HELMHOLTZ

SPITZENFORSCHUNG FÜR
GROSSE HERAUSFORDERUNGEN

Wissenschaft Online
“Forschungsdatenmanagement”

Adina Wagner

Psychoinformatics Lab, INM-7

Institute of Experimental Psychology, HHU Düsseldorf

ReproNim/INCF fellow

incf International Neuroinformatics Coordinating Facility

ReproNim

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)

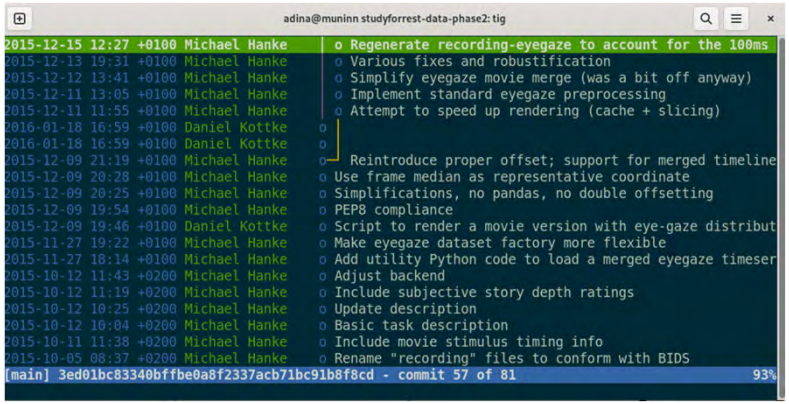
DOI 10.5281/zenodo.7155599

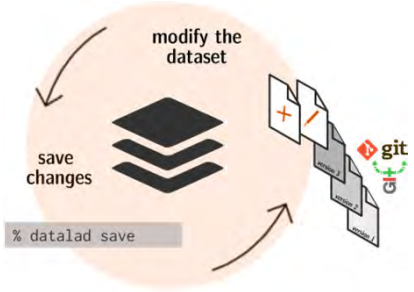
Datenmanagement im INM-7

