DataCat

Heinrich Heine Universität Düsseldorf

JÜLICH Forschungszentrum

STEPHAN HEUNIS[1], CHRISTIAN MOENCH[1], BENJAMIN POLDRACK[1], & MICHAEL HANKE[1,2]

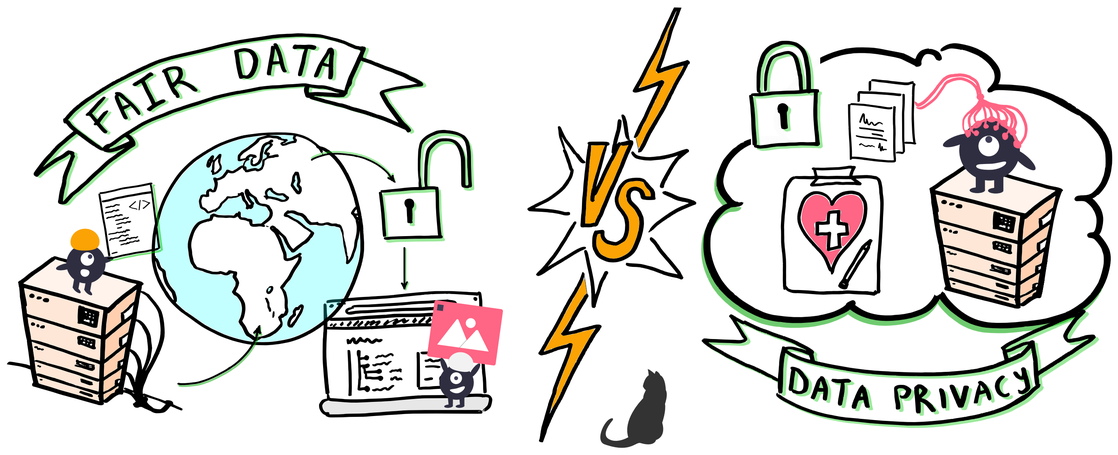
[1] INSTITUTE OF NEUROSCIENCE AND MEDICINE, BRAIN & BEHAVIOUR INM-7, RESEARCH CENTER JUELICH, GERMANY

[2] INSTITUTE OF SYSTEMS NEUROSCIENCE, MEDICAL FACULTY, HEINRICH HEINE UNIVERSITY DUESSELDORF, GERMANY

GENERATE A USER-FRIENDLY DATA BROWSER FROM STRUCTURED METADATA USING DATALAD-CATALOG

1 THE CHALLENGE

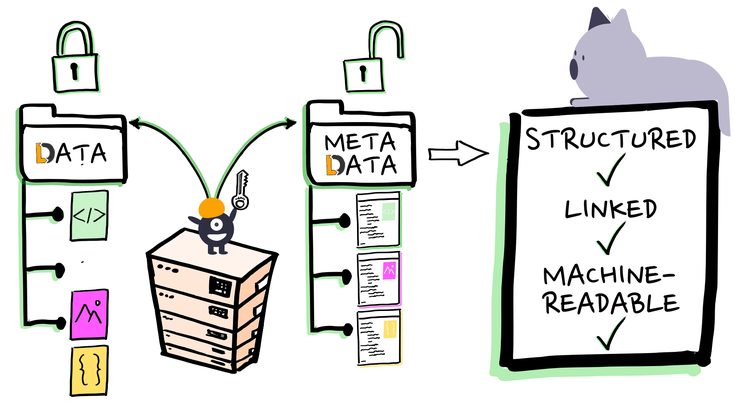
THE IMPORTANCE AND BENEFITS OF MAKING RESEARCH DATA FINDABLE, ACCESSIBLE, INTEROPERABLE, AND REUSABLE (I.E. FAIR) ARE CLEAR[1]. BUT OF EQUAL IMPORTANCE IS OUR ETHICAL AND LEGAL OBLIGATIONS TO PROTECT THE PERSONAL DATA PRIVACY OF RESEARCH PARTICIPANTS.



