ScaDS.AI DRESDEN LEIPZIG

CENTER FOR SCALABLE DATA ANALYTICS AND ARTIFICIAL INTELLIGENCE

NFDI 4 BIOIMAGE

NATIONAL RESEARCH DATA MANAGEMENT INFRASTRUCTURE FOR MICROSCOPY AND BIOIMAGE ANALYSIS

# 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.





https://www.doi.org/10.5281/zenodo.10970869

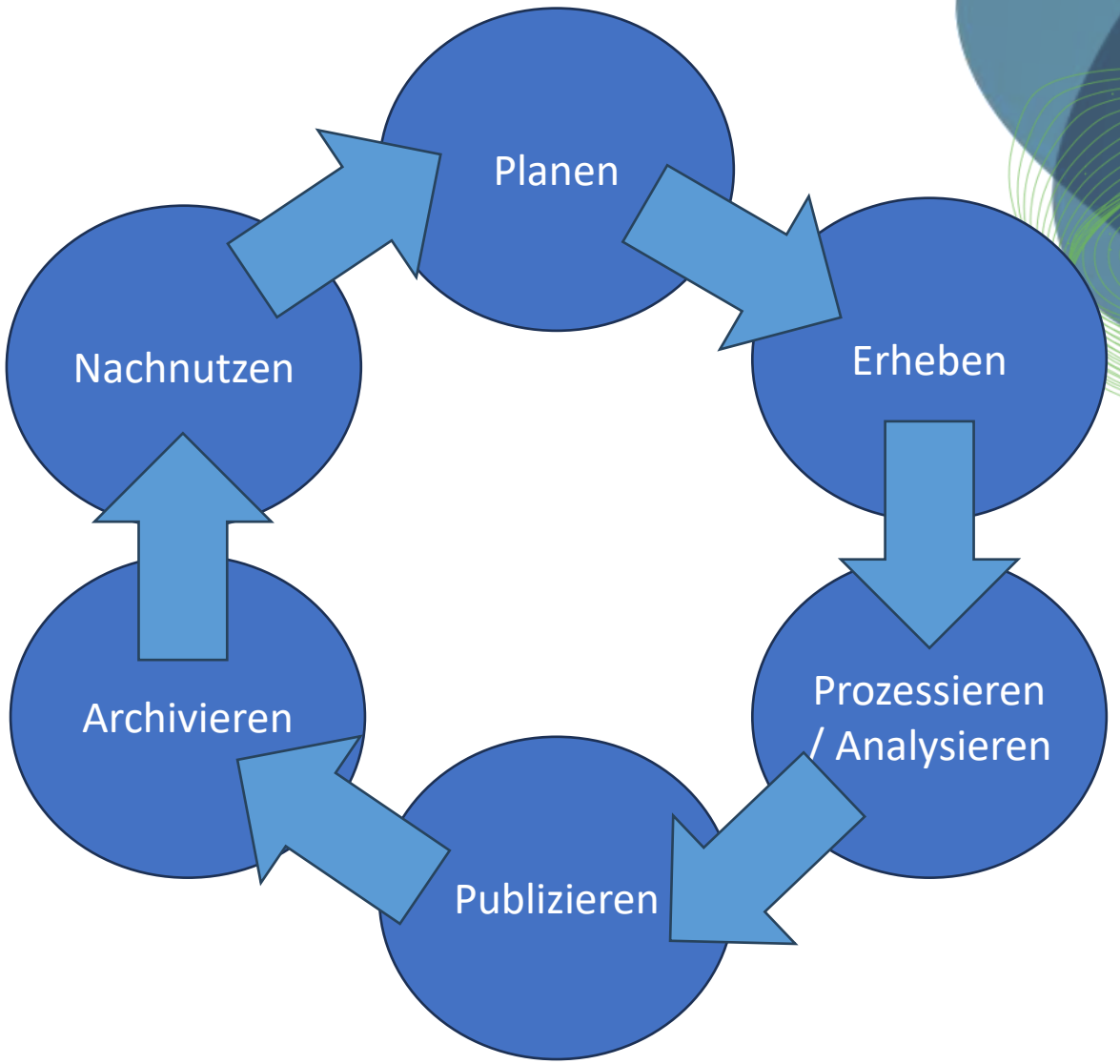
