

Datenmanagement

Robert Haase



GEFÖRDERT VOM



Bundesministerium
für Bildung
und Forschung

SACHSEN



Diese Maßnahme wird gefördert durch die Bundesregierung aufgrund eines Beschlusses des Deutschen Bundestages. Diese Maßnahme wird mitfinanziert durch Steuermittel auf der Grundlage des von den Abgeordneten des Sächsischen Landtags beschlossenen Haushaltes.

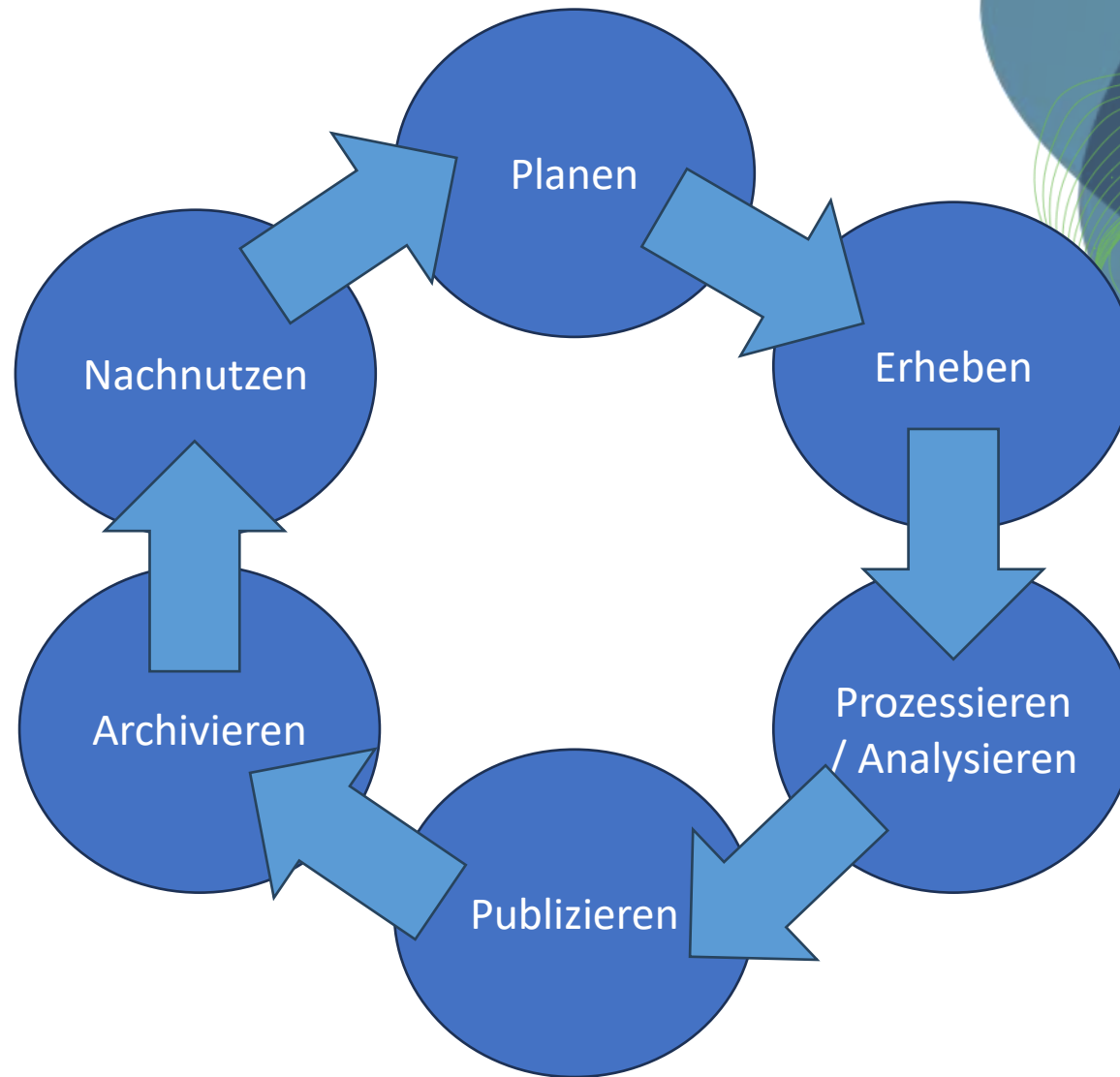
Datenmanagement

- Alle Aktivitäten, Prozesse, Begriffe, Personen die im Zusammenhang mit Daten stehen
 - Verarbeitung
 - Speicherung
 - Organisation
 - Veröffentlichung
 - ...
- Im Alltag: Der Umgang mit Daten



Lebenszyklus

- Idealerweise sind Prozesse mit Daten in einem iterativen Zyklus eingebettet



Lebenszyklus

- Kosten
- Nutzen
- Qualitätsziele
- Strategische Entscheidungen



Datenmanagement

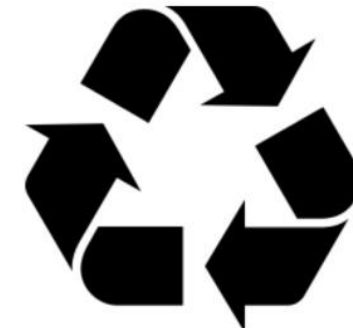
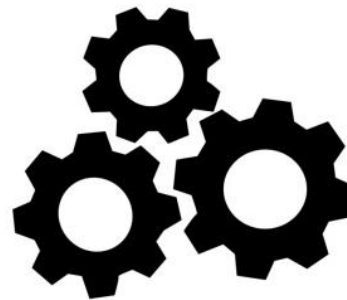
- Oft genanntes strategisches Ziel:

