Data management and stewardship in environMENTAL

Providing a FAIR data infrastructure

Sven Twardziok environMENTAL seminar 3 06.07.2022



Data Management and Bioinformatics Services Digital Health Center, BIH@Charite



Ulrike TaronData manager



Marcel JentschData manager



Philipp Strubel
Software
developer



Joshua WaackApprentice

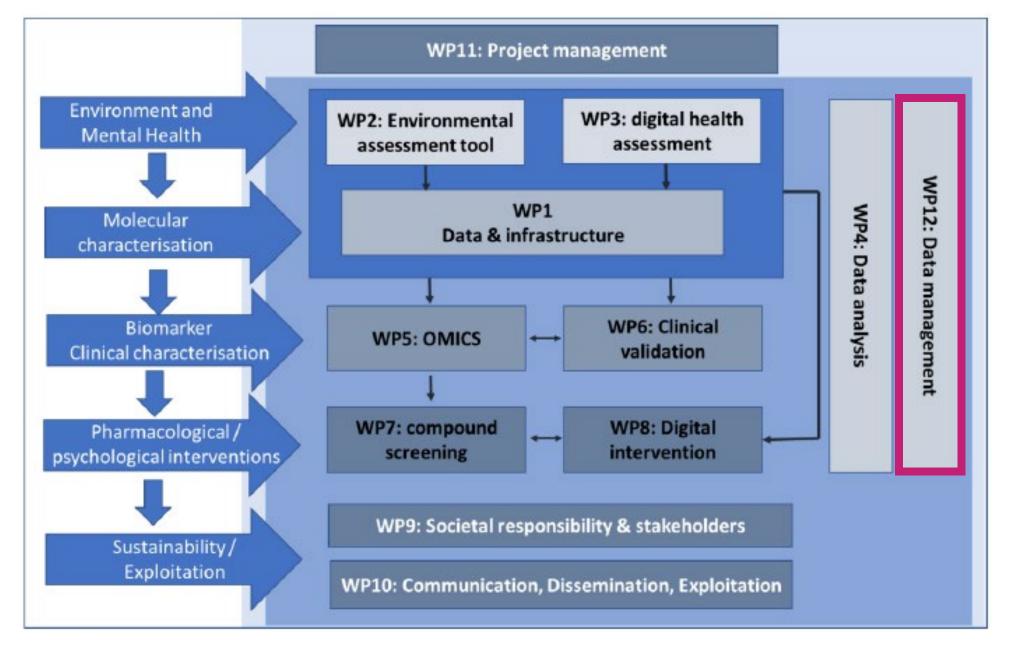


Valentin
Schneider-Lunitz
Bioinformatician



Sven Twardziok
Bioinformatician,
group leader







WP12: Data Management

Objectives

- 1. Project wide data management following the FAIR principles
- 2. Providing a data infrastructure at BIH
- 3. Distribution of samples for analysis and integration of deep phenotyping data

Deliverables:

- D12.1 Data management plan (M6)
- D12.3 Provision of data infrastructure (M12)
- • •



Data Management Plan (DMP)

