AFFORD deliverables WP2.3.2: Open Research Data (ORD) hub in Gitlab pages

https://crsuzh.pages.uzh.ch/afford_page/

Gorka Fraga Gonzalez , 22.09.2023

Content

Overview

'Front-end'

- Hub - for ORD sharing
- Wiki - collaborative documentation

'Back-end'

- Data flow – link preview data to sources
- Metadata - create standards

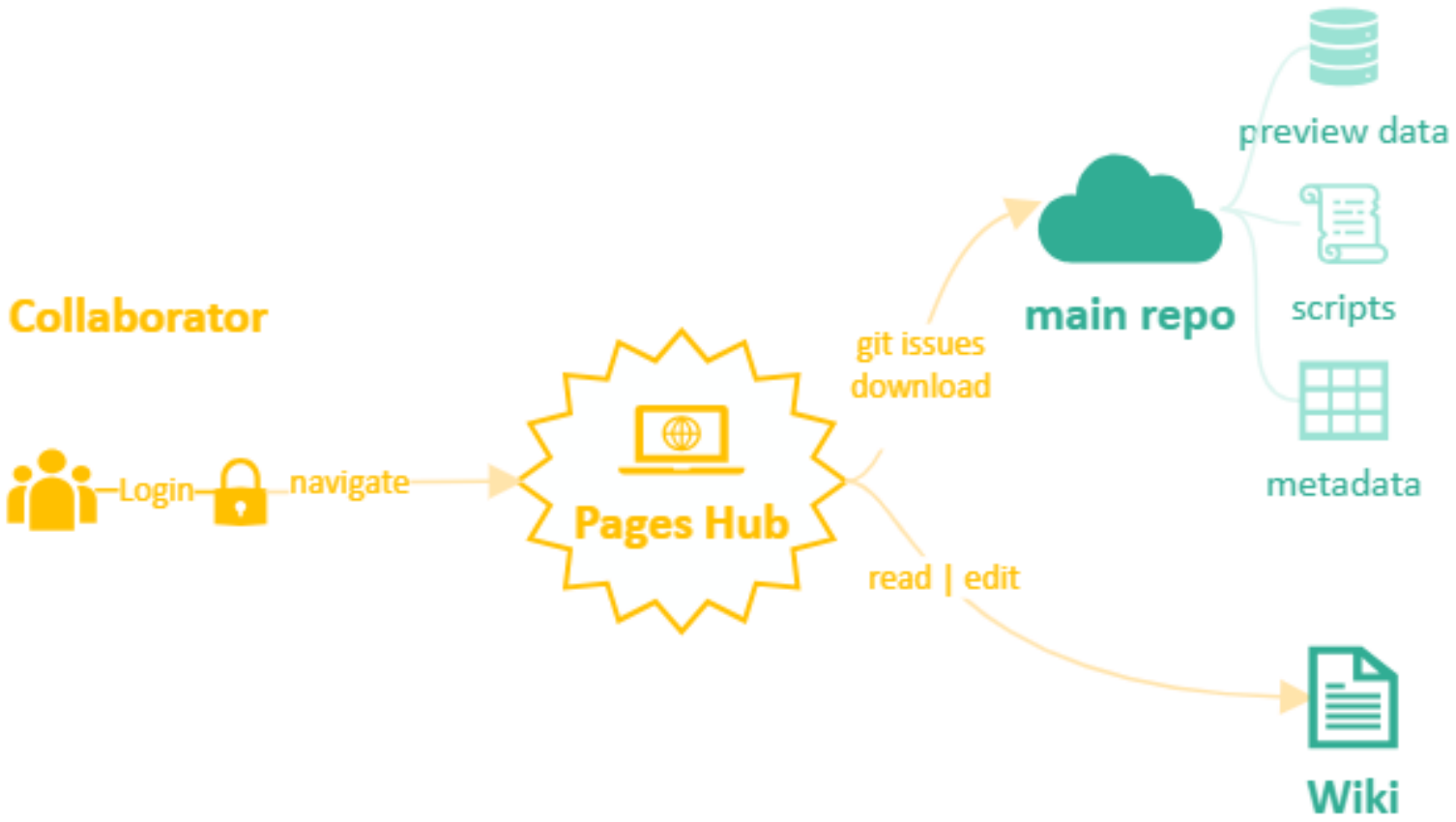
Summary & Next steps …

Overview of Gitlab workflow

Owner - A lab or researcher managing a project or a series of data packages. They could have a **Gitlab group** set up e.g., *gitlab.uzh.ch/interfacegroup*

Repository scope - e.g., Data from Interface group for the Fluid Dynamics CNS Sinergia project will include several synchrotron data packages from Japan, France, etc.
e.g., gitlab.uzh.ch/interfacegroup/FluDynCNS

Overview of Gitlab workflow



Collaborator - A lab or researcher in direct collaboration with the owner e.g., other members of the Fluid Dynamics CNS project

Hub

What?

- **Central hub** to metadata, documents, preview data and source data links

Why?

- Improve **findability**, **accessibility, interoperability, showcasing data**

- **Version control** and avoid duplicate files and data transfer

How?

- **Gitlab repository** with access public or **restricted** to collaborators

Hub

In-browser Hub Demo: https://crsuzh.pages.uzh.ch/afford_page/

Highlight:

- Arrangement by tab
- Selection and copying of multiple rows
- Simple coding (R Markdown streamlining in progress)
- Data table additional extensions
- Continuous integration (show repo)

Wiki

What?

- **Central documentation** site (e.g., Fluid dynamics – Interface)

Why?

- Same as hub: version control, findability, accessibility

- *Wiki* is **easy to edit** by collaborators

- *Markdown* keeps formatting simple and with high compatibility

How?

- **Markdown files** in **Wiki** through **Gitlab repository**

Wiki

In-browser Wiki Demo:

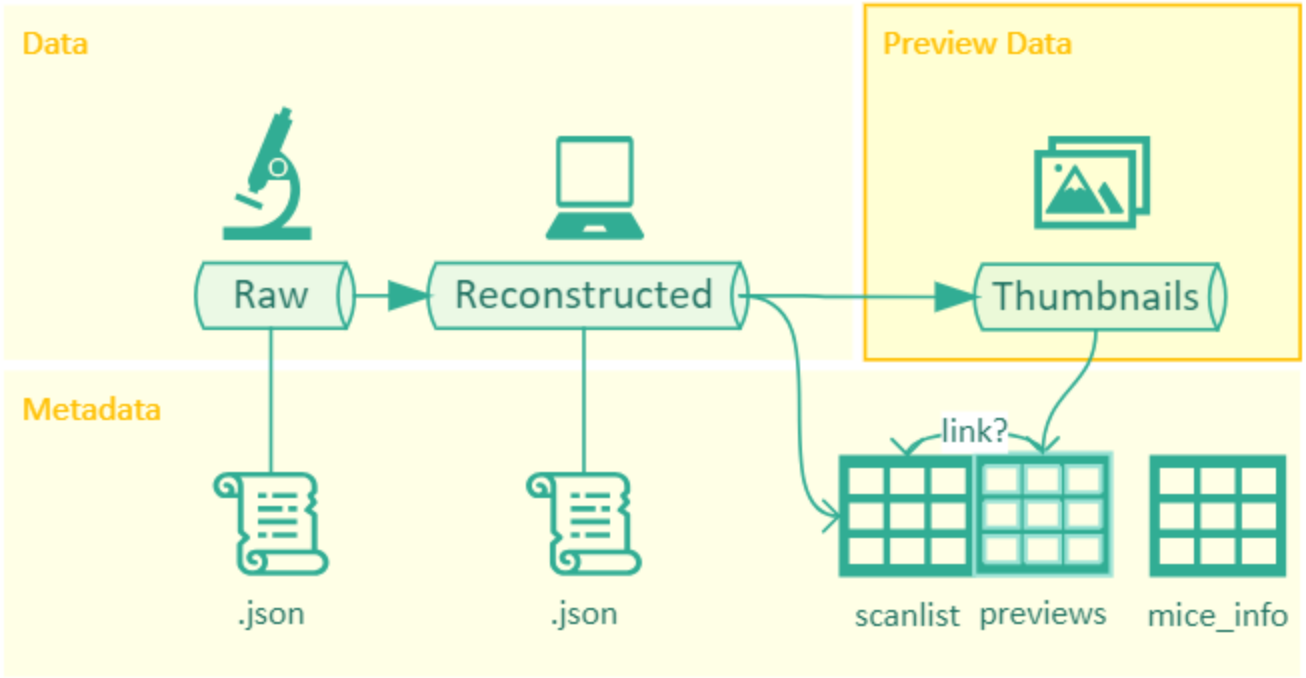
https://gitlab.uzh.ch/crsuzh/afford_page/-/wikis/Home