# Datenmanagement

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


Figure made by Dall-E 3

# Lebenszyklus

- Idealerweise sind Prozesse mit Daten in einem iterartiven Zyklus eingebettet

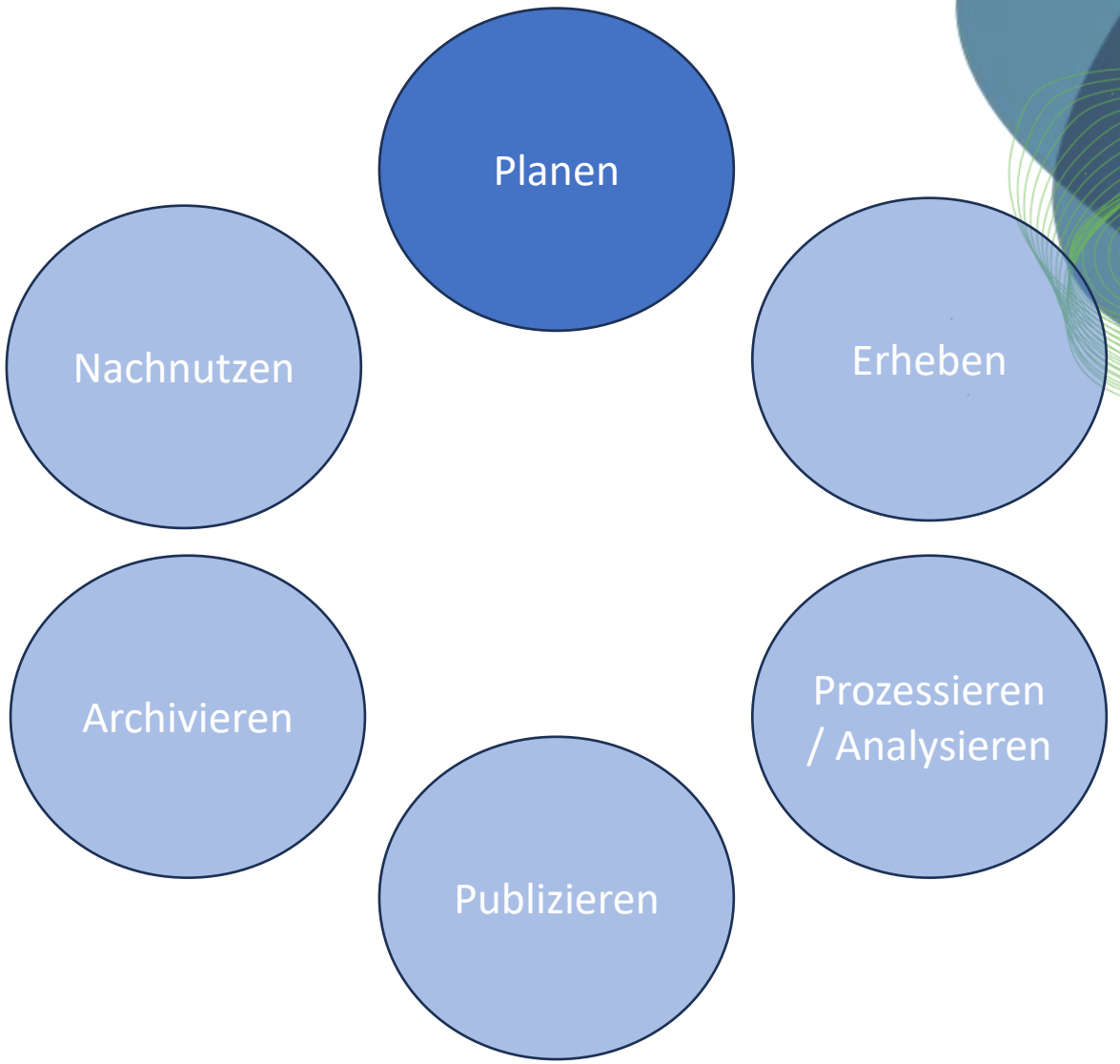





Adaptiert von:
https://forschungsdaten.info/themen/informieren-und-planen/datenlebenszyklus/