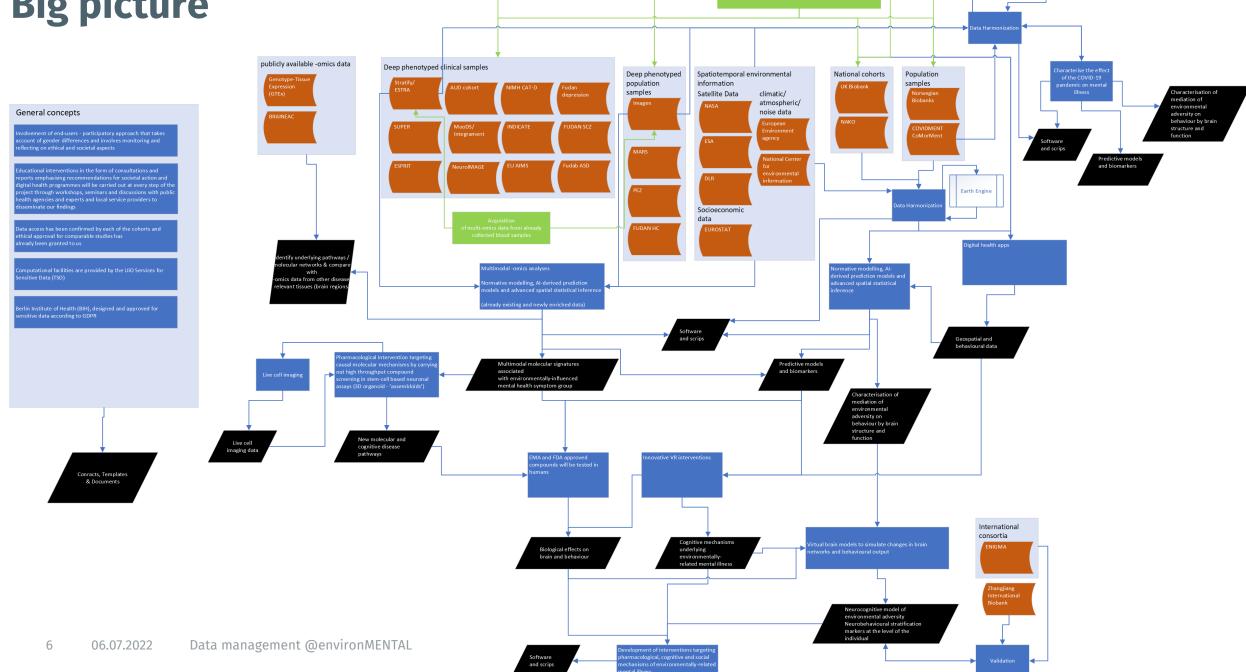
A DMP describes the data management life cycle for the data to be collected, processed and/or generated. As part of making research FAIR, a DMP should include information on¹:

- handling of research data during and after the end of the project
- what data will be collected, processed and/or generated?
- which methodology and standards will be applied
- whether data will be shared/made open access and
- how data will be curated and preserved (including after the end of the project).

¹Guidelines on FAIR Data Management in Horizon 2020



Big picture



Earth Engine

FAIR

Findable, Accessible, Interoperable, Reusable

- FAIR principles contain 15 rules to **define** good data management
- **Goal:** share data in machine readable way (aka. Fully AI-Ready)
- Focus on reuse of research data but also applicable for medicine
- FAIRness can be easily validated by selfassessment or automatic validation

SCIENTIFIC DATA

OPEN Comment: The FAIR Guiding Principles for scientific data management and stewardship

set of stakeholders—representing academia, industry, funding agencies, and scholarly publishers—have

Received: 10 December 2015 come together to design and jointly endorse a concise and measureable set of principles that we refer Accepted: 12 February 2016 to as the FAIR Data Principles. The intent is that these may act as a guideline for those wishing to enhance the reusability of their data holdings. Distinct from peer initiatives that focus on the human scholar, the FAIR Principles put specific emphasis on enhancing the ability of machines to automatically find and use the data, in addition to supporting its reuse by individuals. This Comment is the first formal publication of the FAIR Principles, and includes the rationale behind them, and some exempla

Supporting discovery through good data management

Good data management is not a goal in itself, but rather is the key conduit leading to knowledge discovery and innovation, and to subsequent data and knowledge integration and reuse by the mmunity after the data publication process. Unfortunately, the existing digital ecosystem rrounding scholarly data publication prevents us from extracting maximum benefit from our sereacric investments (e.g., ref. 1). Partially in response to this, science funders, publishers and governmental agencies are beginning to require data management and stewardship plans for data generated in publicly funded experiments. Beyond proper collection, annotation, and archival, data stewardship includes the notion of 'long-term care' of valuable digital assets, with the goal that the defined, and is generally left as a decision for the data or repository owner. Therefore, bringing slarity around the goals and desiderata of good data management and stewardship, and defining simple guideposts to inform those who publish and/or preserve scholarly data, would be of great utility

