

Wissenschaft Online “Forschungsdatenmanagement”

Adina Wagner

Psychoinformatics Lab, INM-7

Institute of Experimental Psychology, HHU Düsseldorf

ReproNim/INCF fellow



DOI

10.5281/zenodo.7155599



Acknowledgements

Software

- Joey Hess (git-annex)
- The DataLad team & contributors,
- insb. Yaroslav Halchenko, Michael Hanke

Wissenschaft und FDM

- Psychoinformatics Lab & INM-7
- insb. Laura Waite, Gosia Wierzba, Michael Hanke, Alex Waite, Felix Hoffstädter, Simon Eickhoff

Förderer



Kollaborationen



Neurowissenschaftliche Daten

Herausforderungen und Lösungen im Forschungsdatenmanagement

- Potentiell **proprietäre Formate** bei Hardware und Analysesoftware
- **Heterogene Daten** aus komplexen, multimodalen Messungen
- Große **Datenmengen** durch große Stichproben und detaillierte Messungen
- Viele Daten gelten als Gesundheitsdaten im Rahmen der **DSGVO** mit **besonderem Datenschutz**
- **Viel-schrittige Analyse Pipelines** (Vervielfachung der Datenmenge & “researchers degrees of freedom”)
- Umfangreiches und wachsendes Ökosystem von **Open Source** Formaten und Softwarelösungen
- Offene, community-led Entwicklung von **(Meta-)Datenstandards** (Beispiele: BIDS, OpenMINDS, NWB, ...) und
- Offener **Datenaustausch** ist etablierte Praxis
- Initiativen zum **DSGVO-konformen Datenaustausch** (openbrainconsent)

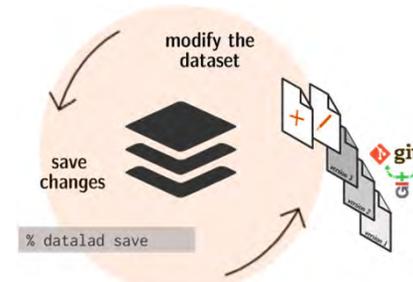
Datenmanagement im INM-7

Data

datalad.org
doi.org/10.21105/joss.03262

- Open Source Tool für Datenmanagement und -Publikation, basierend auf Git/git-annex
- DataLad Datasets können
 - digitale Daten jedes Typs und jeder Größe versionieren,
 - dezentral publizieren (zB GitHub, Open Science Framework, AWS S3, ...)
 - von lokaler oder nicht-lokaler Infrastruktur konsumiert (installiert) werden
 - Prozessierungsschritte reproduzierbar aufzeichnen und wiederholen,
 - Datenverfügbarkeit, -integrität, und Zugriff auf (einzelne) Dateien gewährleisten ohne Dateninhalt lokal speichern zu müssen

```
adina@muninn studyforrest-data-phase2: tig
2015-12-15 12:27 +0100 Michael Hanke  o Regenerate recording-eyegaze to account for the 100ms
2015-12-13 19:31 +0100 Michael Hanke  o Various fixes and robustification
2015-12-12 13:41 +0100 Michael Hanke  o Simplify eyegaze movie merge (was a bit off anyway)
2015-12-11 13:05 +0100 Michael Hanke  o Implement standard eyegaze preprocessing
2015-12-11 11:55 +0100 Michael Hanke  o Attempt to speed up rendering (cache + slicing)
2016-01-18 16:59 +0100 Daniel Kottke
2015-12-09 21:19 +0100 Michael Hanke  o Reintroduce proper offset; support for merged timeline
2015-12-09 20:28 +0100 Michael Hanke  o Use frame median as representative coordinate
2015-12-09 20:25 +0100 Michael Hanke  o Simplifications, no pandas, no double offsetting
2015-12-09 19:54 +0100 Michael Hanke  o PEP8 compliance
2015-11-27 19:22 +0100 Michael Hanke  o Script to render a movie version with eye-gaze distribut
2015-11-27 18:14 +0100 Michael Hanke  o Make eyegaze dataset factory more flexible
2015-10-12 11:43 +0200 Michael Hanke  o Add utility Python code to load a merged eyegaze timeser
2015-10-12 11:19 +0200 Michael Hanke  o Adjust backend
2015-10-12 10:25 +0200 Michael Hanke  o Include subjective story depth ratings
2015-10-12 10:04 +0200 Michael Hanke  o Update description
2015-10-11 11:38 +0200 Michael Hanke  o Basic task description
2015-10-05 08:37 +0200 Michael Hanke  o Include movie stimulus timing info
2015-10-05 08:37 +0200 Michael Hanke  o Rename "recording" files to conform with BIDS
[main] 3ed01bc83340bffe0a8f2337acb71bc91b8f8cd - commit 57 of 81 93%
```