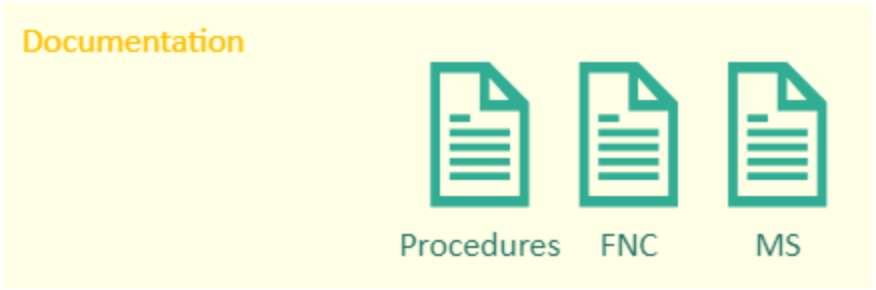
Highlights:

- Track versions history, ‘last-edited-by’ info
- Rich text editor
- Note: it is a ‘hidden’ repository separate from Gitlab pages

Data and metadata flow

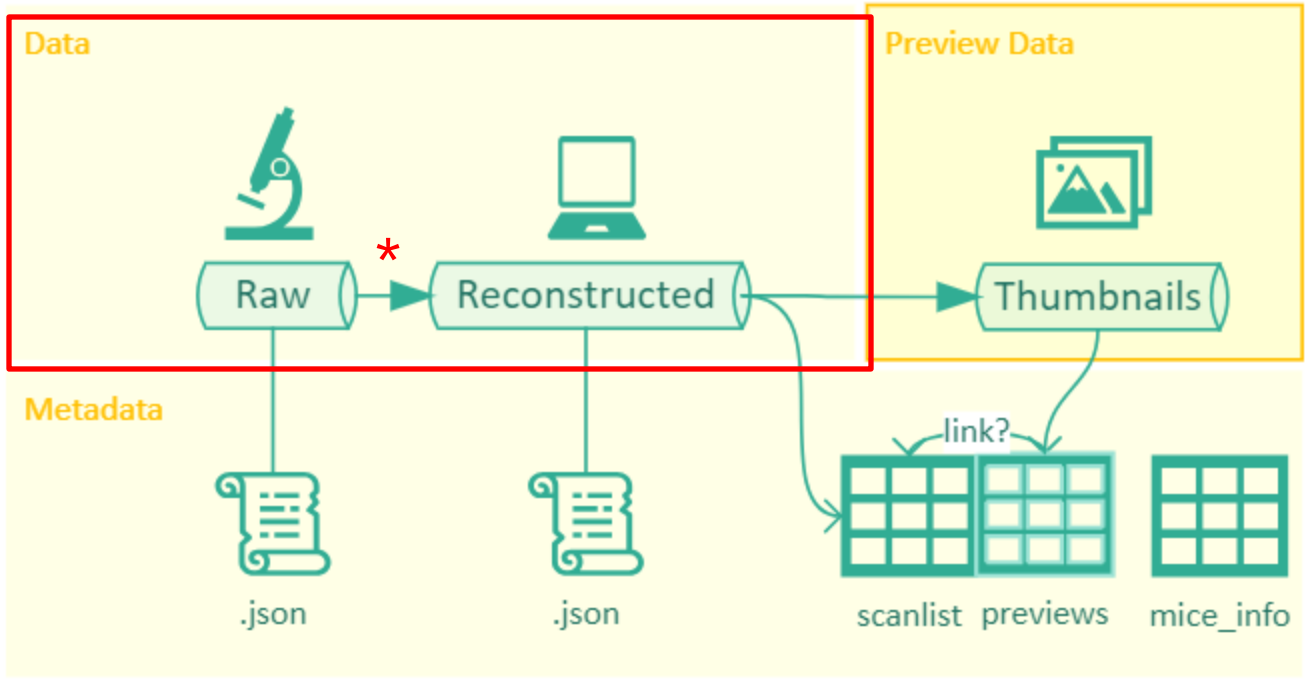
Per data package (e.g., Japan-Spring 2023)

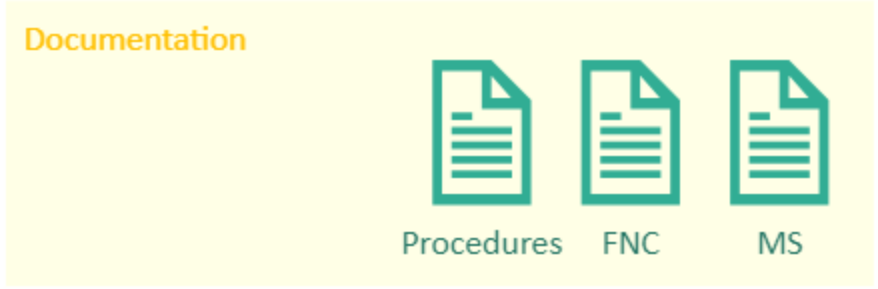




Data and metadata flow

Per data package (e.g., Japan-Spring 2023)