This article describes four foundational principles—Findability, Accessibility, Interoperability, and Reusability—that serve to guide data producers and publishers as they navigate around these obstacles, thereby helping to maximize the added-value gained by contemporary, formal scholarly digital publishing. Importantly, it is our intent that the principles apply not only to data' in the conventional sense, but also to the algorithms, tools, and workflows that led to that data' in the scholarly digital research objects'—from data to analytical pipelines—benefit from application of these principles, since all components of the research process must be available to ensure

There are numerous and diverse stakeholders who stand to benefit from overco There are infinerous and diverse staken localisms who stand to benefit from overcoming these obstacles researchers wanting to share, get credit, and reuse each other's data and interpretations; professional data publishers offering their services; software and tool-builders providing data analysis and processing services such as reusable workflows; funding agencies (private and public) increasingly

spondence and requests for materials should be addressed to B.M. (email: barend.mons@dtls.nl). #A full list of authors and their affiliations appears at the end of the paper

SCIENTIFIC DATA | 3:160018 | DOI: 10.1038/sdata.2016.18

Wilkinson, M. D. et al. The FAIR Guiding Principles for scientific data management and stewardship. Sci. Data 3:160018 doi: 10.1038/sdata.2016.18 (2016).



FAIR

Findable, Accessible, Interoperable, Reusable



Findable: store metadata in public catalogues and use persistent identifiers for data/metadata



Interoperable: use common file formats and apply standard vocabularies and ontologies



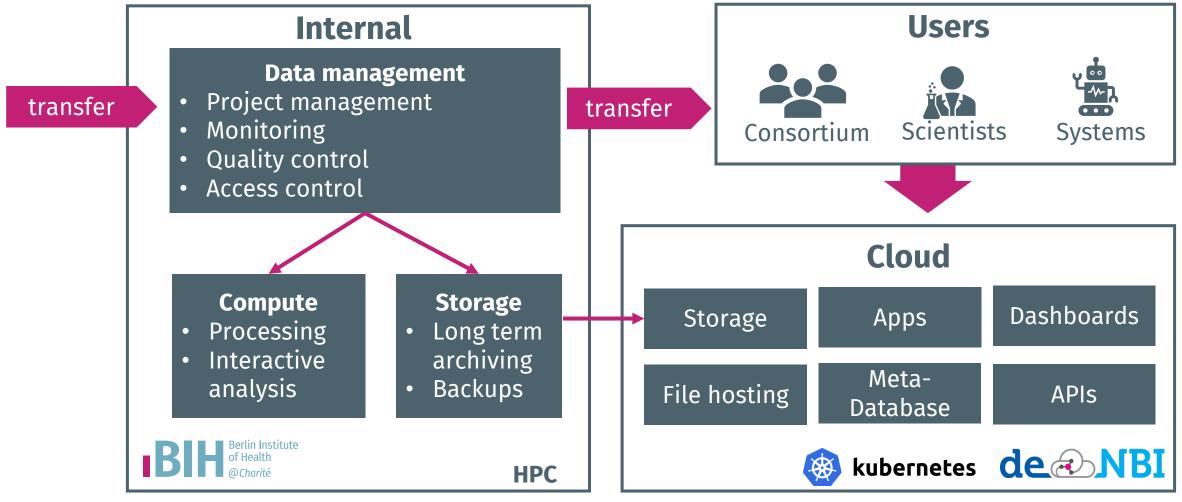
Accessible: store data in long term storage and make data accessible (controlled)



Reusable: provide rich information about data creation and data usage information



Data Infrastructure @BIH FAIR data infrastructure





Meta Database Findable & Reusable

Meta-database

environMENTAL catalog:

- Internal & external data
- Tools
- Software & scripts
- SOPs
- Documentation
- Standards



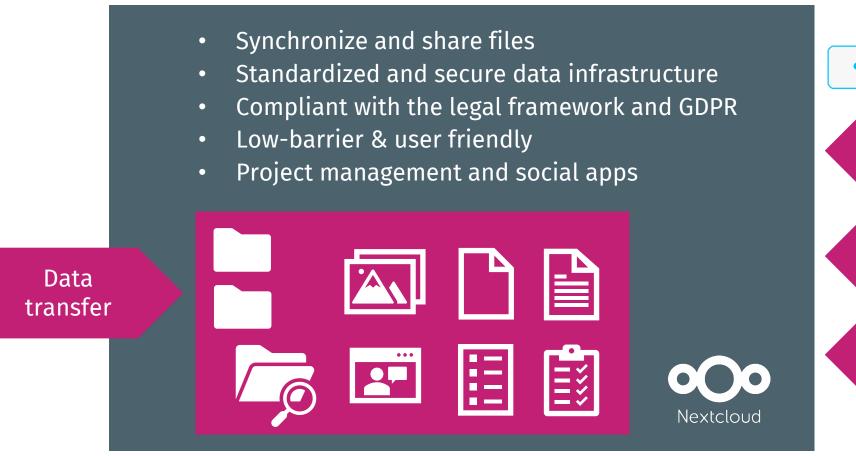
Search text

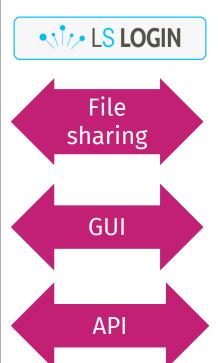
- Digital Object Identifier
- Description
- Access rules
- Location & links
- Usage terms
- Licenses
- APIs
- Publications



PsyShare: File Hosting Service

Accessible & Reusable







BIH Cloud Infrastructure



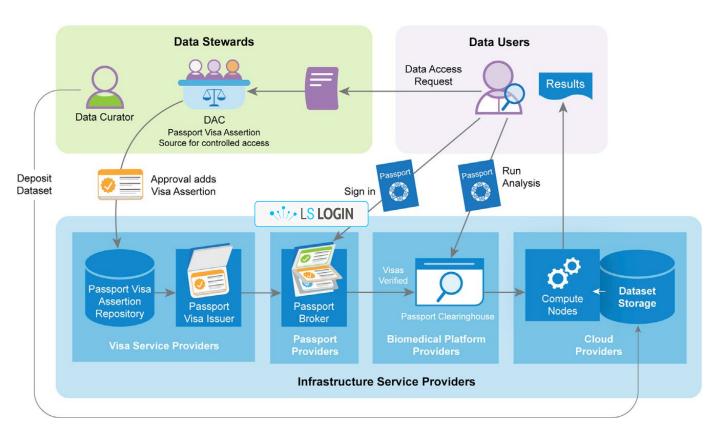


Federated Access Contol Accessible & Reusable



Life-Science AAI + GA4GH passports: manage users and access in a federated framework

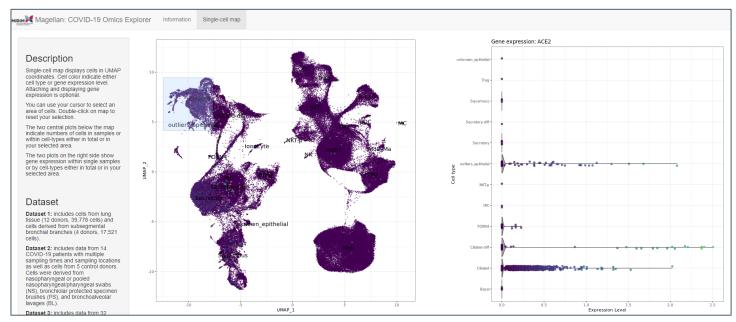
- Federated login system developed by EOSC-life and ELIXIR
- GA4GH Passports define standard for setting roles and access rights