F
indable

A
ccessible

I
nteroperable

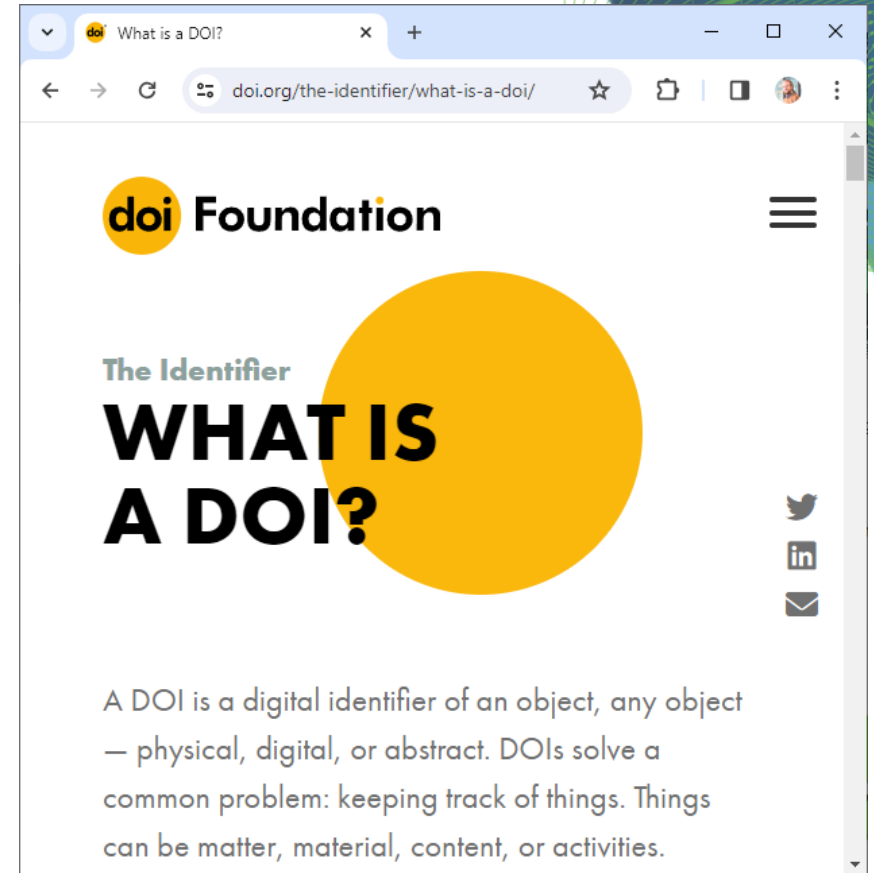
R
eusable



Die FAIR-Prinzipien

Findable

- F1. (Meta)daten sind verbunden mit einem global eindeutigem Identifier
 - Universal Resource Identifier (URI)
 - Digital Object Identifier (DOI)
- F2. Daten sind mit “reichen Metadaten” beschrieben
- F3. Metadaten beinhalten die DOI, die sie beschreiben
- F4. (Meta)daten sind in einer durchsuchbaren Resource registriert



Universal Resource Identifiers

- Welche dieser Links sind *URIs*?

<https://twitter.com/haesleinhuepf/status/891596662782779392>

<https://doi.org/10.5281/zenodo.28325>

<https://opendata.leipzig.de/dataset/vornamenstatistik-2023>

<https://www.leipzig.de/>

Digital Object Identifier

- Welche dieser Links sind Digital Object Identifier?

<https://twitter.com/haesleinhuepf/status/891596662782779392>

<https://doi.org/10.5281/zenodo.28325>

<https://opendata.leipzig.de/dataset/vornamenstatistik-2023>

<https://www.leipzig.de/>

Resource Identifiers

- Unique Identifier zeigen immer auf die gleichen Daten

Straßennetz, Stadt Leipzig

Das Straßennetz der Stadt Leipzig. Der Datensatz beinhaltet die Straßennamen, Von- und Bis-Abschnitte, Stadtbezirke der Abschnitte, die Zugehörigkeit zum klassifizierten Netz (Bundes-, Staats-, Kreisstraße), die Straßensart (Ortsstraße, Feldweg usw.), die Erschließung (Haupt-, Neben-, Anliegernetz), den Bausträger und die Länge in Meter. Die Daten werden direkt aus dem System abgerufen und sind daher stets aktuell.

Daten und Ressourcen

- Das Straßennetz im CSV-Format
- Das Straßennetz im GeoJSON-Format
- Das Straßennetz im GeoPackage-Format
- WFS-GetCapabilities

Zusätzliche Informationen

Feld	Wert
Ansprechpartner	Verkehrs- und Tiefbauamt, Stadt Leipzig
E-Mail	vta@leipzig.de
Verwaltungsebene	kommunale Ebene
Gemeindename	Leipzig, Stadt
Ausgestellt	2021-08-20
Aktualisiert	2024-01-17

Gemeindename	Leipzig, Stadt
Ausgestellt	2021-08-20
Aktualisiert	2024-01-17

Das ist also kein Unique Identifier

Resource Identifiers

opendata.leipzig.de/dataset/strassennetz-stadt-leipzig/resource/28e48a6e-9ae0-49c2-a86a-fa85b47cf4c0

Stadt Leipzig

Datensätze Organisationen Kategorien Anwendungen Über uns

Organisationen / Verkehrs- und Tiefbauamt / Straßennetz, Stadt Leipzig / Das Straßennetz im GeoJSON-Format

Das Straßennetz im GeoJSON-Format

Herunterladen

URL: <https://geodienste.leipzig.de/l3/OpenData/wfs?REQUEST=getFeature&typeName=Open...>

Das Straßennetz im GeoJSON-Format

Text Map viewer GeoJSON

Vollbildschirm Einbettung

```
{"type": "FeatureCollection", "features": [{"type": "Feature", "id": "Strassen_Segmente.372216", "geometry": {"type": "Multil
```

Technisch gesehen ist das kein URI

Möglicherweise ist er aber nicht in einer globalen Resource registriert.

Findbarkeit

- Unser zukunftsgerichtetes Selbst wird uns danken.

Kannst Du Dich noch an deinen Vortrag in 2021 erinnern?

Wo sind die Folien?

Online,
open access!

Sharing and licensing material | x +

f1000research.com/slides/10-519

F1000Research

Search

SUBMIT YOUR RESEARCH

BROWSE GATEWAYS & COLLECTIONS HOW TO PUBLISH ABOUT BLOG MY RESEARCH SIGN IN

Home » Browse » Sharing and licensing material

SLIDES

NOT PEER REVIEWED

VIEW FULL SCREEN

PowerPoint P... 1 / 28 24%

Code Slides Text Data ...

