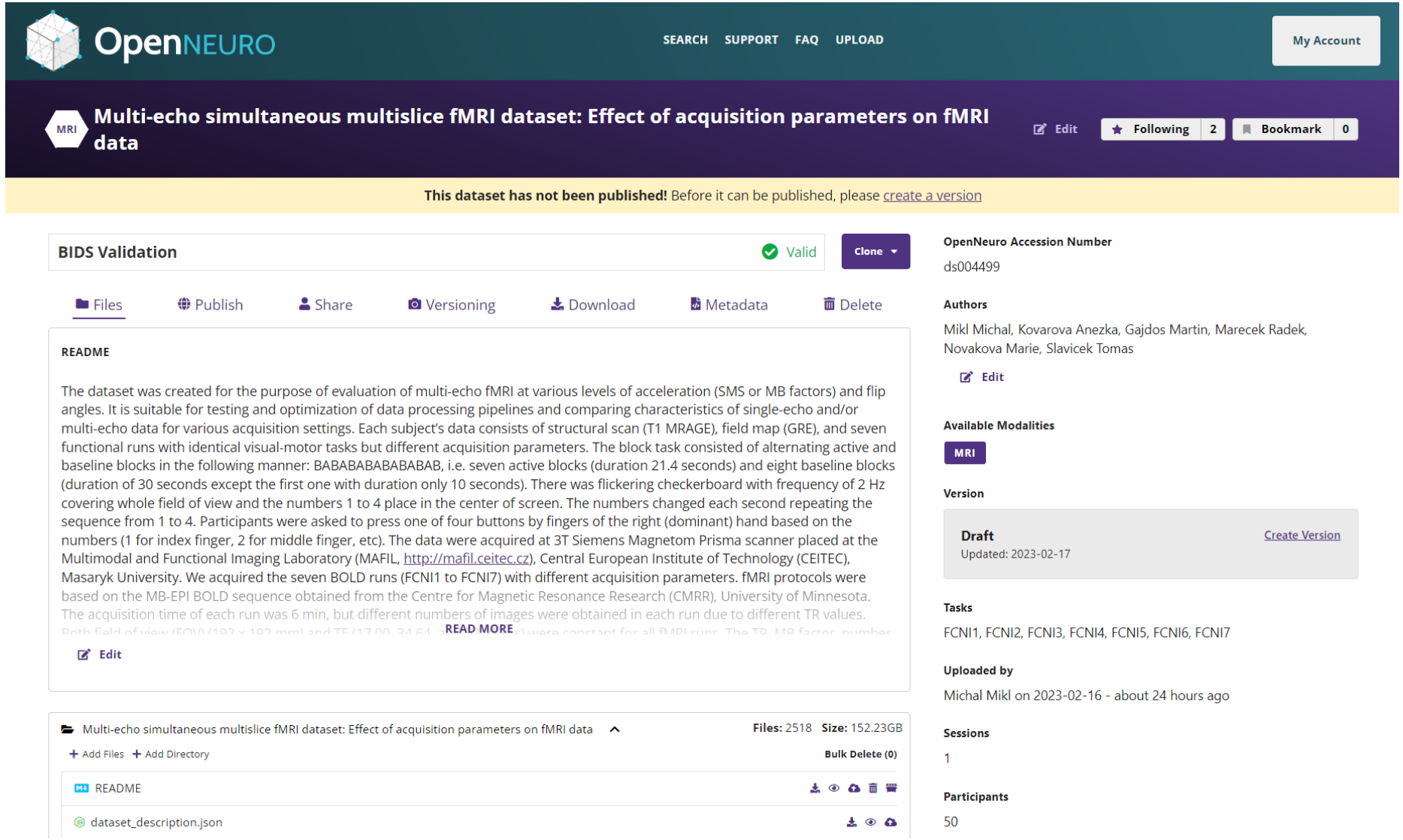
Upload do repozitáře

- https://openneuro.org/
- Po nahrání aktualizace DOI (URL)





Ověření záznamu DOI



ITEC

Díky za pozornost!

Otázky?