[HIER einfach im Browser ausprobieren](#)

Datenmanagement im INM-7

Data



Laura Waite, Data steward

- interne & externe Datensätze werden standardisiert (zB BIDS), gemäß Datenschutz gesichert (zB Zugriffsrechte), & zentral auffindbar gemacht
- aktuell ~400TB (nur INM-7)

INM-7 Docs

Juseless Learning **Datasets** On-site Digital Admin News

Ethics

Overview

INM-7 SUPERDATASET

Overview

The **INM-7 superdataset** is the main entry point for all data holdings centrally managed for the institute. It includes both internally and externally collected data that have been cleaned-up and normalized. The superdataset is hosted on **JuGit**, improving discoverability, coordination, and issue tracking. The actual content and file information is stored on Juseless' **bulk storage** infrastructure, where data is **password protected** and access is limited to align with the various **Data Use Agreements**.

Data is structured according to the following categories:

- Original:** Data in their original form, as acquired from upstream. This includes raw and derivative data.
- Processed:** Data that are a result of on-site (pre)processing.
- Archived:** The outcome of the **archival process**.
- Containers:** Computational pipelines set up as DataLad container datasets.

datasets_repo

Project information

Repository

Issues 61

Merge requests 0

CI/CD

Security & Compliance

Deployments

Packages and registries

Infrastructure

Monitor

Analytics

Wiki

Snippets

update Data Use Certifications for ABCD dataset post 2022 access renewal
Laura Waite authored 1 week ago

Name	Last commit	Last update
hcp	update Data Use Certifications for HCP-Lifespan datasets	1 week ago
abd @ 3e1c5605	update Data Use Certifications for ABCD dataset post 2022 acces...	1 week ago
ale @ 92d6ad0d	update original dataset READMEs to include citation guidelines	6 months ago
camcan @ bedb851b	update original dataset READMEs to include citation guidelines	6 months ago
enki @ 8e945309	[DATALAD] Recorded changes	2 years ago
enki-pheno @ f4bb3a5c	update original dataset READMEs to include citation guidelines	6 months ago
fcp @ 9a718436	add README for fcp	6 months ago



Datenmanagement im INM-7

- Abgeschlossene Projekte werden archiviert oder publiziert (z.B. Metadatenkatalog; FDM challenge Dataverse-Integration)

The screenshot shows the Jülich Data portal interface. At the top, it identifies the 'Institute of Neuroscience and Medicine - Brain and Behavior (INM-7)'. The main heading is 'Age differences in predicting working memory performance from network-based functional connectivity'. Below this, there is a 'Description' section with a brief summary of the dataset, a 'Subject' section, and a 'Change View' section with 'Table' and 'Tree' options. On the right side, there are buttons for 'Access Dataset', 'Edit Dataset', 'Link Dataset', 'Contact Owner', and 'Share'. A 'Dataset Metrics' section shows '0 Downloads'.