Sharing and licensing material
Robert Haase
June 30th 2021

This material is licensed by Robert Haase, PoL Dresden under the CC-BY 4.0 license <https://creativecommons.org/licenses/by/4.0/>

TECHNISCHE UNIVERSITÄT DRESDEN

Metrics | 411 Views | 60 Downloads

DOWNLOAD 30.92 MB

SHARE CITE

PART OF THE GATEWAY

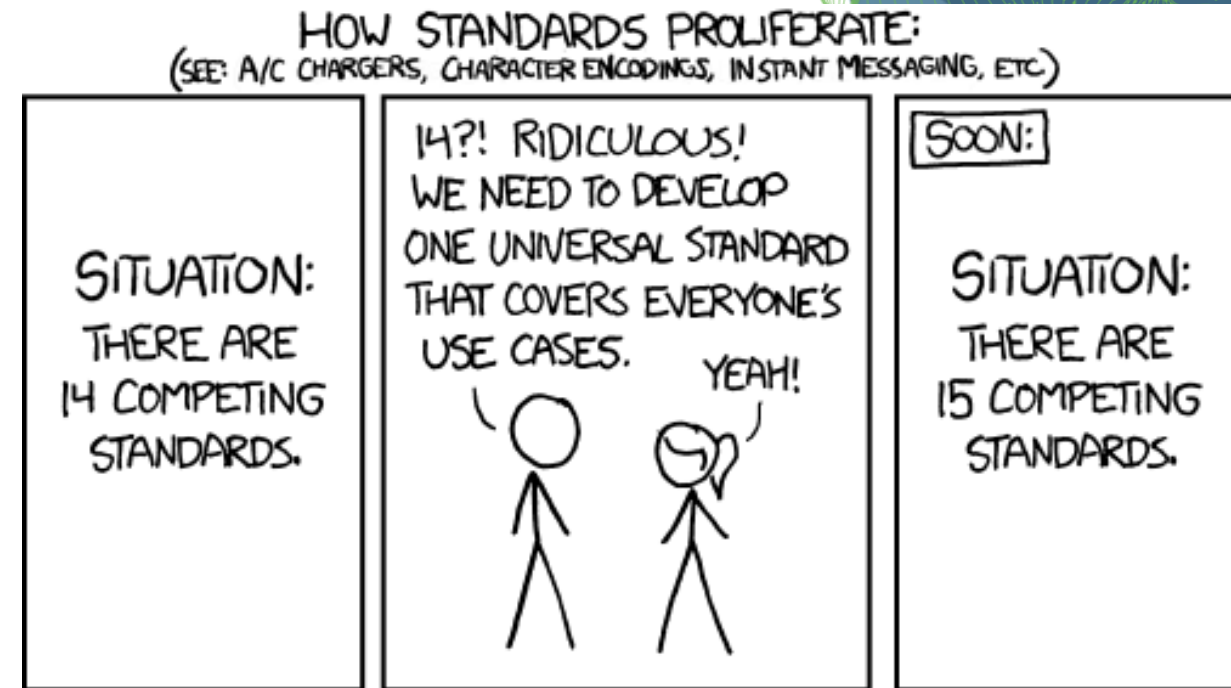
neubias NEUBIAS - the Bioimage Analysts Network

BROWSE BY RELATED SUBJECTS

Artificial intelligence
Computer and information sciences
Electrical engineering

Die FAIR-Prinzipien

- Accessible / Verfügbar / Zugreifbar
- A1. (Meta)daten können über ein Standardisiertes Protokoll empfangen werden
 - A1.1 Das Protokoll ist offen, frei und universell implementierbar
 - A1.2 Das Protokoll erlaubt Authentifizierung und Autorisierung, wenn erforderlich
- A2. Metadaten sind verfügbar, auch wenn die Daten selbst nicht mehr verfügbar sind



Accessibility

- Beispiel: Geodaten
 - GeoJSON

The image shows two browser windows. The top window is the OpenData portal for Leipzig, displaying a list of datasets. The 'Straßennetz, Stadt Leipzig' dataset is selected, and the 'GeoJSON' format is highlighted with a blue arrow. The bottom window shows the 'Pretty-print' view of the GeoJSON data for a specific street segment, 'Kändlerstraße'.

```
{ "type": "FeatureCollection", "features": [ { "type": "Feature", "id": "Strassen_Segmente.372216", "geometry": { "type": "MultiLineString", "coordinates": [ [ [ [ 309616.4829, 5688099.341 ], [ 309750.8633, 5688129.5935 ] ] ] }, "geometry_name": "geom", "properties": { "objectid": 372216, "str": "Kändlerstraße", "str_nr": "06242", "abs_nr17": "06242062420622400", "abs_nr": 4, "von_str": "Kändlerstraße", "von_str_nr": "06242", "bis_str": "Krakauer Straße", "bis_str_nr": "06224", "stadtbez": "W", "klass_netz": null, "kat_sstrgs": "0", "sstrgsname": "Ortsstraße", "kat_ersch1": 3, "erschlname": "D3", "baulast": "Stadt Leipzig", "lance": "138", "entwurf": null, "entwurf_text": null, "freigegeben": null } } ] }
```


Restricted Access

- Das A in FAIR steht nicht zwingend fuer Open-Access

blobs.tif

Published March 18, 2024 | Version v1

Dataset **Restricted**

0 VIEWS 0 DOWNLOADS

Haase, Robert^{1,2}

This dataset contains blobs.tif, which was published before as blobs.gif as part of ImageJ's example images. The dataset is public-domain, available online in png format as well: <https://samples.fiji.sc/blobs.png>

This record in Zenodo serves demonstrating that data can be published with closed access.

Files

Restricted

The record is publicly accessible, but files are restricted to users with access.

Citations

Show Literature (0) Dataset (0) Software (0)

Search for citation ... Search

blobs.tif

Published March 18, 2024 | Version v1

Dataset **Restricted**

0 VIEWS 0 DOWNLOADS

Haase, Robert^{1,2}

This dataset contains blobs.tif, which was published before as blobs.gif as part of ImageJ's example images. The dataset is public-domain, available online in png format as well: <https://samples.fiji.sc/blobs.png>

This record in Zenodo serves demonstrating that data can be published with closed access.

Files

blobs.tif

Version v1

10.5281/zenodo.10829230

Mar 18, 2024

Cite all versions? You can cite all versions by using the DOI 10.5281/zenodo.10829229. This DOI represents all versions, and will always resolve to the latest one. [Read more.](#)