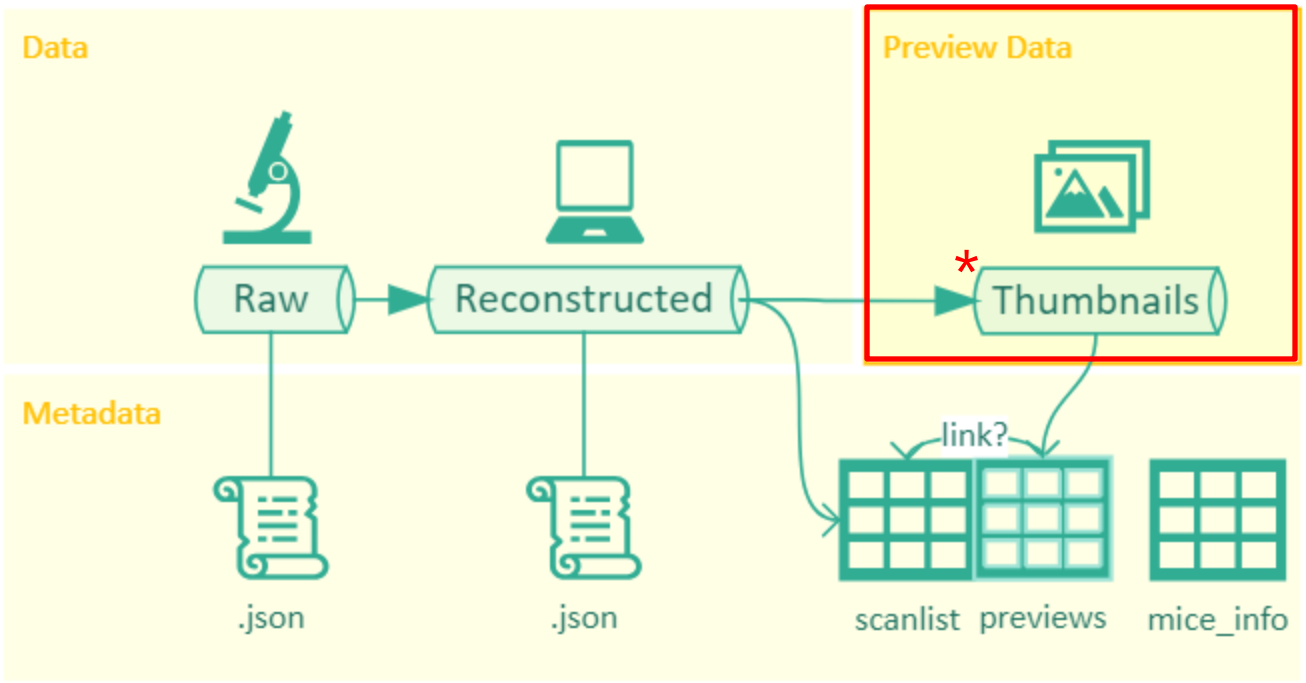


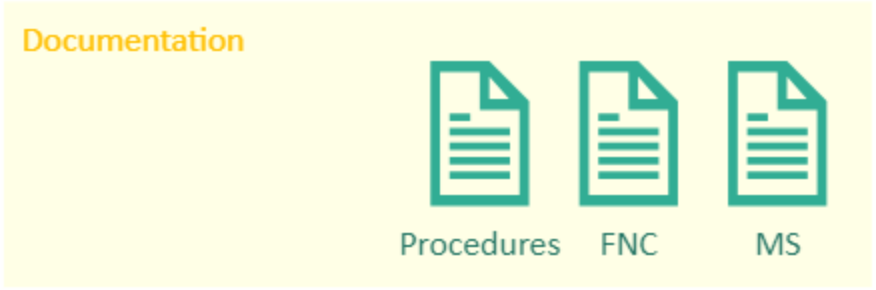
The recorded **data** may undergo minimal preprocessing or **reconstruction** to be useable

- * The **script** to do this should be provided and documented with all parameters
- * **Metadata** of reconstructed data should contain the relevant reconstruction parameters (see JSON)

Data and metadata flow

Per data package (e.g., Japan-Spring 2023)





'Preview data' is generated to assess quality, validate, explore results, share with collaborators: that is the data showed in the local HUB

* The **script** to create this data should be provided / documented in the repo

The **link** between the preview file and the source data should be **preserved** (e.g., in metadata tables)

Metadata should also indicate **what exactly** the thumbnail shows (e.g., center slice n. xx)

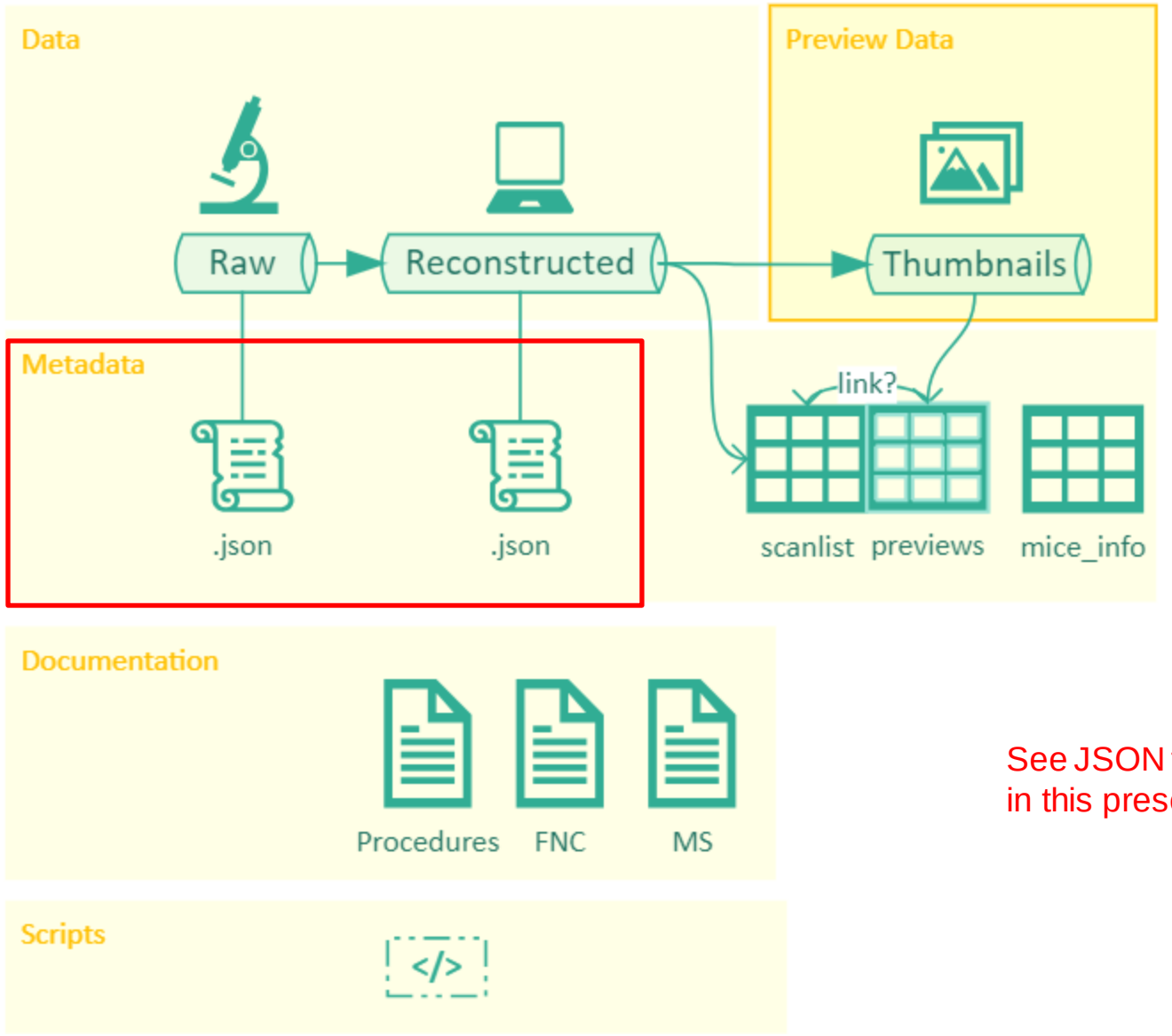
? e.g., Slice 749 or slice 750 of the reconstructed image? Wiki info vs power shell

! Now we have 3 scripts to go from reconstructed to preview data: one powershell script to rename(?) some images and then two R scripts to compress them to jpg and extract filename metadata for the hub table.

This could be simplified to a single script-→ store in *utilities*

Data and metadata flow

Per data package (e.g., Japan-Spring 2023)