The screenshot shows the DataLad catalog page for the 'studyforrest_multires3t' dataset. It includes the dataset title, authors (Adina Weges, Christian Münch, Laura Walte, Michael Hanke), version information, and a license. There are buttons for 'Download with DataLad' and 'Export metadata'. A 'Description' section provides details about the dataset's origin and purpose. A 'Keywords' section lists terms like 'handedness', 'staffw', 'age', 'BIDS', 'Neuroscience', 'gender', 'subject', 'Studyforrest', and 'orientation'. A 'Properties' section shows 'Subjects: 7', 'Sessions: 0', and 'Tasks: 1'. At the bottom, there is a table with columns 'Property' and 'Value(s)'. The table contains entries for Subject, Task, Acquisition, Run, and Modality.

Property	Value(s)
Subject	["20", "17", "16", "06", "10", "21", "09"]
Task	["orientation"]
Acquisition	["y30", "y20", "y14", "y20mb"]
Run	[1, 2, 3, 4, 5, 6, 7, 8, 9, 10]
Modality	["T1w"]

The screenshot shows an issue page for the 'SpEx Dataset'. The issue title is 'SpEx Dataset' and it was created 21 hours ago by 'Camilleri, Julia'. The main content of the issue is a template for archiving datasets, including instructions on how to provide metadata and a code block with example text and comments. The code block contains the following text:

```
#<!-- METADATA START --> # DO NOT DELETE THIS LINE
# All example text is surrounded with <ex> and </ex>. Please replace the example
# text including the <ex> and </ex> with your data. Delete all non-applicable
# lines and sections.

# Attention: indentation is important and must be preserved!

# information on the study the to-be-archived dataset was created for
study:
# short name or label
name: SpEx - Speech and executive function
```

On the right side of the issue page, there are sections for 'Add a to do', 'Assignee' (Camilleri, Julia), 'Labels' (in metadata valid), 'Milestone' (None), 'Due date' (None), 'Time tracking' (No estimate or time spent), 'Confidentiality' (Not confidential), and 'Lock issue' (Unlocked).

Die Herausforderung: Datenmanagement at scale

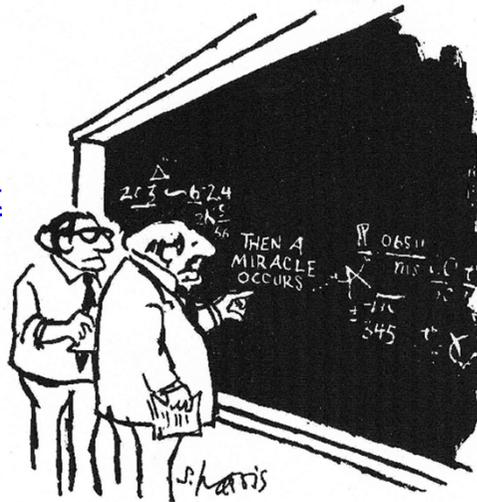
Technische Komplexität darf Reproduzierbarkeit nicht verhindern



- Gehört zu den einflussreichsten und größten Datensätzen
- Technische Herausforderungen machen Reproduzier- & Nachvollziehbarkeit (Stichwort: FAIRness) schwierig

Warum FAIR?

NARPS Study ([Botvinik-Nezer et al., 2020](#)): 70 independent research groups, investigating 9 hypothesis, on the same data: Consistent conclusions for four hypothesis



"I think you should be more explicit here in step two."