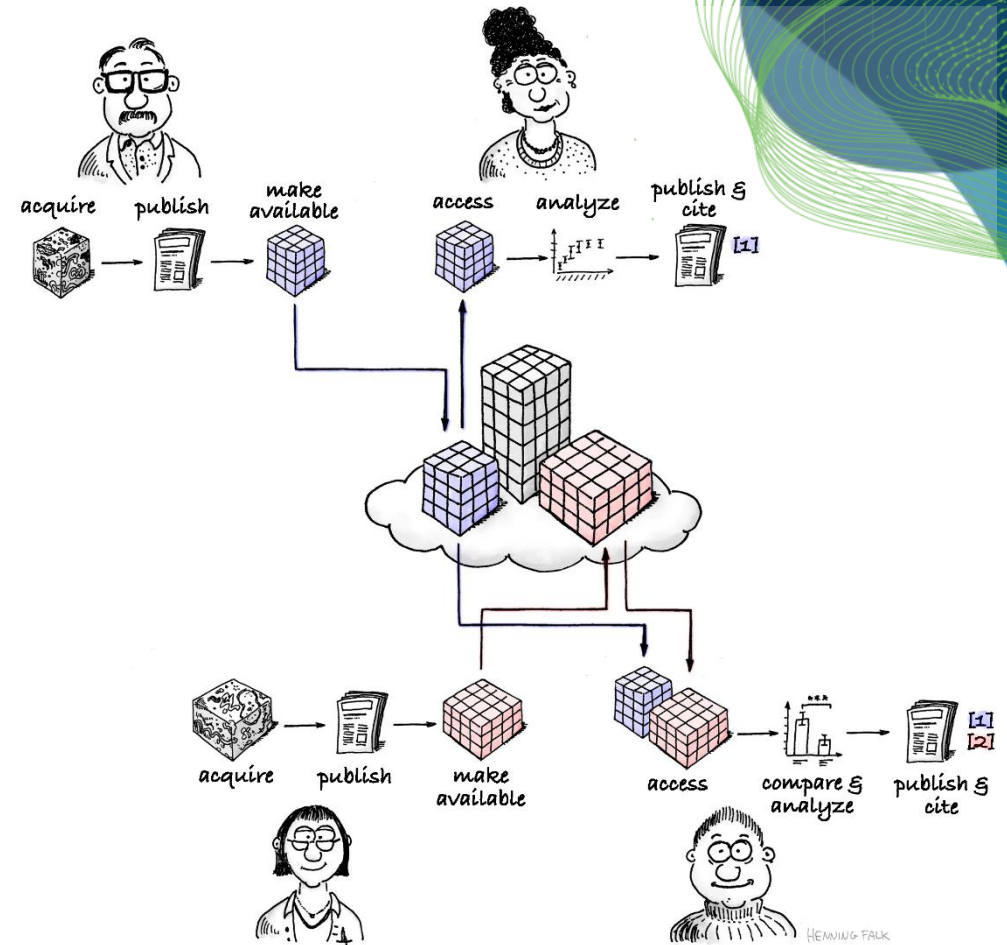
Die FAIR-Prinzipien

- Interoperable
 - I1. (Meta)daten sind formalisiert in einer zugänglichen, gemeinsamen, breit angewandten Sprache, geeignet für Wissensrepräsentation
 - I2. (Meta)daten nutzen ein Vokabular, das ebenfalls den FAIR-Prinzipien unterliegt
 - I3. (Meta)daten referenzieren andere qualifizierte (Meta)daten



Die FAIR-Prinzipien

- Reusable / Wiederverwendbar
- R1. (Meta)daten sind reich an vielfaeltigen, akkuraten und relevanten Attributen
- R1.1. (Meta)daten werden mit einer klaren und verfuegbaren Nutzungslizenz versehen
- R1.2. (Meta)daten sind stets mit der detaillierten Herkunft erfasst
- R1.3. (Meta)daten folgen gemeinschaftlich definierten Standards



Lizensierung

Willkommen - Open Data-Portal de x +
opendata.leipzig.de

Stadt Leipzig

Datensätze Organisationen Kategorien Anwendungen Über uns Nutzung

OPEN DATA-PORTAL DER STADT LEIPZIG

Datensatz suchen, z.B. Umwelt

266 Datensätze 32 Organisationen 13 Gruppen 9 Anwendungen

Nutzung - Open Data-Portal de x +
opendata.leipzig.de/pages/usage

Stadt Leipzig

Datensätze Organisationen Kategorien Anwendungen Über uns Nutzung Hackathons

Nutzung

Mit dem Zugriff auf die Daten dieser Internetseiten und Plattform stimmen Sie den Nutzungsbedingungen zu.

Die Stadt Leipzig veröffentlicht Daten für die weitere Nutzung durch Bürgerinnen und Bürger, die Wirtschaft, die Medien, die Wissenschaft und sonstige Institutionen.

Für die Datensätze gilt, soweit nicht anders gekennzeichnet, die Lizenz: [Datenlizenz Deutschland](#) in der aktuellen Fassung mit folgenden Bedingungen:

- Die Stadt Leipzig möchte über neue Anwendungen und Services informieren, die auf den Daten dieser Plattform verwendet werden. Die Stadt Leipzig ist daher berechtigt, Informationen über solche Anwendungen und Services zu veröffentlichen und für eine Berichterstattung zu verwenden. Der Service auf opendata.leipzig.de gestattet die Entwicklung von Services auf opendata.leipzig.de gestatteten Leistungen und Daten ausdrücken.

by-2-0 - GovData
govdata.de/dl-de/by-2-0

Startseite Anmelden / Registrieren FAQ Kontakt

GOV DATA
Das Datenportal für Deutschland

Daten Showroom SPARQL Informationen Blog

DL-DE->BY-2.0

Datenlizenz Deutschland – Namensnennung – Version 2.0

(1) Jede Nutzung ist unter den Bedingungen dieser „Datenlizenz Deutschland – Namensnennung – Version 2.0“ zulässig.

Die bereitgestellten Daten und Metadaten dürfen für die kommerzielle und nicht kommerzielle Nutzung insbesondere

1. vervielfältigt, ausgedruckt, präsentiert, verändert, bearbeitet sowie an Dritte übermittelt werden;
2. mit eigenen Daten und Daten Anderer zusammengeführt und zu selbständigen neuen Datensätzen verbunden werden;
3. in interne und externe Geschäftsprozesse, Produkte und Anwendungen in öffentlichen und nicht öffentlichen elektronischen Netzwerken eingebunden werden.