Craig Voisin, et. Al., GA4GH Passport standard for digital identity and access permissions, Cell Genomics, Volume 1, Issue 2, https://doi.org/10.1016/j.xgen.2021.100030

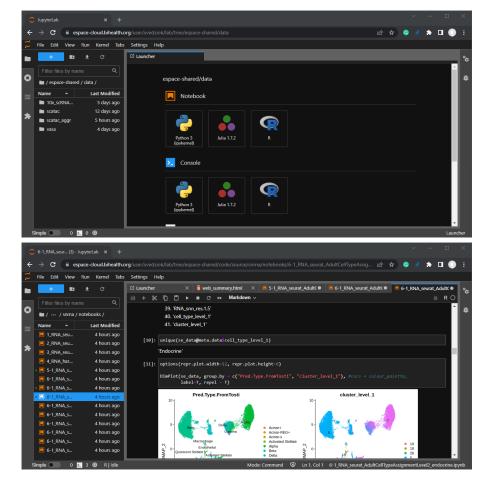


Cloud Workspaces Interoperable



https://digital.bihealth.org/

Chua, R.L., Lukassen, S., Trump, S. *et al.* COVID-19 severity correlates with airway epithelium–immune cell interactions identified by single-cell analysis. *Nat Biotechnol* **38,** 970–979 (2020). https://doi.org/10.1038/s41587-020-0602-4



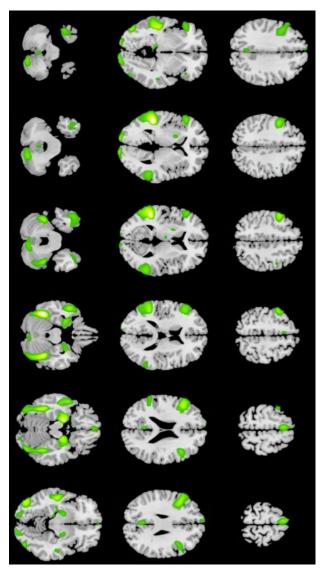
https://espace-cloud.bihealth.org

Horizon 2020 project for "Expression and Spatial analysis Pancreas Atlas Consortium Europe", ESPACE



IMAGEN, STRATIFY and ESTRA @BIH





European research projects

Longitudinal neurobiological basic research

Influence of biological, psychological, and environmental factors during adolescence on brain development and mental health

IMAGEN: over 2000 adolescents and their parents

STRATIFY: over 400 young adults with major depression, schizophrenia, alcohol use disorders and eating disorders

ESTRA: eating disorders and comorbid mental illnesses

Demographics, neuropsychological assessments, medical questionnaires, MR neuroimaging and genomics

Data collected over a 10 year period

Covid-19 Follow-Up (IMACOV/STRATICO)

Image source: nature.com/articles/s41562-020-0846-5



Summary

- Goal: manage all data according to FAIR principles
- Open data management plan
- Infrastructure at BIH:
 - IMAGEN, STRATIFY&ESTRA cohorts
 - Long term data archiving
 - PsyShare: file hosting service
 - Cloud workspaces (Jupyter, Apps, APIs)
- Connection to external infrastructures (e.g. EBRAINS)



Vielen Dank

Sven TwardziokGroup leader at the Digital Health Center

Berlin Institute of Health (BIH)

Charité - Universitätsmedizin Berlin Berlin Institute of Health (BIH) Digital Health Center | Health Data Kapelle-Ufer-2 | 10117 Berlin | Germany Phone: +49 30 450 543105

www.bihealth.org





EBRAINS European shared digital research infrastructure

- 1. Open to integrating brain research
- 2. Developed and powered by the EU-funded Human Brain Project
- 3. Provides digital tools and services
 - Data and Knowledge
 - Atlases
 - Simulations
 - other