“The variety of methodological & analytical choices is not the enemy to computational reproducibility, **the challenge lies in encoding those degrees of freedom in a standardized, ideally machine-readable way**”
[Gilmore et al., 2017](#)

Die Herausforderung: Datenmanagement at scale

Article | [Open Access](#) | [Published: 11 March 2022](#)

FAIRly big: A framework for computationally reproducible processing of large-scale data

[Adina S. Wagner](#) , [Laura K. Waite](#), [Małgorzata Wierzba](#), [Felix Hoffstaedter](#), [Alexander Q. Waite](#), [Benjamin Poldrack](#), [Simon B.](#)

[Eickhoff](#) & [Michael Hanke](#)

[Scientific Data](#) 9, Article number: 80 (2022) | [Cite this article](#)

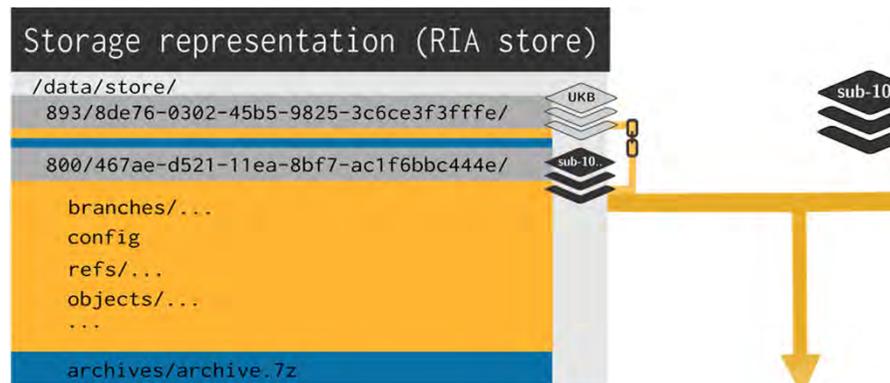


- Kombiniert etablierte Open Source Software (Software-container, Job scheduling, Versionskontrolle)
- Nutzt Prinzipien der Softwareentwicklung für Datenanalysen:
 - Daten, Code, & Softwareumgebungen sind “Abhängigkeiten”
 - Parallel ausgeführte Teilanalysen werden aus eigenen “Feature Branches” zusammengeführt
- Erlaubt Überprüfung computationaler Reproduzierbarkeit von Ergebnissen, unabhängig von ursprünglich genutzter Infrastruktur

FAIRly big: interne Datenrepräsentation

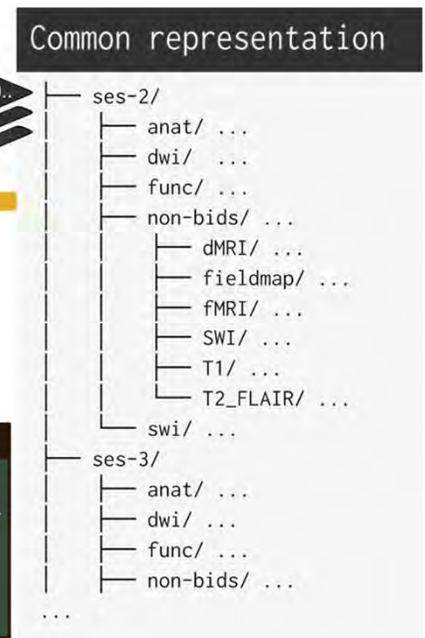
- Speicherplatz- & Dateianzahl-sparsame interne RIA Repräsentation von DataLad Datasets
- Erlaubt auch verschlüsselte Speicherung und Transport zur Erfüllung von Datenschutzanforderungen

“Backend”