(2) Bei der Nutzung ist sicherzustellen, dass folgende Angaben als Quellenvermerk enthalten sind:

1. Bezeichnung des Bereitstellers nach dessen Maßgabe,
2. der Vermerk „Datenlizenz Deutschland – Namensnennung – Version 2.0“ oder „dl-de/by-2-0“ mit Verweis auf den Lizenztext

Lizensierung

Datenlizenz Deutschland – Namensnennung – Version 2.0

- (1) Jede Nutzung ist unter den Bedingungen dieser „Datenlizenz Deutschland – Namensnennung – Version 2.0“ zulässig.
- Die bereitgestellten Daten und Metadaten dürfen für die kommerzielle und nicht kommerzielle Nutzung insbesondere
 1. vervielfältigt, ausgedruckt, präsentiert, verändert, bearbeitet sowie an Dritte übermittelt werden;
 2. mit eigenen Daten und Daten Anderer zusammengeführt und zu selbständigen neuen Datensätzen verbunden werden;
 3. in interne und externe Geschäftsprozesse, Produkte und Anwendungen in öffentlichen und nicht öffentlichen elektronischen Netzwerken eingebunden werden.
- (2) Bei der Nutzung ist sicherzustellen, dass folgende Angaben als Quellenvermerk enthalten sind:
 1. Bezeichnung des Bereitstellers nach dessen Maßgabe,
 2. der Vermerk „Datenlizenz Deutschland – Namensnennung – Version 2.0“ oder „dl-de/by-2-0“ mit Verweis auf den Lizenztext unter www.govdata.de/dl-de/by-2-0 sowie
 3. einen Verweis auf den Datensatz (URI).
- Dies gilt nur soweit die datenhaltende Stelle die Angaben 1. bis 3. zum Quellenvermerk bereitstellt.
- (3) Veränderungen, Bearbeitungen, neue Gestaltungen oder sonstige Abwandlungen sind im Quellenvermerk mit dem Hinweis zu versehen, dass die Daten geändert wurden.

Quiz

- Datenlizenz Deutschland - Namensnennung - Version 2.0 ist besonders ähnlich zu

CC0



CC-BY



CC-BY-NC

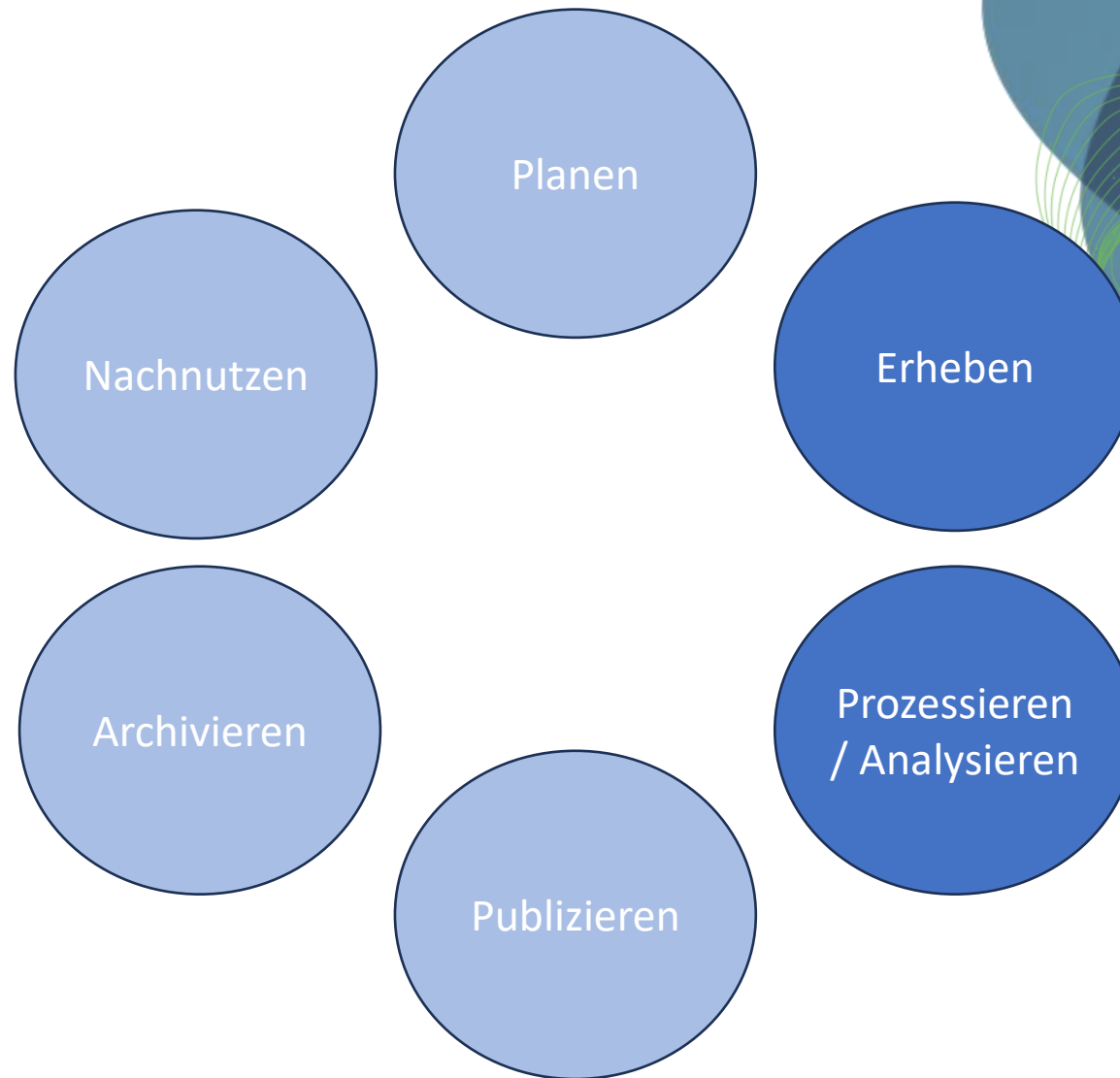


CC-BY-SA



Lebenszyklus

- Arten von Daten
- Rahmenbedingungen
 - Nutzungsrechte
- Technologie / IT Infrastruktur
- Backup



Arten von Daten

- Strukturierte Daten
 - Tabellen, Datenbanken
- Unstrukturierte Daten
 - Texte, Emails, Videos
- Semi-strukturierte Daten
 - Frageboegen