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

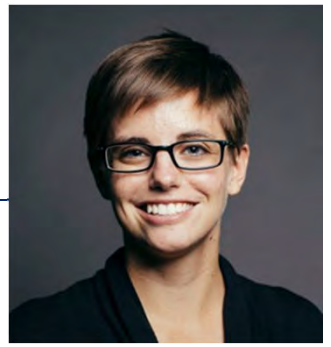




HIER einfach im Browser ausprobieren

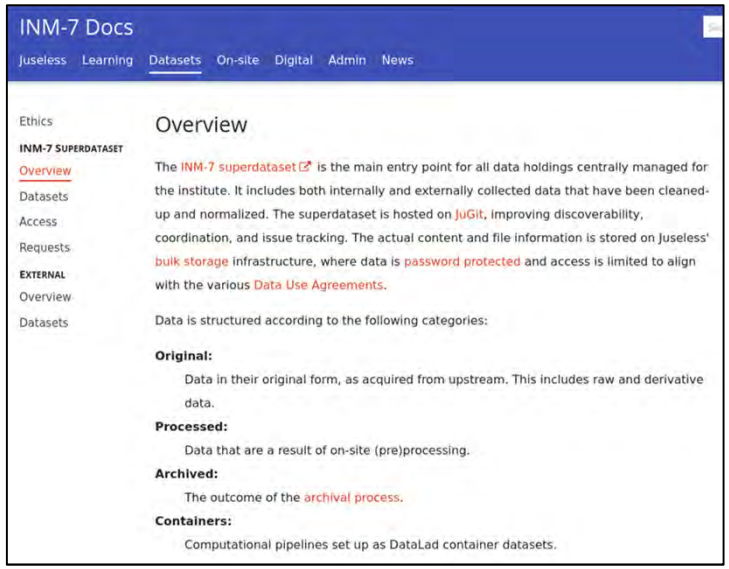
Datenmanagement im INM-7

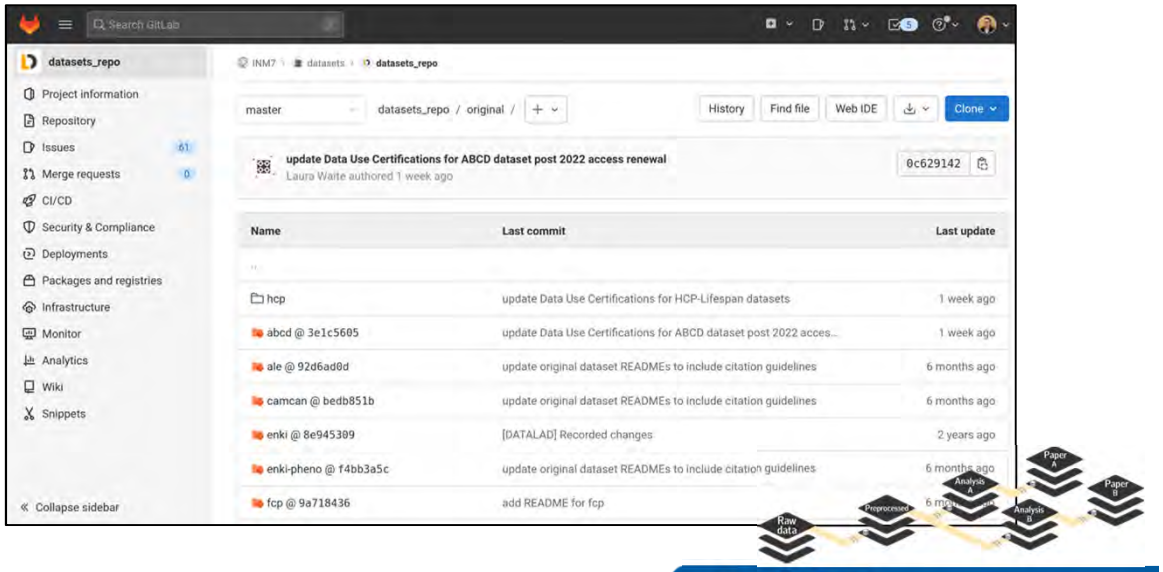
Data



Laura Waite, Data steward

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

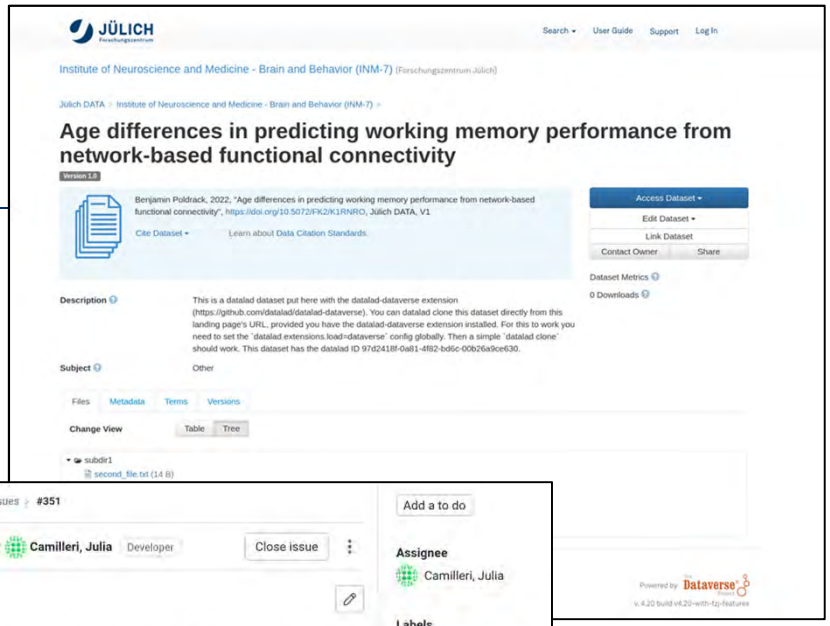


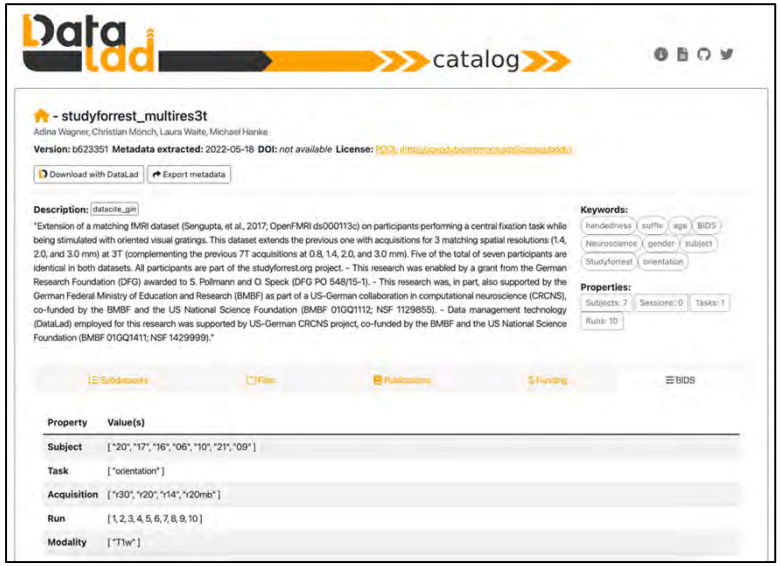


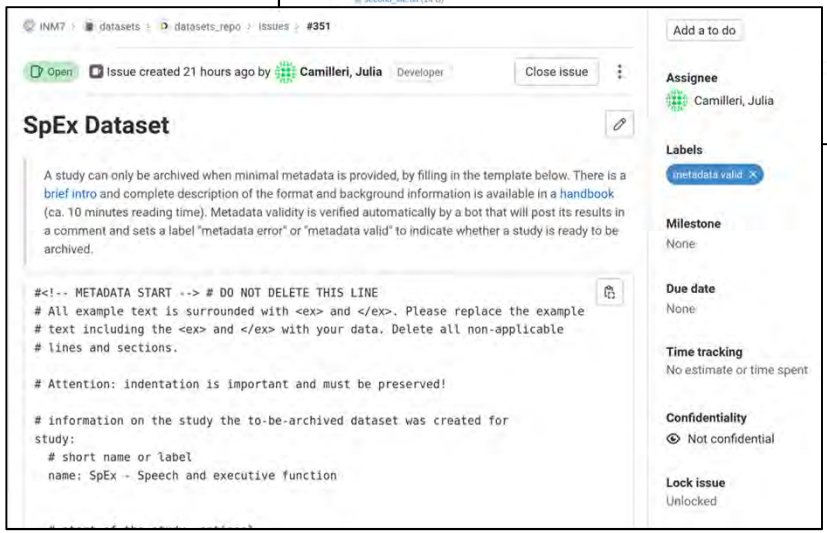
DOI 10.5281/zenodo.7155599