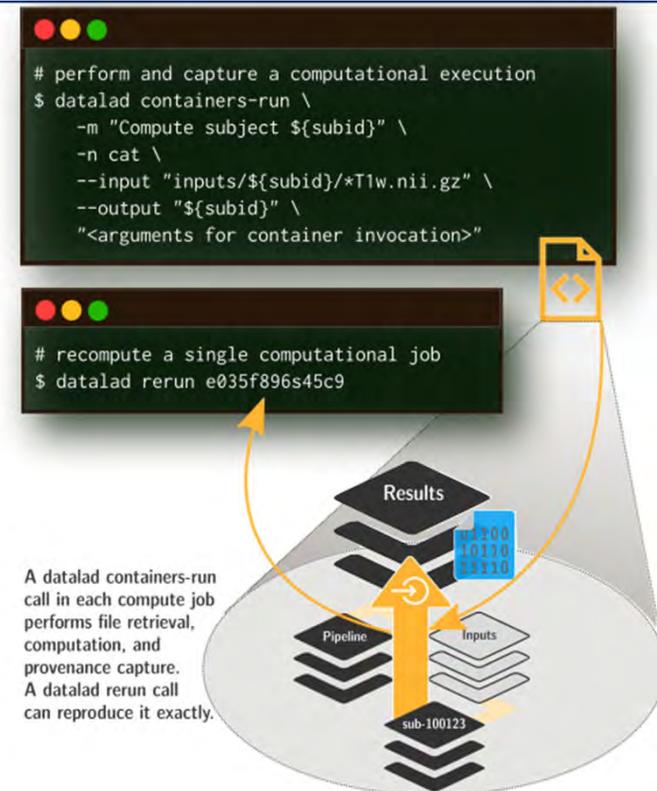
```
# clone one participant dataset from the store
$ datalad clone ria+file:///store#800467ae-d521-11ea-8bf7-ac1f6bbc444e sub-10..
$ cd sub-10..
# retrieve all of the subjects data from the archives in
# the store
$ datalad get .
```

“Frontend”



FAIRly big: “Ephemeral workspaces”

- (Teil-)Analysen definieren Abhängigkeiten (benötigte Daten, Code, Software). Aus diesen Abhängigkeiten werden “kurzlebige” Analyseumgebungen für jede Berechnung erstellt (ge-bootstrapped).
- Wenn die Berechnung in frisch erstellten Umgebungen erfolgreich ist, sind Abhängigkeiten hinreichend definiert.
- Hinreichend definierte Abhängigkeiten ermöglichen Wiederherstellung der Analyseumgebung auf anderer Infrastruktur