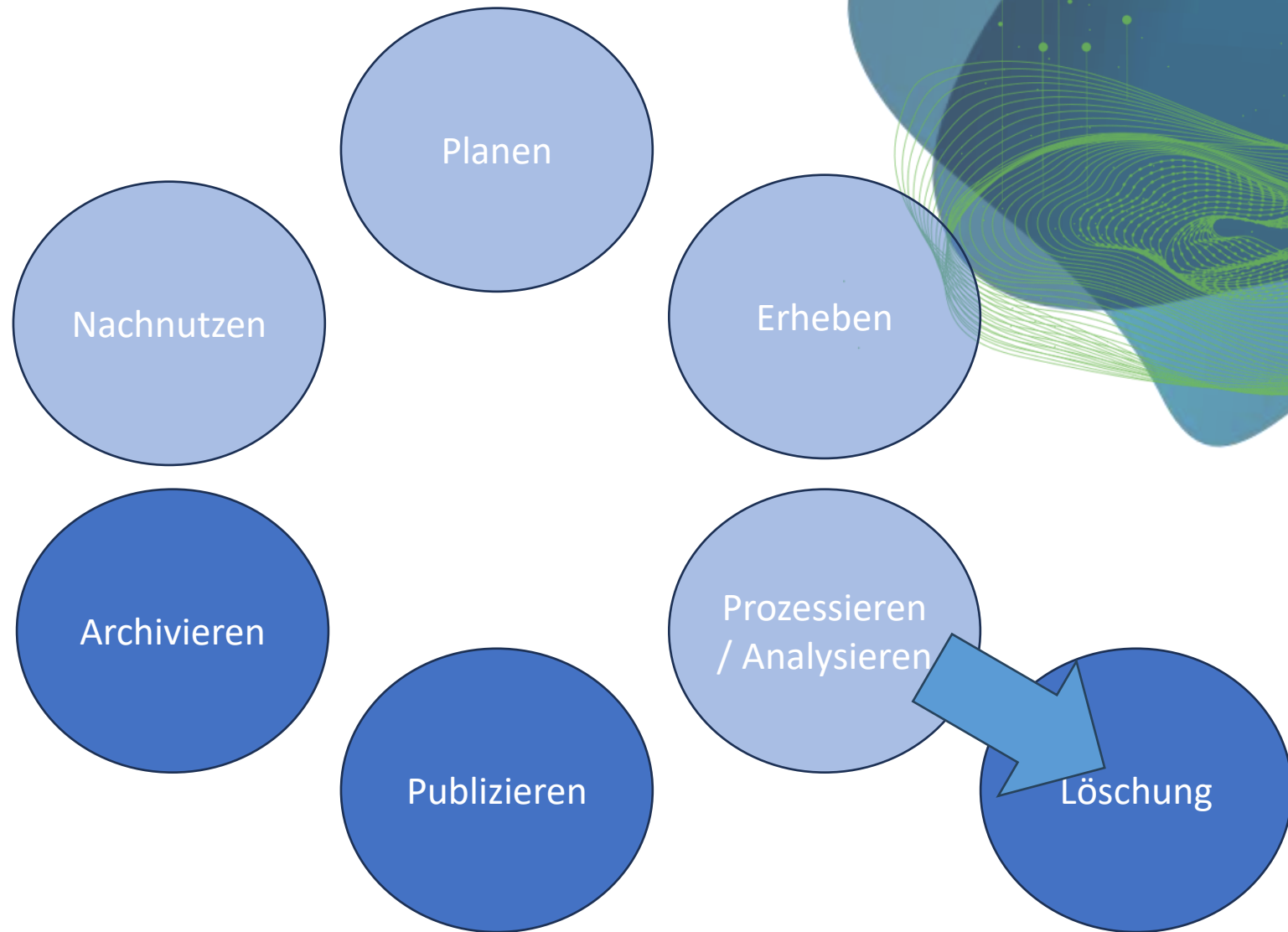
Arten von Daten

- Offen zugängliche Daten
- Forschungsdaten
- Personenbezogene Daten
- Geheime Daten



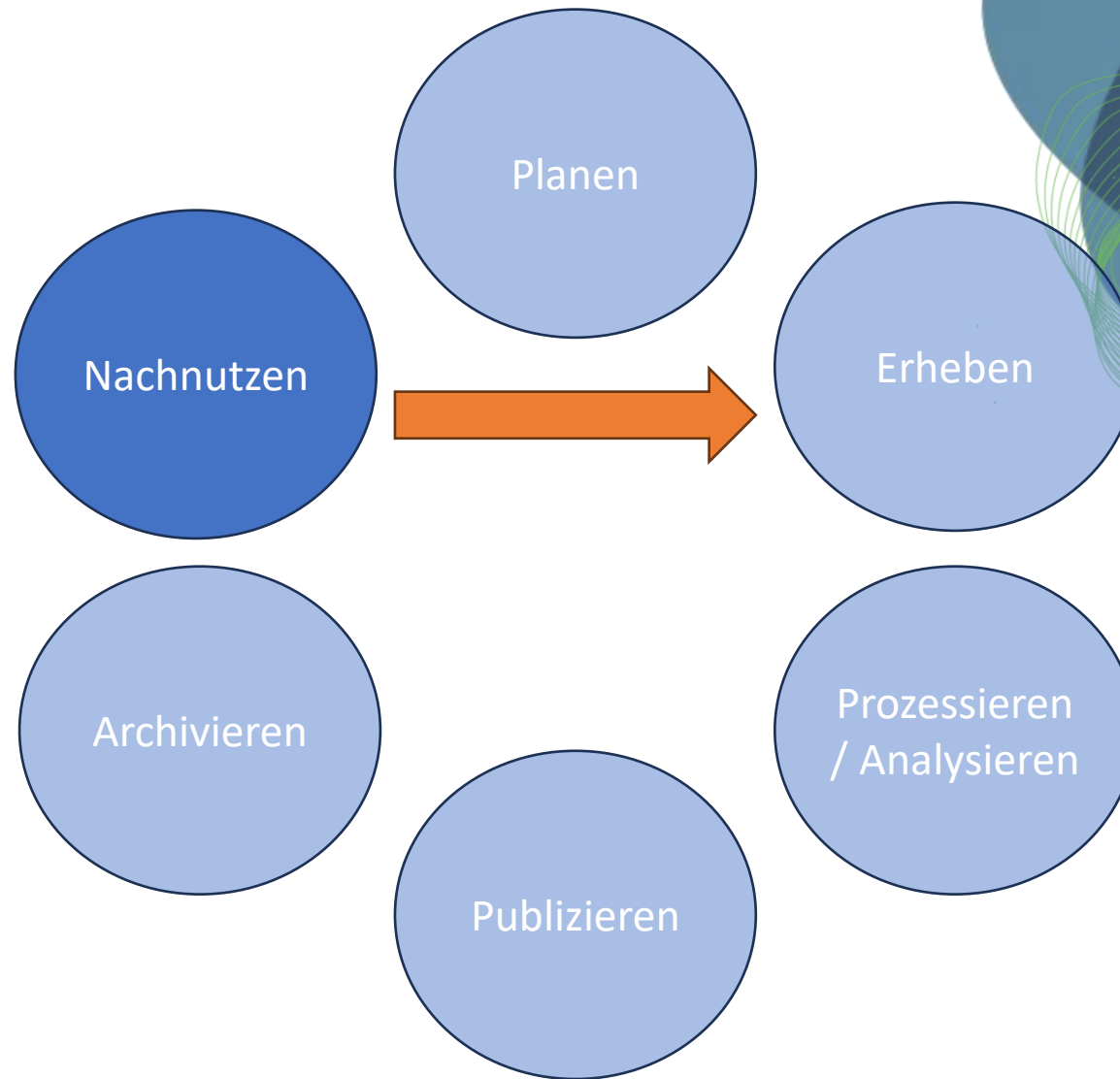
Lebenszyklus

- Veröffentlichungsrechte
- Gesetzliche Vorschriften
- Autorenschaften
- Registrierung (-> Findbarkeit)