Datenmanagement im INM-7

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







Die Herausforderung: Datenmanagement at scale

Technische Komplexität darf Reproduzierkeit nicht verhindern

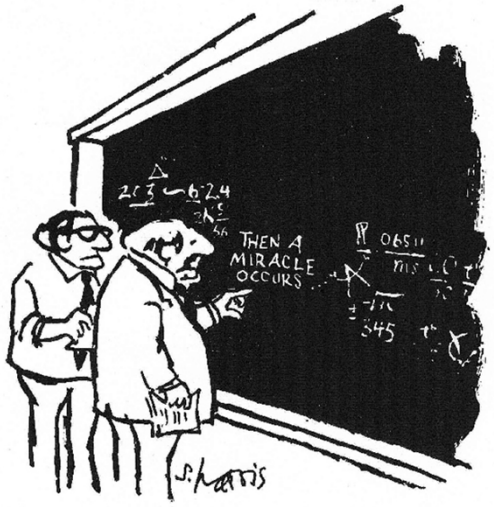
biobank uk

Improving the health of future generations

- 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

DOI 10.5281/zenodo.7155599

Die Herausforderung: Datenmanagement at scale

Article | Open Access | Published: 11 March 2022

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

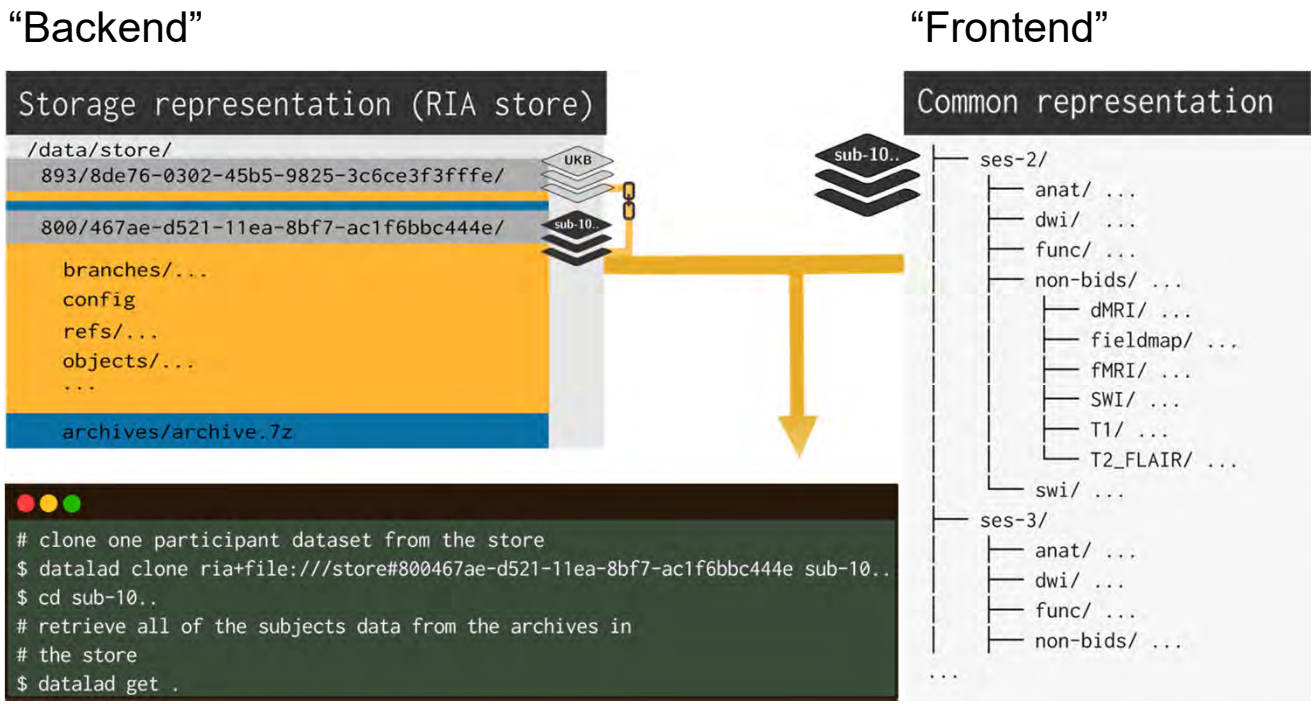
Adina S. Wagner, Laura K. Waite, Małgorzata Wierzba, Felix Hoffstaedter, Alexander Q. Waite, Benjamin Poldrack, Simon B. Eickhoff & Michael Hanke

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

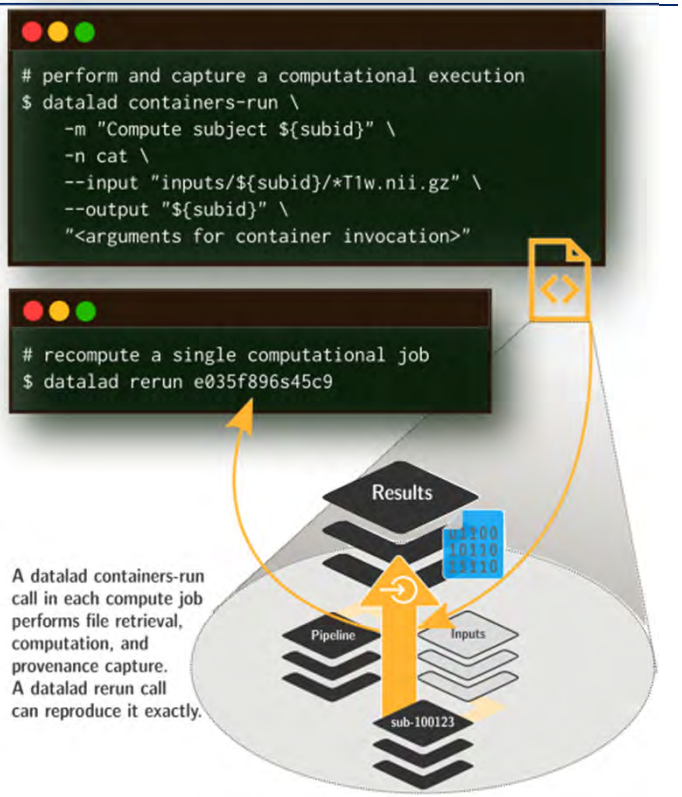
- Speicherplatz- & Dateianzahl-sparsame interne RIA Repräsentation von DataLad Datasets
- Erlaubt auch verschlüsselte Speicherung und Transport zur Erfüllung von Datenschutz-anforderungen