JSON metadata files to accompany each data file

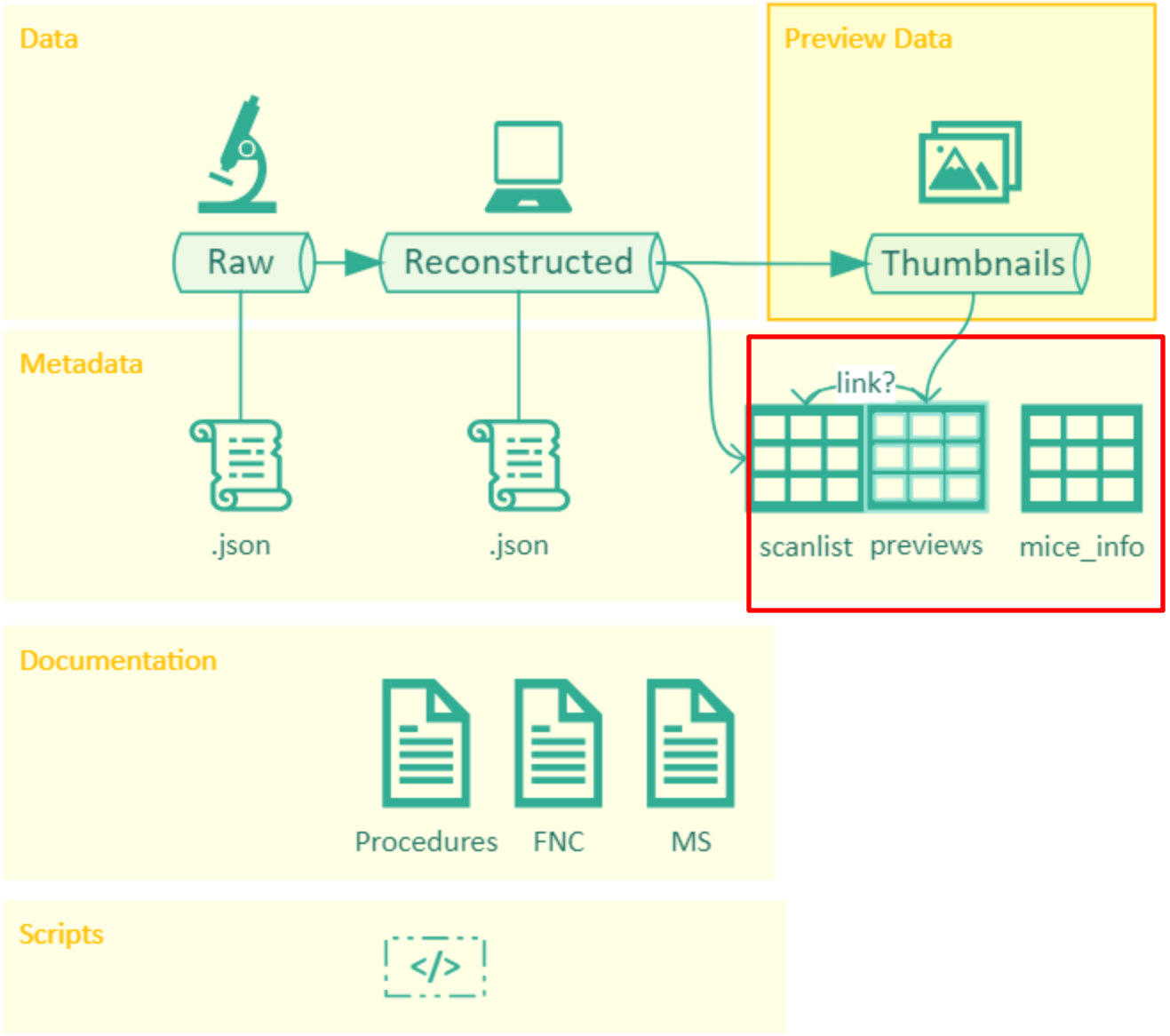
- Basic data **provenance** information (e.g., where, by who, when it was collected)

- More advanced provenance info: Recording parameters, software, hardware, and anything that happened to the file from recording

See JSON format proposals later in this presentation

Data and metadata flow

Per data package (e.g., Japan-Spring 2023)



Metadata tables

Wide-formatted tables

Summarizing the relevant files of the data package

- *Scanlist* files describing source recordings
- *Previews metadata* that can be rendered interactive linking to preview images (HUB)
- *Mice_info* specimen information (wide format, one specimen per row)

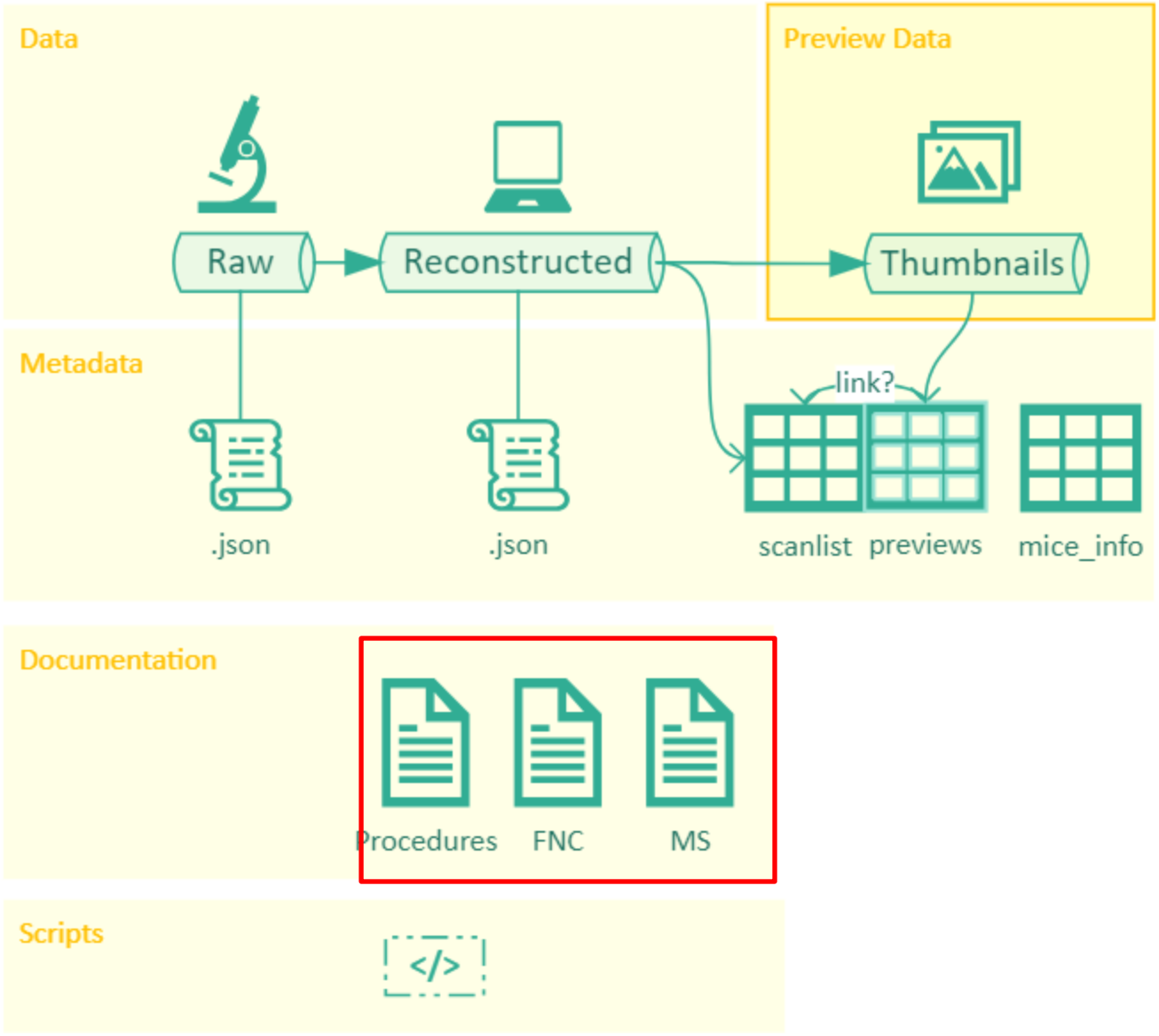
- *Additional* tables may be necessary for lab tests, control measurements, etc.

A **codebook** should clarify variable names for each table

Located at Gitlab repository: **no duplicates** should exist

Data and metadata flow

Per data package (e.g., Japan-Spring 2023)