Lebenszyklus

- Potentieller Mehrgewinn
- Nachhaltigkeit
- Wichtig: **Lizensierung**



Lizensierung

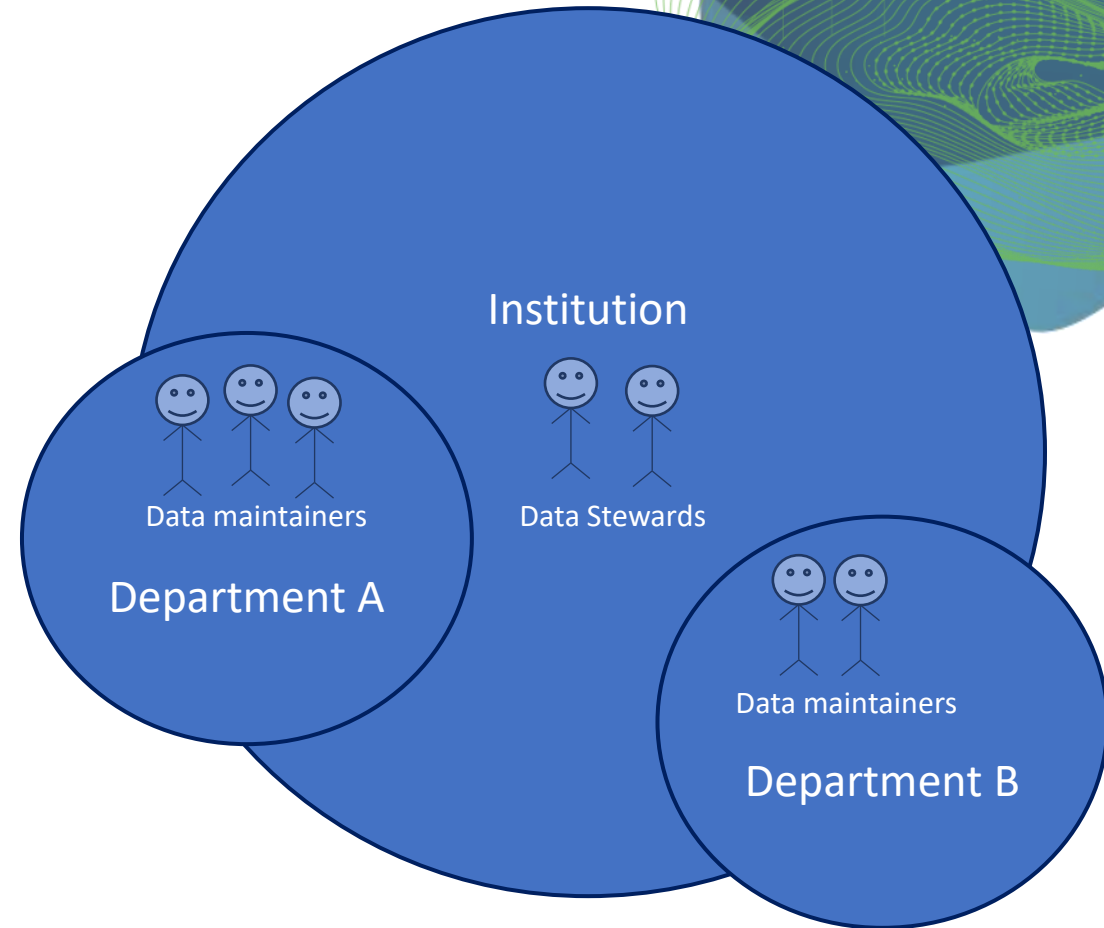
- Session gestern @ DataWeek Leipzig

The screenshot shows the Zenodo record page for 'Sharing & Licensing' by Robert Haase. The page includes the Zenodo logo, a 'Manage record' button, and the publication date 'Published April 12, 2024 | Version v2'. The title 'Sharing & Licensing' is prominently displayed, along with the author's name 'Haase, Robert^{1,2}'. A 'Show affiliations' button is visible. The abstract text discusses Open Science, Open Source, Open Access, and FAIR principles. At the bottom, there is a 'Files' section showing a PDF file named 'DataWeek_Sharing+Licensing.pdf'.

The screenshot shows the PDF viewer interface for 'DataWeek_Sharing+Licensing.pdf'. The viewer displays the first page of the document, which features logos for ScaDS.Atl, NFDI4 BIOIMAGE, and GloBIAS. The main content includes the title 'Sharing & Licensing' and the author's name 'Robert Haase'. There are callout boxes for 'Slides', 'Code', 'Text', and 'Data'. The footer contains a CC-BY 4.0 license notice and a DOI link: <https://doi.org/10.5281/zenodo.10966230>. The file size is listed as 85.4 MB.

Was ist **gutes** Datenmanagement?

- Klar definierte Verantwortlichkeiten und Prozesse (Governance)
- Gute Kommunikation
- Dediziertes Personal
- Ansprechpartner mit Expertise ("Data Stewards")



Rollen / Job-Profile

Domaenen-Spezialist/in

- Fokussiert auf Fragestellung, i.d.R. mit direktem Bezug zur Realität
- Beispiele: Geowissenschaftler/in, Stadtplaner/in

Datenanalyst/in

- Fokussiert auf Methoden zur Prozessierung / Analyse von Daten
- Beispiele: Statistiker/in, Datenwissenschaftler/in

IT-Spezialist/in

- Fokussiert auf IT Infrastruktur zur Haltung und Prozessierung von Daten
- Beispiele: Informatiker/in, IT-Sicherheitsexpert/in

Datenmanagementplan (DMP)

- Beschreibt in der Regel den IST-Zustand einer Datenumgebung (Oder zumindest einen realisierbaren Zustand)
- Administrative Informationen (Projektname, Datenurheber*in, weitere Mitwirkende, Kontakt, Förderprogramm usw.)
- Projekt- und Datensatzbeschreibung
 - Datentypen, -formate, -umfang
 - Angaben zu Metadaten und Standards
 - Qualitätsstandards

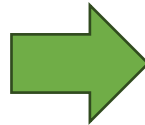
Datenmanagementplan (DMP)