DOI 10.5281/zenodo.7155599

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

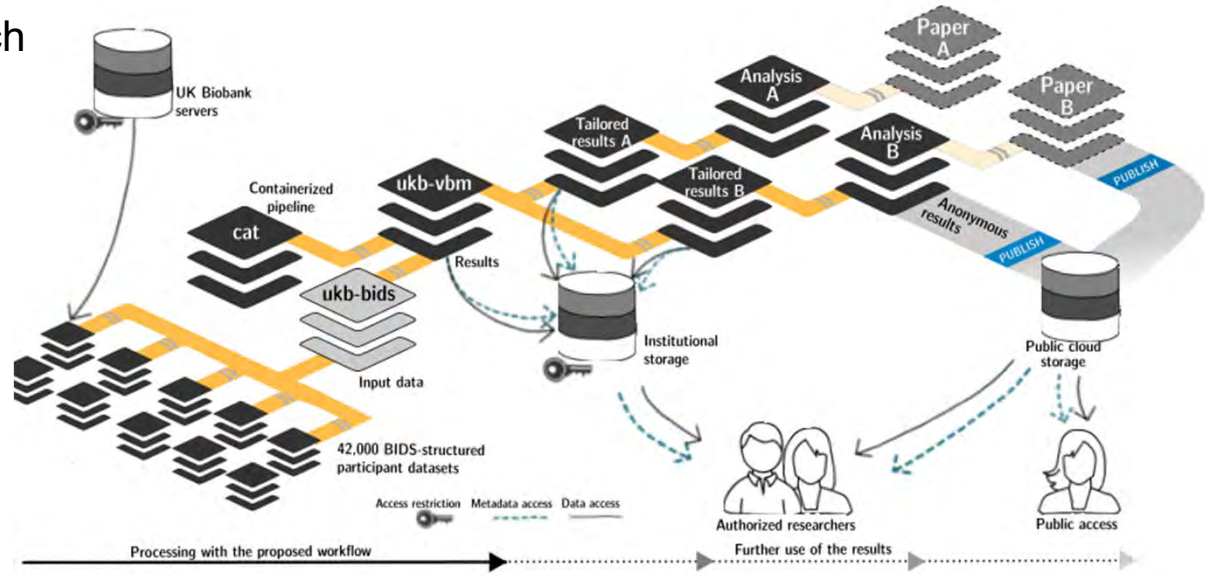


DOI 10.5281/zenodo.7155599

FAIRly big: Transparent und reproduzierbar

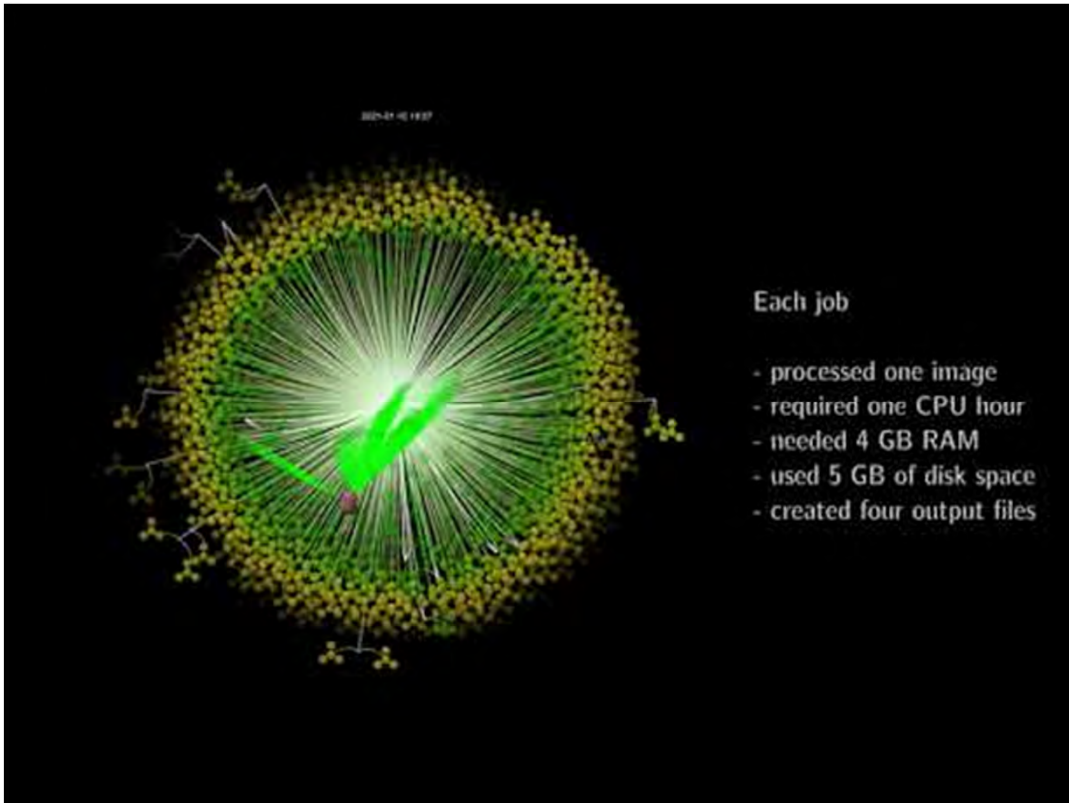
Felix Hoffstädter

- Die Ergebnisse des Frameworks machen die Prozesshistorie so **transparent** und automatisch **nachvollzierbar** wie es der Datenschutz zulässt
- Für uns und andere mittlerweile **Standard** zur internen Daten-aufbereitung oder öffentlichem Daten-austausch



DOI 10.5281/zenodo.7155599

Weiterführende Links - Vielen Dank für Ihre Aufmerksamkeit



Tutorial

Talk

Visualisierung

DOI 10.5281/zenodo.7155599

[Zusatzfolien nachfolgend]