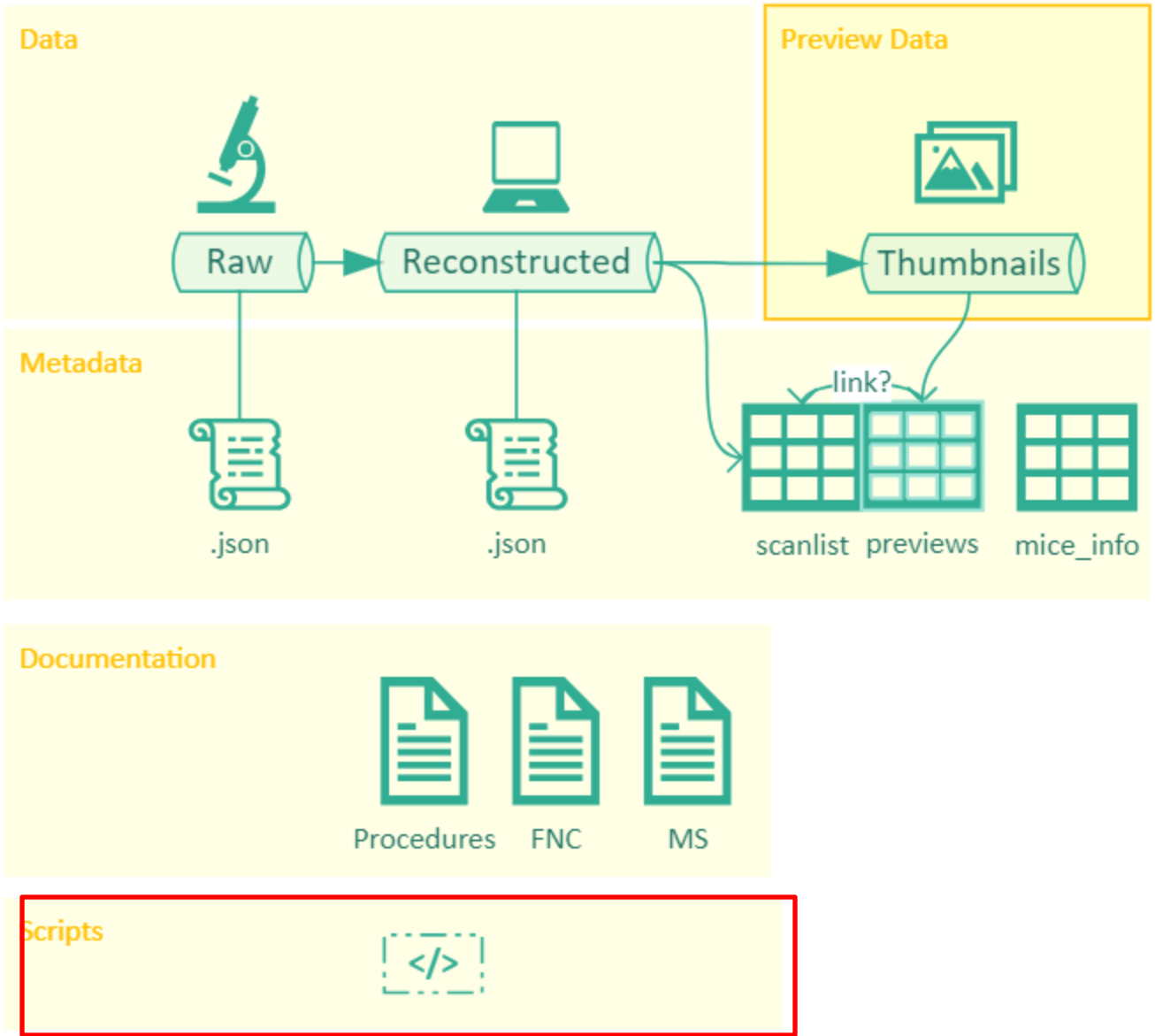


Documentation
Lab procedures, file naming convention and metadata specifications to clarify how data, metadata and scripts are organized

Also in the Gitlab repository

Data and metadata flow

Per data package (e.g., Japan-Spring 2023)



Scripts

All code to:

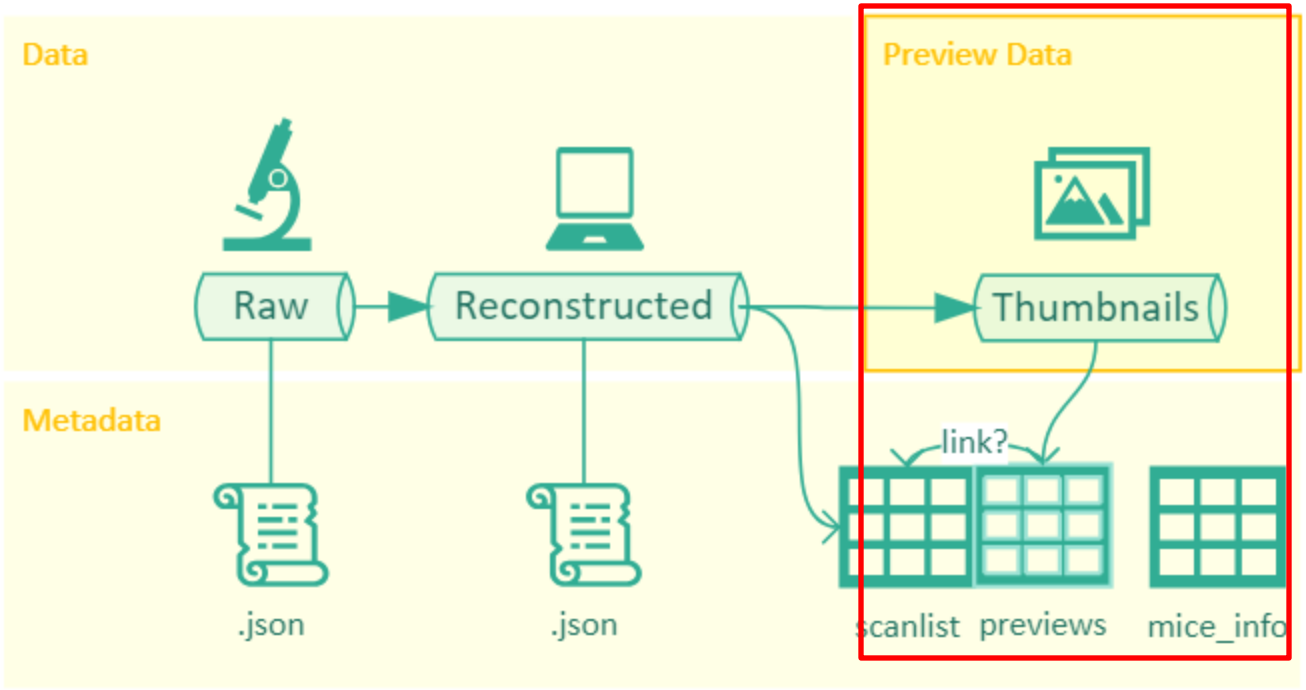
- organize the data package
- generate metadata json and tables
- Generate thumbnails
- Rename or do other file edits
- render the Hub

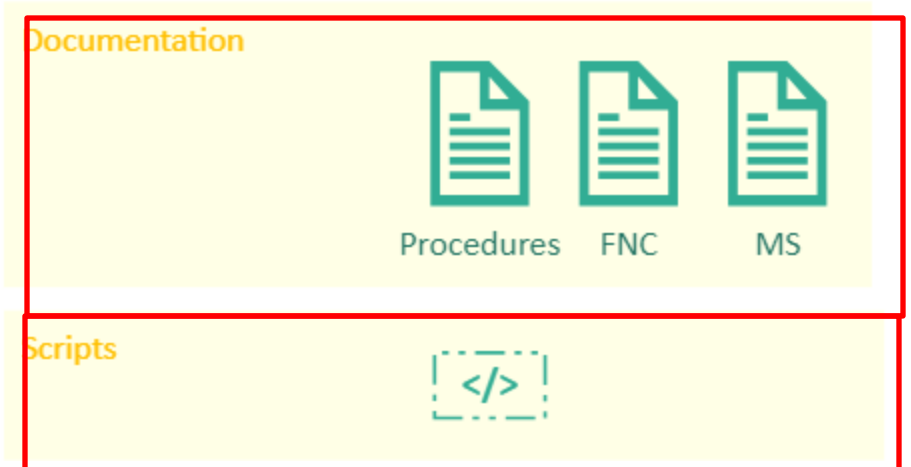
It can be stored in the same Gitlab repository as the website with the Hub

Code snippets or small scripts can also be written in the wiki as part of the documentation

Data and metadata flow

Per data package (e.g., Japan-Spring 2023)





In Gitlab Repo

Summary of a data package content

Per data package (e.g., Japan-Spring 2023)		
File	Content	Formatting Issues
Data		
Raw	Directly measured from instruments	
Reconstructed	Minimally preprocessed so that it can be interpreted	
Thumbnails	Preview images for sharing, inspection, QA, etc	
Documentation		Markdown recommended (e.g., in wiki)
Procedures	How the data was collected and reconstructed	"
FNC	Filename convention	"
MS	Metadata specifications: fields descriptions, how tables were generated	"
Metadata tables		Incl. codebook sheet (variable dictionary) No color coding for interoperability. Wide-format
Scanlist *.xlsx	List of available files and their recording parameters. If applicable, the corresponding thumbnail files	-- *rendered interactive for navigation in the hub*
Mice info *.xlsx	Additional info per specimen	
Lab control *.xlsx	Additional info on lab calibration, tests, etc	
hash tables	Archived files and hashes for verification	
Metadata json (per data file)		
<datafilename>.json	To be stored together with each individual data file, basic data provenance tracking (e.g., where, what equipment, who collected, basic rec parameters)	Script-generated: Common info hardcoded, file-specific info derived from data

Metadata standards and JSON files

What?