SO WE ARE STRUCK WITH THIS APPARENT CONTRADICTION: HOW CAN WE SHARE OUR DATA OPENLY, YET KEEP IT SECURE AND PROTECTED? SHOULD WE ERR ON THE SIDE OF FAIR DATA, OR OF DATA PRIVACY? OR HOW CAN WE BALANCE THE ASSUMED TRADEOFF BETWEEN THESE ASPETCS?

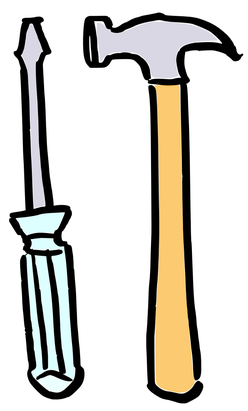
2 THE OPPORTUNITY

STRUCTURED, LINKED, AND MACHINE-READABLE METADATA PRESENTS A POWERFUL OPPORTUNITY TO ADDRESS THIS CONTRADICTION. METADATA PROVIDES BOTH HIGH-LEVEL INFORMATION ABOUT OUR RESEARCH DATA (SUCH AS STUDY DESIGN, PARTICIPANT INFORMATION, AND DATA ACQUISITION PARAMETERS) AND LOWER-LEVEL DESCRIPTIVE ASPECTS OF EACH FILE IN THE DATASET (SUCH AS FILENAMES, RELATIVE PATHS, SIZES, AND FORMATS). WITH THIS METADATA, WE CAN CREATE AN ABSTRACT REPRESENTATION OF A FULL DATASET THAT IS SEPARATE FROM THE ACTUAL DATA CONTENT. THIS MEANS THAT THE CONTENT CAN BE STORED SECURELY TO MAINTAIN PRIVACY, WHILE THE METADATA CAN BE SHARED OPENLY TO COMPLY WITH FAIR PRINCIPLES.



3 THE TOOLSET

THESE IDEALS ARE ACHIEVABLE IN PRACTICE WITH A FREE AND OPEN SOURCE TOOLSET. DATALAD[2] CAN BE USED FOR DECENTRALISED MANAGEMENT OF DATA AS LIGHTWEIGHT, PORTABLE AND EXTENSIBLE REPRESENTATIONS. DATALAD-METALAD CAN EXTRACT STRUCTURED HIGH- AND LOW-LEVEL METADATA AND ASSOCIATE IT WITH DATASETS OR WITH INDIVIDUAL FILES. AND AT THE END OF THE PIPELINE, DATALAD-CATALOG CAN TURN THE STRUCTURED METADATA INTO A USER-FRIENDLY DATA BROWSER!



DATALAD

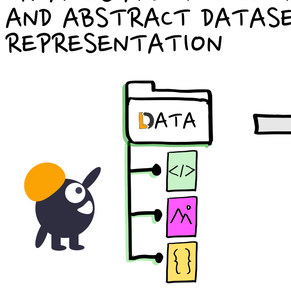
CREATES A LIGHTWEIGHT AND ABSTRACT DATASET REPRESENTATION

DATALAD METALAD

EXTRACTS AND AGGREGATES METADATA FOR DATASETS AND FILES

DATALAD CATALOG

GENERATES A DATA BROWSER FROM METADATA



4 THE CATALOG

METADATA EXTRACTED FROM VARIOUS SOURCES CAN BE AGGREGATED AND ADDED TO A CATALOG. INCOMING METADATA IS FIRST VALIDATED AGAINST A CATALOG-SPECIFIC SCHEMA BEFORE THE CATALOG ITSELF IS GENERATED OR INDIVIDUAL ENTRIES ARE ADDED. THE OUTPUT IS A SET OF STRUCTURED METADATA FILES AND A VUEJS-BASED BROWSER INTERFACE THAT HANDLES AUTOMATIC RENDERING. THIS CAN BE HOSTED AND SERVED AS STANDALONE CONTENT WITHOUT REQUIRING ANY INFRASTRUCTURE-SPECIFIC BUILDING STEPS

METADATA e.g. via DataLad

METADATA e.g. via BIDS

METADATA via ANY CUSTOM OR STANDARD ROUTE

METADATA VALIDATION + CATALOG GENERATION

{jsonschema}

```
# catalog create
# catalog validate
# catalog add
# catalog set-super
# catalog remove
# catalog serve
```