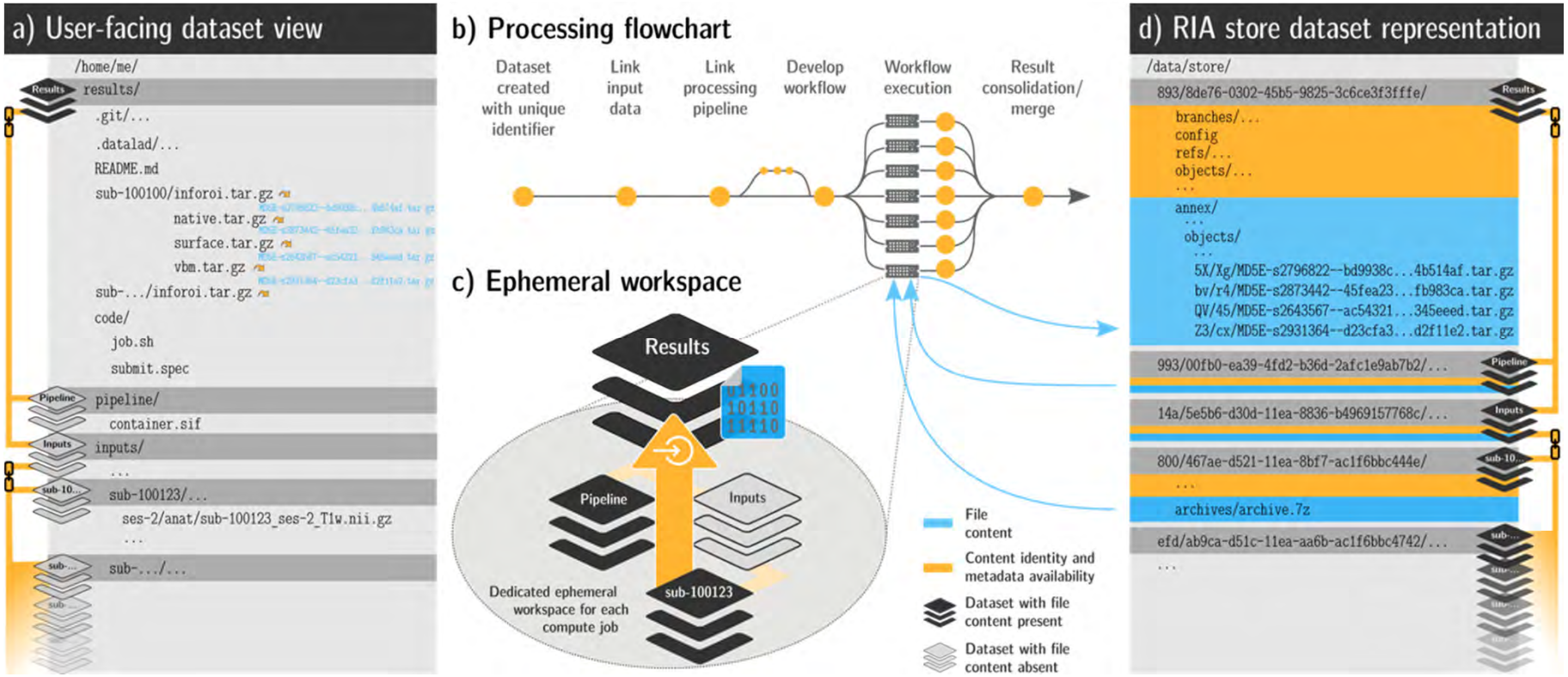
DOI 10.5281/zenodo.7155599

Provenienz

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

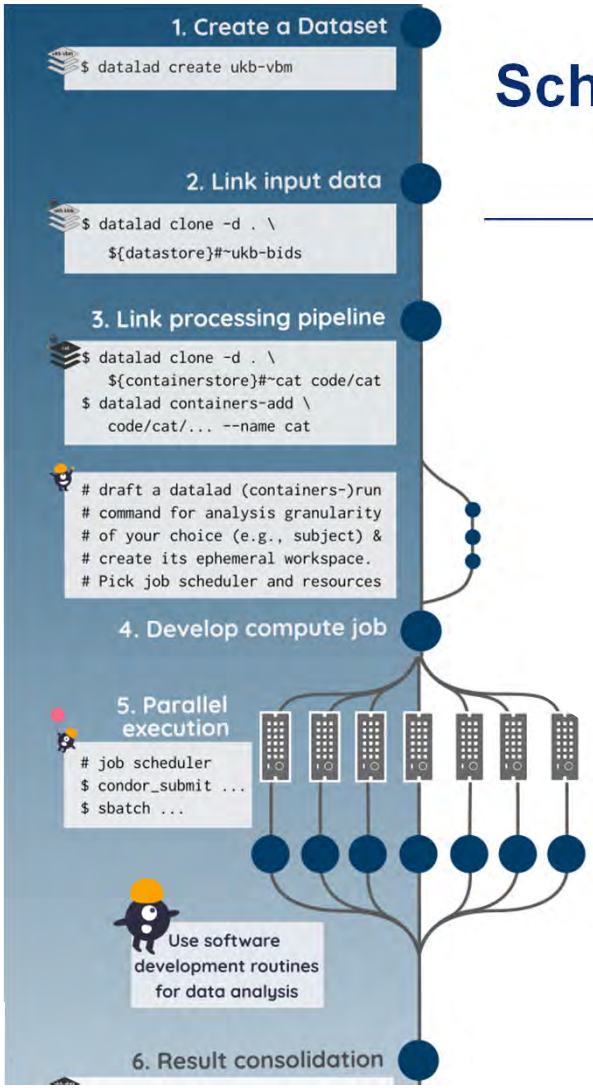
DOI 10.5281/zenodo.7155599

Schematische Übersicht



DOI 10.5281/zenodo.7155599

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)