- Metadata files stored together with recorded datasets

Why?

- Metadata tables are not stored with each data record (e.g., spec-01 folder) and do not contain all info per data object or file
- *Why .JSON?* Concise, compact, and easy to parse and generate

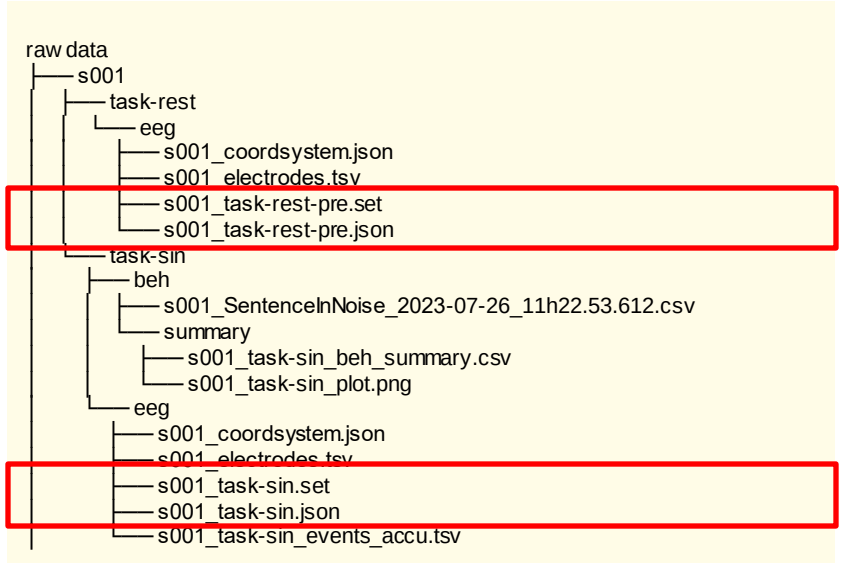
How?

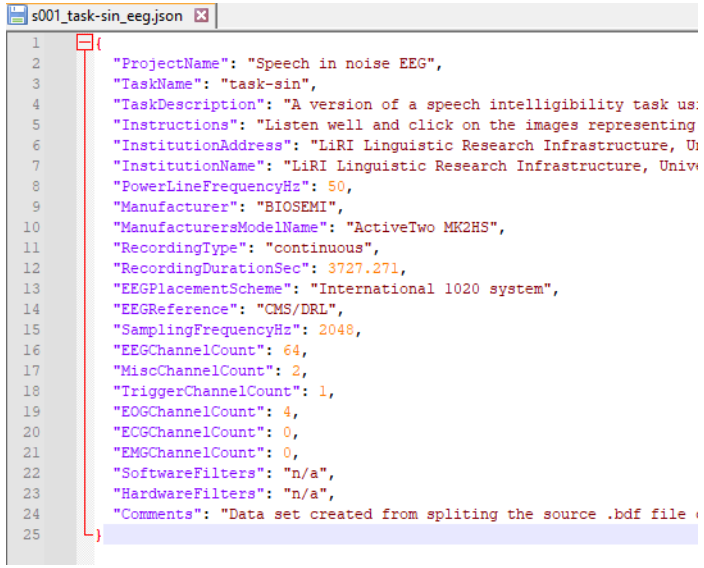
- Manual generation is not recommended (scriptable with any language) :
- Hardcoded info (e.g., project title) + data-derived info (e.g., duration)

Metadata standards and JSON files

Example of an EEG project following BIDS guidelines

Each EEG dataset (.set) comes with a .json file (with same name). Provides common info to all project files as well as file-specific info (e.g., duration)



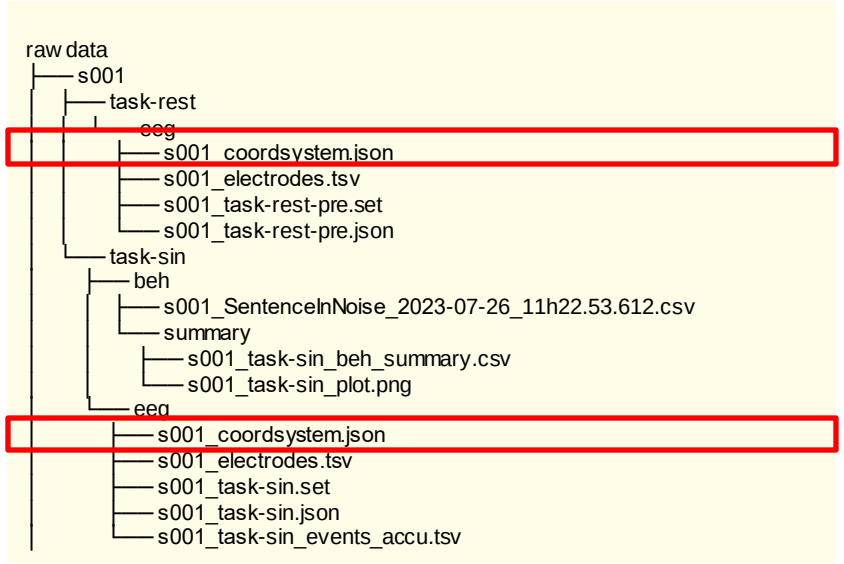


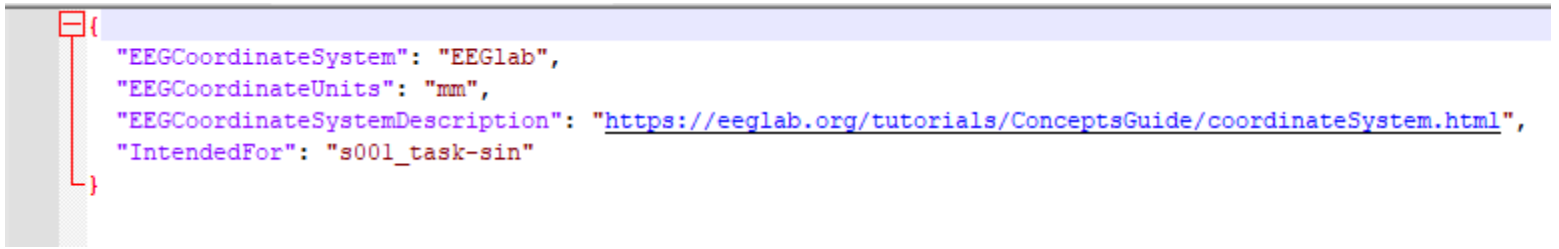
Metadata standards and JSON files

Example of an EEG project following BIDS guidelines

Additional json files containing details on the recording system

E.g., coordinate system, to interpret the coordinates provided in metadata file *'s001_electrodes.tsv'*