- Wie werden die Daten verarbeitet? (SOPs)
- Wie werden Daten geteilt, publizieren?
- Archivierung und Backup der Daten
- Rollen, Verantwortlichkeiten
- Monitoring / Controlling / Qualitätssicherung
- Kosten

Datenmanagementplan (DMP)

- Trick: ChatGPT ist gut darin Dokumente zu strukturieren
- Wichtig: Sie sind für den Inhalt des DMPs verantwortlich, nicht ChatGPT

In our project we investigate the underlying physical principles for Gastrulation in Tribolium castaneum embryo development. Therefore, we use light-sheet microscopes to acquire 3D timelapse imaging data. We store this data in the NGFF file format. After acquisition, two scientists, typically a PhD student and a post-doc or group leader look into the data together and decide if the dataset will be analyzed in detail. In case yes, we upload the data to an Omero-Server, a research data management solution specifically developed for microscopy imaging data. Data on this server is automatically backed-up by the compute center of our university. We then login to the Jupyter Lab server of the institute where we analyze the data. Analysis results are also stored in the Omero-Server next to the imaging data results belong to. The Python analysis code we write is stored in the institutional git-server. Also this server is backed up by the compute center. When the project advances, we start writing a manuscript using overleaf, an online service for collaborative manuscript editing based on latex files. After every writing session, we save back the changed manuscript to the institutional git server. As soon as the manuscript is finished and submitted to the bioRxiv, a preprint server in the life-sciences, we also publish the project-related code by marking the project on the git-server as public. We also tag the code with a release version. At the same time we publish the imaging data by submitting a copy of the dataset from the Omero-Server to zenodo.org, a community-driven repository for research data funded by the European Union. Another copy of the data, the code and the manuscript is stored on the institutional archive server. This server, maintained by the compute center, guarantees to archive data for 15 years. Documents and data we published is licensed under CC-BY 4.0 license. The code we publish is licensed BSD3. The entire project and all steps of the data life-cycle are documented in an institutional labnotebook where every user has to pay 10 Euro per month. Four people will work on the project. The compute center estimates the costs for storage and maintenance of the infrastructure to 20k Euro and half a position of an IT specialist. The project duration is four years.



Data Management Plan

Data Description

Our study aims at understanding the fundamental physical principles informing Gastrulation in Tribolium castaneum embryo development. Through this endeavor, we will generate 3D timelapse imaging data gathered from light-sheet microscopes. The data sets are stored using the NGFF file format. In the process of the project, we will create Python-based code for data analysis, research papers drafted on Overleaf, which incorporates the findings and results of our study.

Documentation and Data Quality

Once the required data is acquired, it undergoes a scrutiny process where two scientists, generally a post-doc or group leader along with a PhD scholar, determine whether the data merits a detailed analysis. If affirmed, it will be stored on our Omero-Server – an exclusive data management solution created for microscopy imaging data. The analysis results, Python analysis scripts, and manuscript edits are also stored in this server and always updated after each modification.

Storage and Technical Archiving of the Project

The data on the Omero-Server and the institutional git-server, where we store the Python analysis code and edits to our manuscript, are automatically backed up by our university's compute center. Completed aspects of the project are retained in the institutional archive server for a guaranteed period of 15 years. This archive likewise holds a copy of all published data, code, and manuscripts attached to the project.

Legal Obligations and Conditions

Our published documents and data are licensed under the CC-BY 4.0 license, while the published code follows the BSD3 license. Individuals working on the project and accessing the institutional lab notebook are required to pay a monthly fee of 10 Euros.

Data Exchange and Long-term Data Accessibility

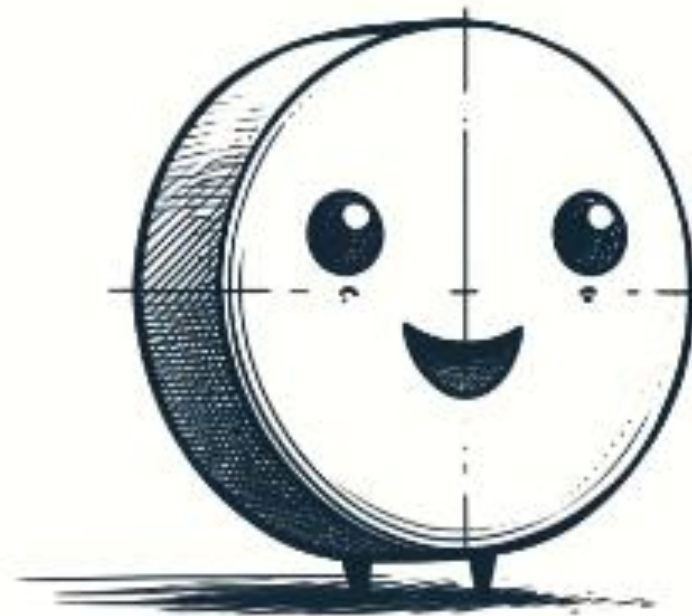
To ensure wider accessibility and visibility, we publish our finalized manuscripts to the bioRxiv – a preprint server primarily geared towards the life-sciences. We further open-source our project-related Python code by switching the settings on our git-server to public and tagging the code with a release version. All imaging data will be made accessible by uploading a copy of the data set from the Omero-Server to zenodo.org – a community-driven repository for research data backed by the European Union.

Responsibilities and Resources

The research group comprises four members who will be actively involved in the project that spans over four years. The compute center estimates the infrastructural costs for both storage and maintenance to be 20k Euros and one-half of an IT specialist's position.

Zusammenfassung

- Daten haben einen Lebenszyklus, idealerweise iterativ.
- Die FAIR-Prinzipien sind ein Katalog von Zielen, die mit professionellem Datenmanagement verfolgt werden können.
- Rollen \neq Job-Profile
- Datenmanagementpläne dienen dazu Projekte vorab zu durchdenken und nachvollziehbar zu dokumentieren.



Acknowledgements

Communities & platforms

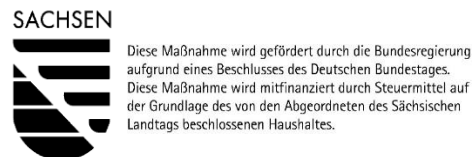


BiAPoL team

- Marcelo Zoccoler
 - Johannes Soltwedel
 - Maleeha Hassan
 - Stefan Hahmann
- Former lab members:
- Ryan George Savill
 - Laura Zigutyte
 - Mara Lampert
 - Allyson Ryan
 - Conni Wetzker
 - Somashekhar Kulkarni
 - Till Korten



Funding



Some Figures were generated using