DataCat

CATALOG METADATA

BROWSER INTERFACE

vue.js

css

html

SERVE

PUSH

5 THE DECENTRALIZED & COLLABORATIVE WORKFLOW

WHILE EXISTING CATALOGING TOOLS OFTEN RELY ON CENTRALISED INFRASTRUCTURE WITH ACCESS TO THE DATA, THE PORTABILITY OF THE DATALAD-BASED TOOLSET ALLOWS DECENTRALIZED AND COLLABORATIVE CATALOG GENERATION AND MAINTENANCE AS THE OPTIMAL STANDARD. INSPIRED BY THE "FAIRLY BIG" PROCESSING WORKFLOW[3], THE FAIRLY BIG CAT WORKFLOW DEMONSTRATES HOW THE STEPS OF DATASET LINKAGE AND DISCOVERY, PARALLEL METADATA EXTRACTION AND CATALOG ENTRY GENERATION, AND ULTIMATELY CATALOG MAINTENANCE CAN BE AUTOMATED WITH THE USE OF TOOLS LIKE DATALAD AND COMMON WORKFLOW MANAGERS.

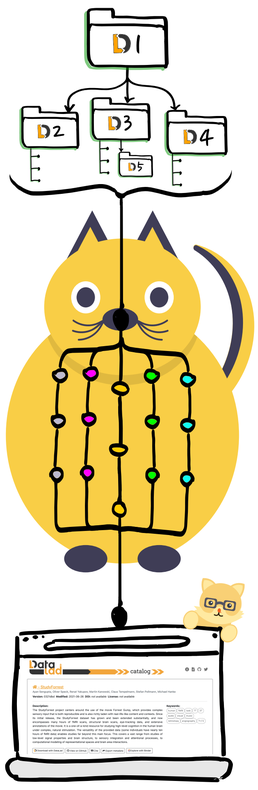
FAIRly BigCat

DATASET LINKAGE AND DISCOVERY

DECENTRALIZED & PARALLEL CATALOG ENTRY GENERATION IN SHORTLIVED WORKSPACES:

- BRANCH PER DATASET
  1. CLONE DATASET
  2. EXTRACT METADATA
  3. TRANSLATE METADATA
  4. ADD TO CATALOG
- OCTOPUS MERGE

BROWSER-BASED DATA CATALOG



TOOLS

Datalad + CUSTOM USER-SPECIFIED PROCEDURE

slurm

HTCondor

Datalad -METALAD -CATALOG + CUSTOM METADATA HANDLERS

YOUR FAVORITE BROWSER

JSH

6 TRY IT OUT!

DATALAD-CATALOG IS A FREE AND OPEN-SOURCE SOFTWARE TOOL, WITH A COMMAND LINE INTERFACE AND PYTHON API, THAT ALLOWS THE CREATION OF USER-FRIENDLY DATA BROWSERS FROM STRUCTURED METADATA. INSTALL, EXPLORE AND CONTRIBUTE VIA THESE LINKS:

INSTALL

```
# pip install
datalad-catalog
```

EXPLORE

DEMO: datalad.github.io/datalad-catalog

TUTORIAL: github.com/datalad/datalad-catalog#tutorial

CONTRIBUTE

DATALAD-CATALOG: github.com/datalad/datalad-catalog

FAIRLY-BIG-CAT: github.com/jsheunis/fairly-big-catalog-workflow

FUNDING

NSF 1912266
NSF 1229999

SPONSORED BY THE Federal Ministry of Education and Research
BMBF 01GQ1905
BMBF 01GQ1411

MOTOR SFB

H2020-EU.3.1.5.3, 945539

H2020-EU.3.1.5.3, 826421

REFERENCES

[1] WILKINSON, M. ET, C. ET AL. (2016). DOI: 10.1038/SDATA.2016.18

[2] HALCHENKO, Y.O. ET AL. (2021). DOI: 10.21105/JOSS.03262

[3] WAGNER, A.S. ET AL. (2022). DOI: 10.1038/S41597-022-01163-2

CONTACT

S.heunis@fz-juelich.de

fMRwhy jsheunis

TALK

DEMO

TUTORIAL

CODE