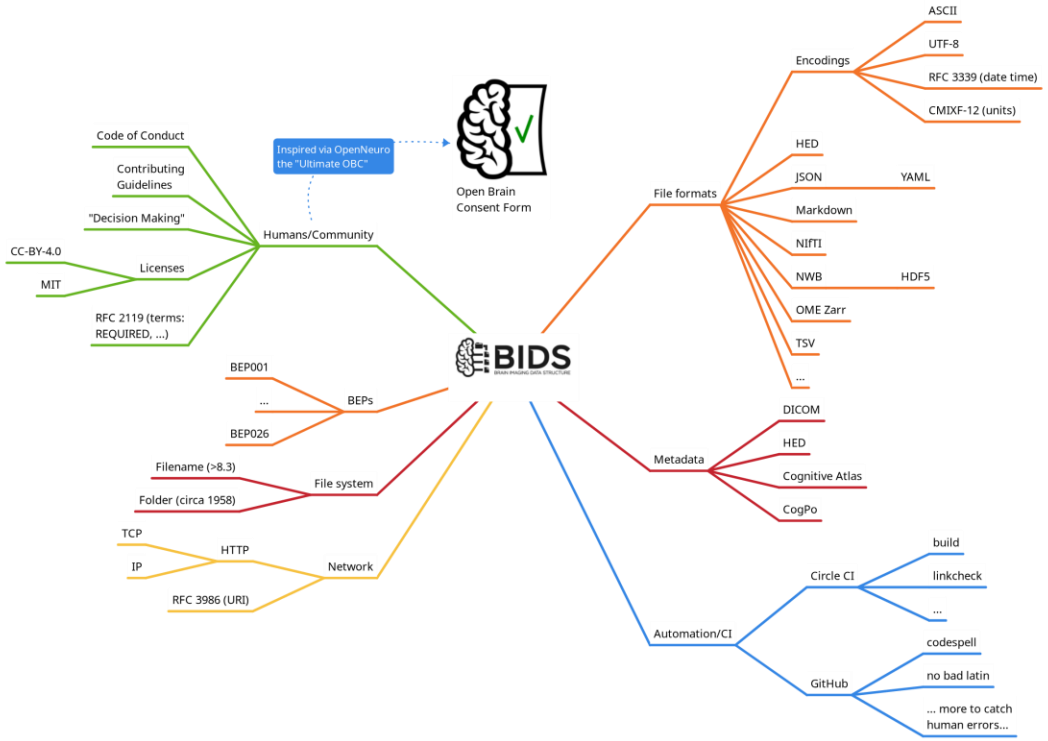


Metadata standards

- Are there existing standards in the field?

Brain Imaging Data Structure https://bids.neuroimaging.io/

"Most advanced and widely used option for structuring files according to MRI sequences and including standardized metadata" Kalantari et al, 2023



Metadata standards

BIDS Extensions Proposals (BEP)

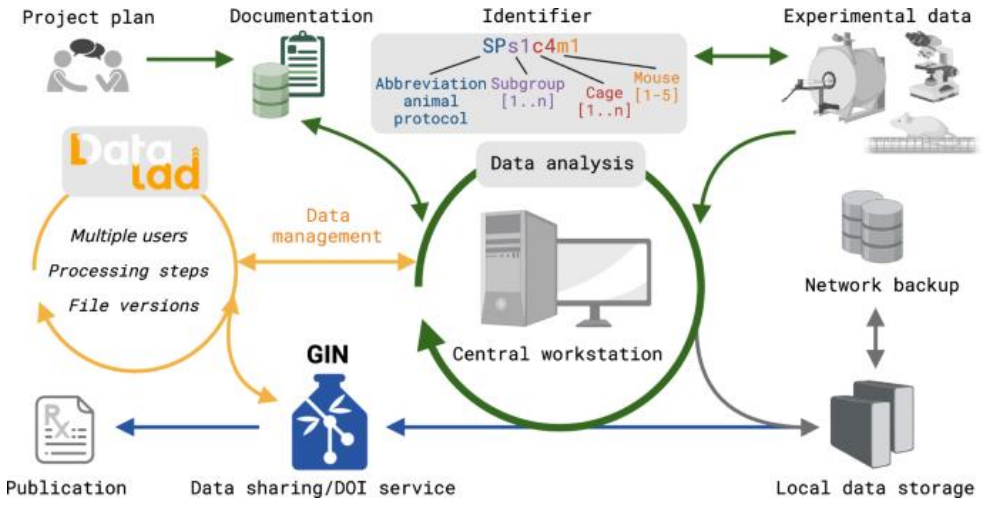
https://bids.neuroimaging.io/get_involved.html#extending-the-bids-specification

- You can propose a BIDS extension to your data **modality**
- Each proposal is an **open google doc** where experts can give comments
- Once accepted it will be **integrated** in the BIDS
- Currently active BEPs: **CT, computational models, microscopy, animal electrophysiology, …**
- Integration with BIDS will facilitate visibility, adoption and version tracking

Metadata standards

- Publications related to BIDS extensions

Kalantari, A., Szczepanik, M., Heunis, S., Mönch, C., Hanke, M., Wachtler, T., & Aswendt, M. (2023). How to establish and maintain a multimodal animal research dataset using DataLad**.** Scientific data, 10(1), 357. https://www.nature.com/articles/s41597-023-02242-8



- **DataLad** opensource software for research data management https://www.datalad.org/
- **YODA** principles for data analysis https://handbook.datalad.org/en/latest/basics/101-127-yoda.html
- **BIDS** compatibility of structure and metadata

Metadata standards

- There are other publications related to BIDS extensions

Bourget, M. H., Kamentsky, L., Ghosh, S. S., Mazzamuto, G., Lazari, A., Markiewicz, C. J., ... & Salo, T. (2022). Microscopy-BIDS: An extension to the Brain Imaging Data Structure for microscopy data. *Frontiers in Neuroscience*, *16*, 871228. https://www.frontiersin.org/articles/10.3389/fnins.2022.871228/full

Next steps on WP2

CRS

- Hub table code generalizable (troubleshoot, streamline, documentation)

- Beginner Tutorials (currently collection for beginners)

Interface group (+ CRS)

- Defining JSON fields based on BIDs proposals and own data
- Clarifying workflow from source data to preview files
- Search if other standards (besides BIDS) exist for data types in FDCNS
- Share hub with FDCNS collaborator to attempt a set up from scratch

Thank you

“SCANLIST” metadata

	A	B	C	D	
1	SPring-8 BL20B2 Beamtime 2023A1208 May 2023				
2	**Scan name**	**Subscan**	**ECG data**	**On Disk No.**	**Recons**
231	JP31-invivo-1scan-phase2-injection-brain			Disk 2	x
232	JP31-invivo-5scan-phase3-postinjection-brain	5 scan		Disk 2	x
233	JP32-invivo-10scan-phase4-reinjection-brain-continuous	10 scan		Disk 2	x
234	JP32-invivo-1scan-phase1-preinjection-brain			Disk 2	x
235	JP32-invivo-1scan-phase1-preinjection-nose			Disk 2	x
236	JP32-invivo-1scan-phase2-injection-brain			Disk 2	x
237	JP32-invivo-1scan-phase3-postinjection-nose-60min			Disk 2	x
238	JP32-invivo-3scan-phase3-postinjection-brain-70min-90min	3 scan		Disk 2	x
239	JP32-invivo-5scan-phase3-postinjection-brain	5 scan		Disk 2	x
240	JP33-invivo-10scan-phase4-reinjection-brain-continuous	10 scan		Disk 2	x
241	JP33-invivo-1scan-phase1-preinjection-brain			Disk 2	x
242	JP33-invivo-1scan-phase1-preinjection-nose			Disk 2	x
243	JP33-invivo-1scan-phase2-injection-brain			Disk 2	x
244	JP33-invivo-1scan-phase3-postinjection-nose-60min			Disk 2	x
245	JP33-invivo-3scan-phase3-postinjection-brain-70min-90min	3 scan		Disk 2	x
246	JP33-invivo-5scan-phase3-postinjection-brain	5 scan		Disk 2	x
247	JP34-invivo-scan1-brain			Disk 2	x
248	JP34-invivo-scan2-nose			Disk 2	x
249	JP34-invivo-scan3-spine			Disk 2	x
250	JP34-invivo-scan4-60scans-brain	60 scan		Disk 2	x
251	purified_water_eppendorf01			Disk 1	x
252	purified_water_eppendorf02			Disk 1	x
253	test01	0		Disk 1	x
254	test02	0		Disk 1	x

“Preview” files metadata (showed in HUB)