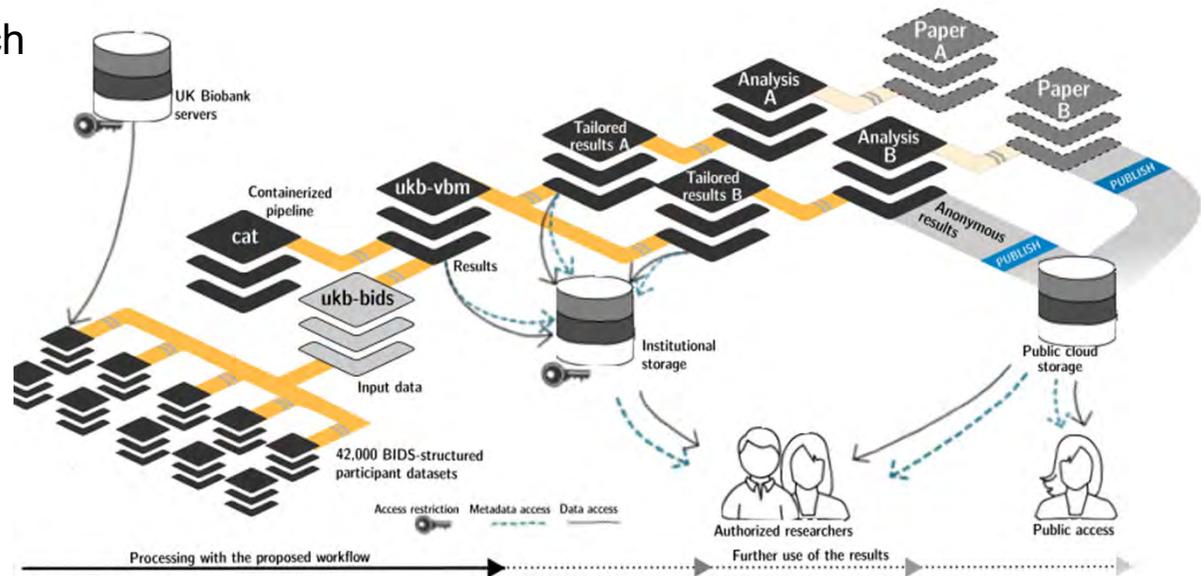


FAIRly big: Transparent und reproduzierbar

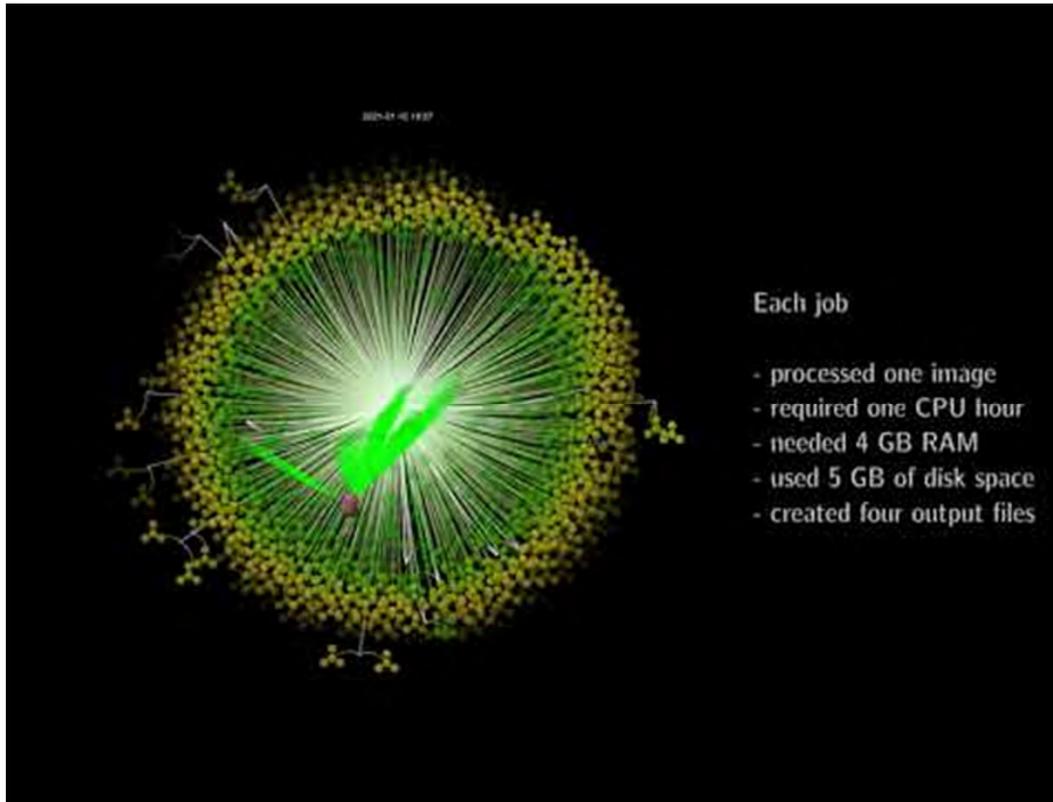


Felix Hoffstädter

- Die Ergebnisse des Frameworks machen die Prozesshistorie so **transparent** und automatisch **nachvollziehbar** wie es der Datenschutz zulässt
- Für uns und andere mittlerweile **Standard** zur internen Datenaufbereitung oder öffentlichem Datenaustausch



Weiterführende Links - Vielen Dank für Ihre Aufmerksamkeit



Tutorial



Talk



Visualisierung



[Zusatzfolien nachfolgend]