	A	B
1	file	specID
601	JP30-invivo-5scan-phase3-postinjection-brain_002_rec00749.jpg	JP30
602	JP30-invivo-5scan-phase3-postinjection-brain_003_rec00749.jpg	JP30
603	JP30-invivo-5scan-phase3-postinjection-brain_004_rec00749.jpg	JP30
604	JP30-invivo-5scan-phase3-postinjection-brain_005_rec00749.jpg	JP30
605	JP31-dead-1scan-phase3-postinjection-brain-65min_rec00749.jpg	JP31
606	JP31-dead-1scan-phase3-postinjection-nose-60min_rec00749.jpg	JP31
607	JP31-invivo-1scan-phase1-preinjection-brain_rec00749.jpg	JP31
608	JP31-invivo-1scan-phase1-preinjection-nose_rec00749.jpg	JP31
609	JP31-invivo-1scan-phase2-injection-brain_rec00749.jpg	JP31
610	JP31-invivo-5scan-phase3-postinjection-brain_001_rec00749.jpg	JP31
611	JP31-invivo-5scan-phase3-postinjection-brain_002_rec00749.jpg	JP31
612	JP31-invivo-5scan-phase3-postinjection-brain_003_rec00749.jpg	JP31
613	JP31-invivo-5scan-phase3-postinjection-brain_004_rec00749.jpg	JP31
614	JP31-invivo-5scan-phase3-postinjection-brain_005_rec00749.jpg	JP31
615	JP32-invivo-10scan-phase4-reinjection-brain-continuous_001_rec00749.	JP32
616	JP32-invivo-10scan-phase4-reinjection-brain-continuous_002_rec00749.	JP32
617	JP32-invivo-10scan-phase4-reinjection-brain-continuous_003_rec00749.	JP32
618	JP32-invivo-10scan-phase4-reinjection-brain-continuous_004_rec00749.	JP32
619	JP32-invivo-10scan-phase4-reinjection-brain-continuous_005_rec00749.	JP32
620	JP32-invivo-10scan-phase4-reinjection-brain-continuous_006_rec00749.	JP32
621	JP32-invivo-10scan-phase4-reinjection-brain-continuous_007_rec00749.	JP32
622	JP32-invivo-10scan-phase4-reinjection-brain-continuous_008_rec00749.	JP32
623	JP32-invivo-10scan-phase4-reinjection-brain-continuous_009_rec00749.	JP32
624	JP32-invivo-10scan-phase4-reinjection-brain-continuous_010_rec00749.	JP32

“SCANLIST” metadata

“Preview” files metadata (showed in HUB)

	A	B	C	D	E
1	SPring-8 BL20B2 Beamtime 2023A1208 May 2023				
2	Scan name	Subsca	ECG	On Disk No.	Reconstructed scan No.
5	JP14-invivo-1prescan-h2-spine		yes	Disk 1	x
6	JP14-invivo-55scan-phase2-maintain	55 scan	yes	Disk 1	x
7	JP15-invivo-10scan-phase1-injection	10 scan	yes	Disk 1	x
8	JP15-invivo-10scan-phase2-maintain	10 scan	yes	Disk 1	x
9	JP15-invivo-10scan-phase3-injection	10 scan	yes	Disk 1	x
00	JP15-invivo-1prescan-h1-bulbs		yes	Disk 1	x
01	JP15-invivo-1prescan-h2-spine		yes	Disk 1	x
02	JP15-invivo-20scan-phase4-maintain	20 scan	yes	Disk 1	x
03	JP15-invivo-30scan-phase4-maintain2	30 scan	yes	Disk 1	x
04	JP16-dead-10scan-phase3-injection	10 scan		Disk 2	x
05	JP16-invivo-10scan-phase1-injection	10 scan		Disk 2	x
06	JP16-invivo-1prescan-h1-bulbs			Disk 2	x
07	JP16-invivo-1prescan-h2-spine			Disk 2	x
08	JP16-invivo-50scan-phase2-maintain	50 scan		Disk 2	1-49
09	JP17-invivo-10scan-phase1-infusion	10 scan		Disk 2	x
				Disk 2	x
				Disk 2	x
				Disk 2	x
				Disk 2	x
				Disk 2	x
15	JP18-invivo-1prescan-h1-bulbs			Disk 2	x
16	JP18-invivo-1prescan-h1-bulbs-v2			Disk 2	x
17	JP18-invivo-1prescan-h2-spine			Disk 2	x
18	JP19-invivo-10scan-phase1-wait	10 scan		Disk 2	x
19	JP19-invivo-1postscan-h1-bulbs			Disk 2	x
20	JP19-invivo-1postscan-h2-spine			Disk 2	x
21	JP19-invivo-1prescan-h1-bulbs			Disk 2	x
22	JP19-invivo-1prescan-h2-spine			Disk 2	x
23	JP19-invivo-1scan-phase2-preinjection			Disk 2	x
24	JP19-invivo-1scan-phase3-postinjection			Disk 2	x
25	JP19-invivo-1scan-phase4-10min			Disk 2	x
26	JP19-invivo-1scan-phase4-20min			Disk 2	x
27	JP19-invivo-1scan-phase4-30min			Disk 2	x
28	JP19-invivo-1scan-phase5-preinjection			Disk 2	x
29	JP19-invivo-1scan-phase6-postinjection			Disk 2	x
30	JP19-invivo-1scan-phase7-10min			Disk 2	x

No files available for JP16 invivo 50 scan. Notes in scanlist file says file is incomplete

	A	B	C
1	file	specID	status
331	JP15-invivo-30scan-phase4-maintain2_026_rec00749.jpg	JP15	invivo
332	JP15-invivo-30scan-phase4-maintain2_027_rec00749.jpg	JP15	invivo
333	JP15-invivo-30scan-phase4-maintain2_028_rec00749.jpg	JP15	invivo
334	JP15-invivo-30scan-phase4-maintain2_029_rec00749.jpg	JP15	invivo
335	JP15-invivo-30scan-phase4-maintain2_030_rec00749.jpg	JP15	invivo
336	JP16-dead-10scan-phase3-injection_001_rec00749.jpg	JP16	dead
337	JP16-dead-10scan-phase3-injection_002_rec00749.jpg	JP16	dead
338	JP16-dead-10scan-phase3-injection_003_rec00749.jpg	JP16	dead
339	JP16-dead-10scan-phase3-injection_004_rec00749.jpg	JP16	dead
340	JP16-dead-10scan-phase3-injection_005_rec00749.jpg	JP16	dead
341	JP16-dead-10scan-phase3-injection_006_rec00749.jpg	JP16	dead
342	JP16-dead-10scan-phase3-injection_007_rec00749.jpg	JP16	dead
343	JP16-dead-10scan-phase3-injection_008_rec00749.jpg	JP16	dead
344	JP16-dead-10scan-phase3-injection_009_rec00749.jpg	JP16	dead
345	JP16-dead-10scan-phase3-injection_010_rec00749.jpg	JP16	dead
346	JP16-invivo-10scan-phase1-injection_001_rec00749.jpg	JP16	invivo
347	JP16-invivo-10scan-phase1-injection_002_rec00749.jpg	JP16	invivo
348	JP16-invivo-10scan-phase1-injection_003_rec00749.jpg	JP16	invivo
349	JP16-invivo-10scan-phase1-injection_004_rec00749.jpg	JP16	invivo
350	JP16-invivo-10scan-phase1-injection_005_rec00749.jpg	JP16	invivo
351	JP16-invivo-10scan-phase1-injection_006_rec00749.jpg	JP16	invivo
352	JP16-invivo-10scan-phase1-injection_007_rec00749.jpg	JP16	invivo
353	JP16-invivo-10scan-phase1-injection_008_rec00749.jpg	JP16	invivo
354	JP16-invivo-10scan-phase1-injection_009_rec00749.jpg	JP16	invivo
355	JP16-invivo-10scan-phase1-injection_010_rec00749.jpg	JP16	invivo
356	JP16-invivo-1prescan-h1-bulbs_rec00749.jpg	JP16	invivo
357	JP16-invivo-1prescan-h2-spine_rec00749.jpg	JP16	invivo
358	JP17-invivo-10scan-phase1-infusion_001_rec00749.jpg	JP17	invivo
359	JP17-invivo-10scan-phase1-infusion_002_rec00749.jpg	JP17	invivo
360	JP17-invivo-10scan-phase1-infusion_003_rec00749.jpg	JP17	invivo
361	JP17-invivo-10scan-phase1-infusion_004_rec00749.jpg	JP17	invivo
362	JP17-invivo-10scan-phase1-infusion_005_rec00749.jpg	JP17	invivo
363	JP17-invivo-10scan-phase1-infusion_006_rec00749.jpg	JP17	invivo
364	JP17-invivo-10scan-phase1-infusion_007_rec00749.jpg	JP17	invivo
365	JP17-invivo-10scan-phase1-infusion_008_rec00749.jpg	JP17	invivo