Provenienz

- Autor und Datum
- Freitextbeschreibung
- Softwarecontainer und -konfiguration
- Daten
- Kommando oder Funktionsaufruf
- Ergebnisse (Pfade & kryptographische Identität)

```
commit e035f896s45c9fac70cn7cc4dbd0dad43907755p
Author: Jane Doe <j.doe@fz-juelich.de>
AuthorDate: Wed Feb 10 18:05:30 2021 +0100
Commit: Jane Doe <j.doe@fz-juelich.de>
CommitDate: Wed Feb 10 18:05:30 2021 +0100

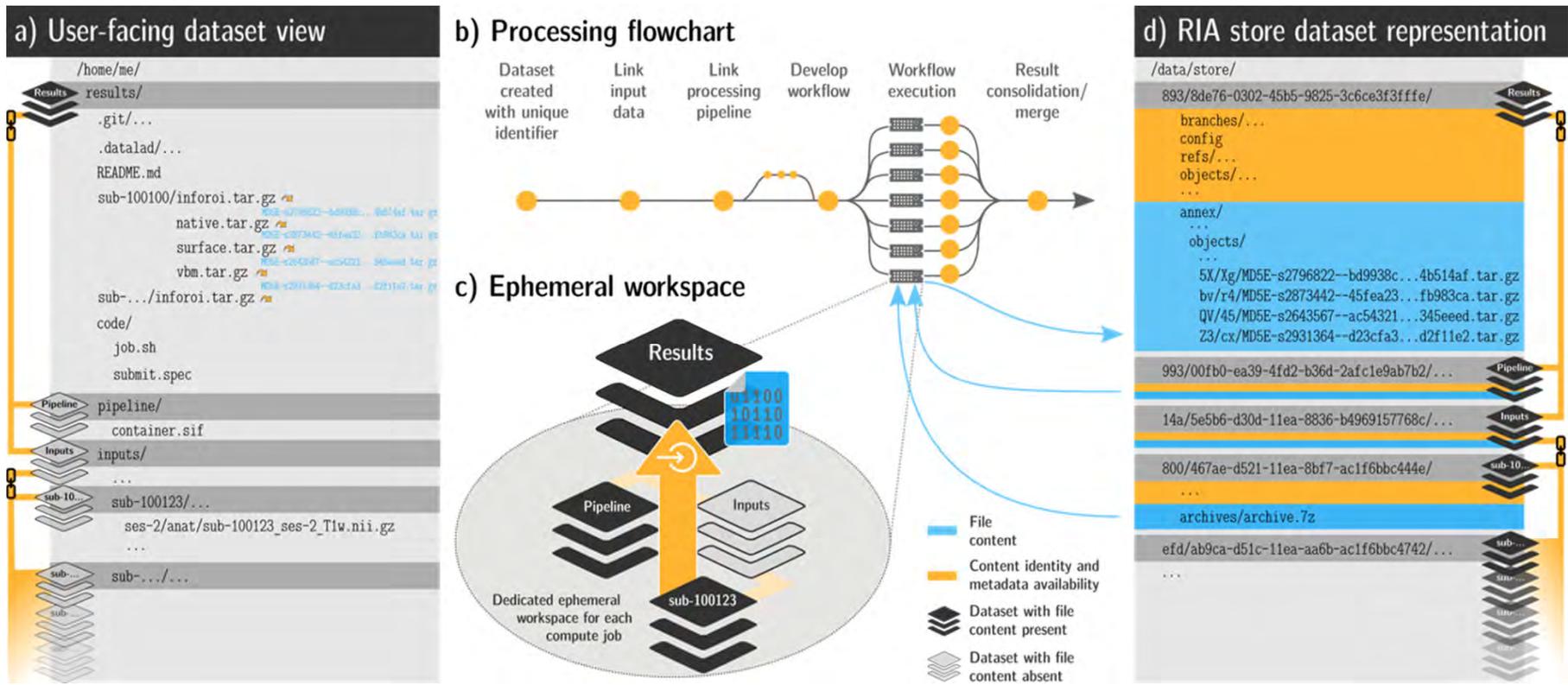
[ATALAD RUNCMD] Compute sub-6025043/ses-2

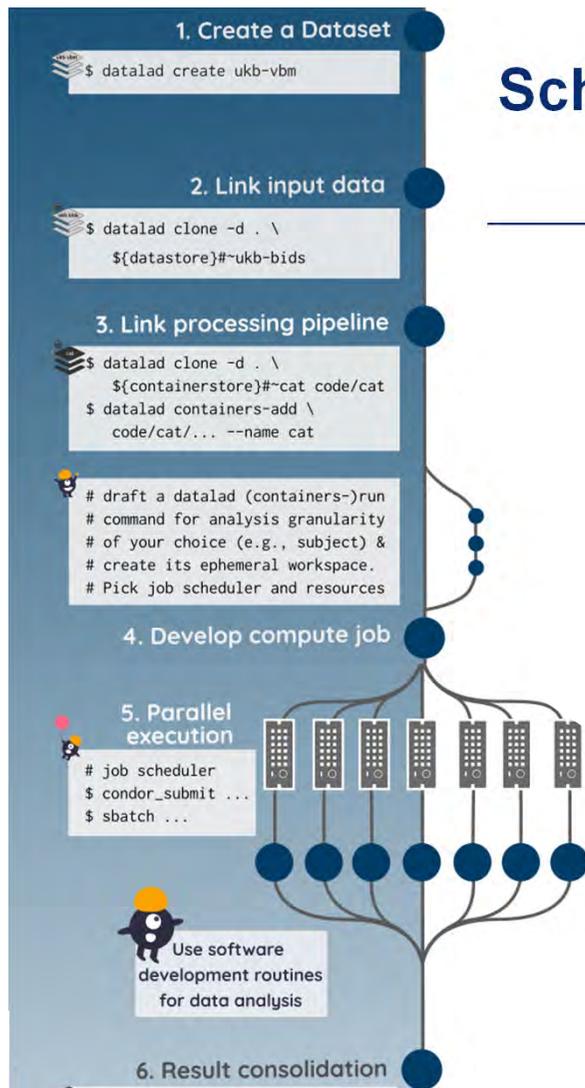
=== Do not change lines below ===
{
  "chain": [],
  "cmd": "singularity exec -B {pwd} --cleanenv code/pipeline/.datalad/
environments/cat/image sh -e -u -x -c [...]"
  "dsid": "8938de76-0302-45b5-9825-3c6ce3f3fffe",
  "exit": 0,
  "extra_inputs": [
    "code/pipeline/.datalad/environments/cat/image"
  ],
  "inputs": [
    "inputs/ukb/sub-6025043/ses-2/anat/sub-6025043_ses-2_T1w.nii.gz",
    "code/cat_standalone_batch.txt",
    "code/finalize_job_outputs.sh"
  ],
  "outputs": [
    "sub-6025043/ses-2"
  ],
  "pwd": "."
}
^^^ Do not change lines above ^^^

---
sub-6025043/ses-2/inforoi.tar.gz | 1 +
sub-6025043/ses-2/native.tar.gz | 1 +
sub-6025043/ses-2/surface.tar.gz | 1 +
sub-6025043/ses-2/vbm.tar.gz | 1 +
4 files changed, 4 insertions(+)
```

 Basic commit metadata
Author, Agent, Date, Time, and Commit Message
Transformations
Command call/ Container parametrization
Software container image
Origin: http://containers.ds.inm7.de/.. Version: dfa6d975ea8888ed33bf714c67
Input data
Origin: http://ukb.ds.inm7.de/.../bids Version: 0c7f0b45140dde1d7291b1572
Expected output data/folder
Captured output data
Path, Content hash

Schematische Übersicht





Schritte des Frameworks

- Das Framework besteht aus 6 Schritten
- Schritt 1-4 werden durch ein Bootstrapping Skript aufgesetzt
- Nutzer bestimmen Natur der Analyse und Ausmaß an Parallelisierung selbst; Bedarf evtl. Anpassung an Limitierung der Infrastruktur (zB Anzahl gleichzeitiger Jobs)