# Lebenszyklus

- Kosten
- Nutzen
- Qualitaetsziele
- Strategische Entscheidungen

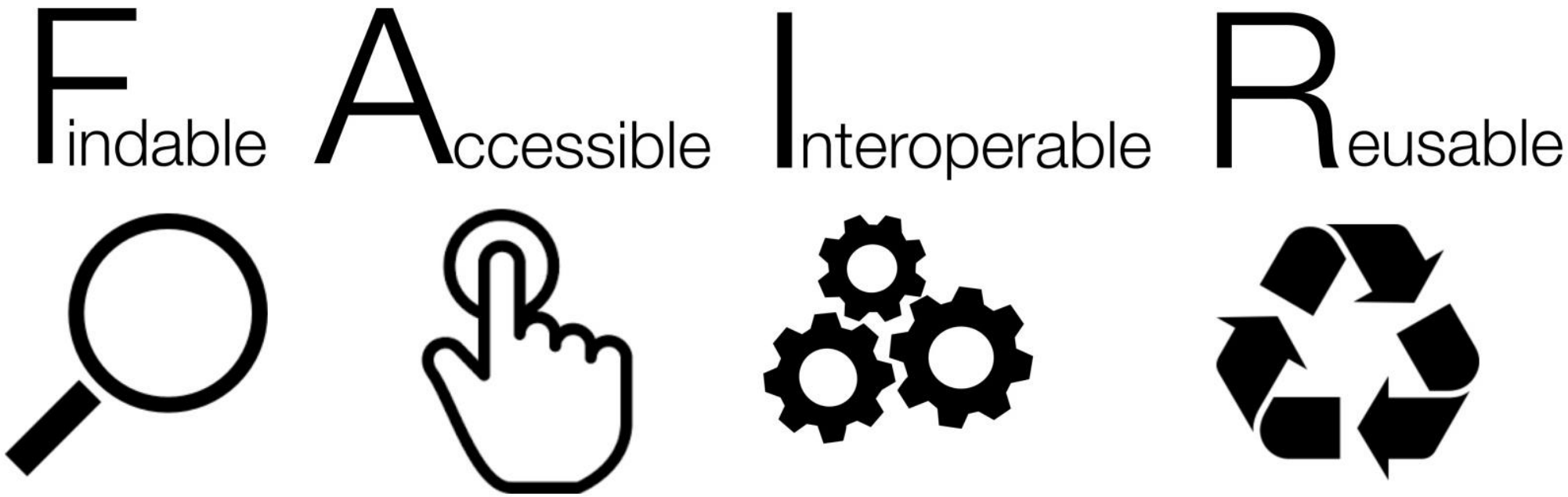





Adaptiert von:
https://forschungsdaten.info/themen/informieren-und-planen/datenlebenszyklus/

# Datenmanagement

- Oft genanntes strategisches Ziel:



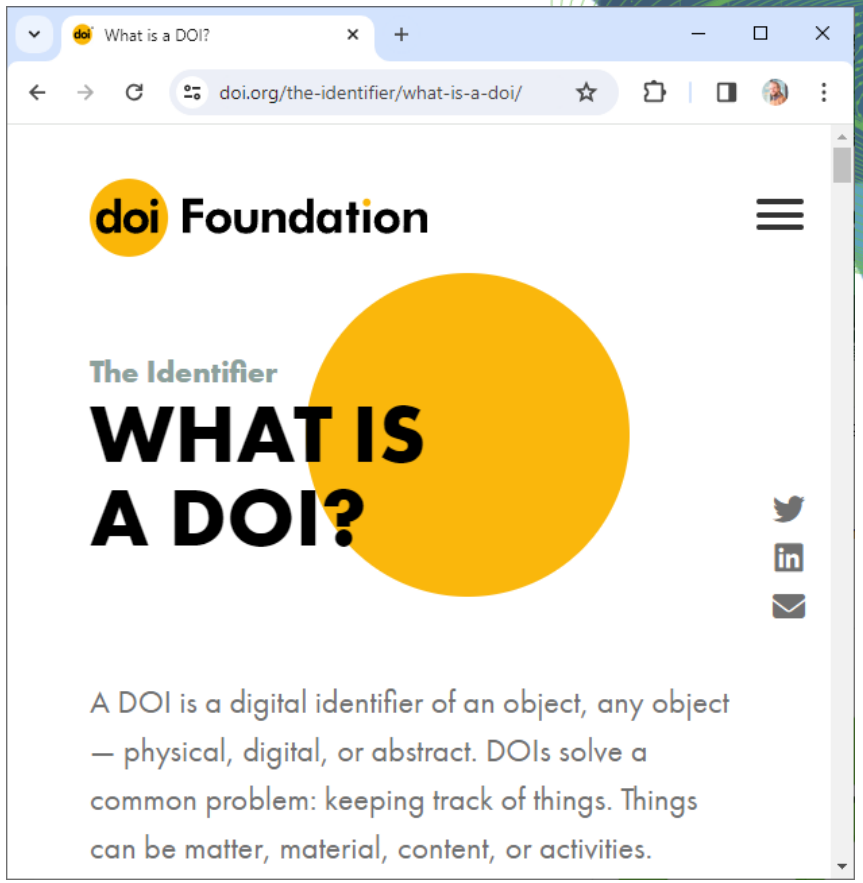

Siehe auch: https://www.go-fair.org/fair-principles/

Figure quelle: Pundir, Sangya. https://commons.wikimedia.org/wiki/File:FAIR_data_principles.jpg lizensiert CC-BY-SA-4.0

# 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





Screenshot source: https://www.doi.org/the-identifier/what-is-a-doi/ licensed CC-BY 4.0
https://www.go-fair.org/fair-principles/
https://biii.eu/introduction-3d-analysis-3d-imagej-suite

# 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

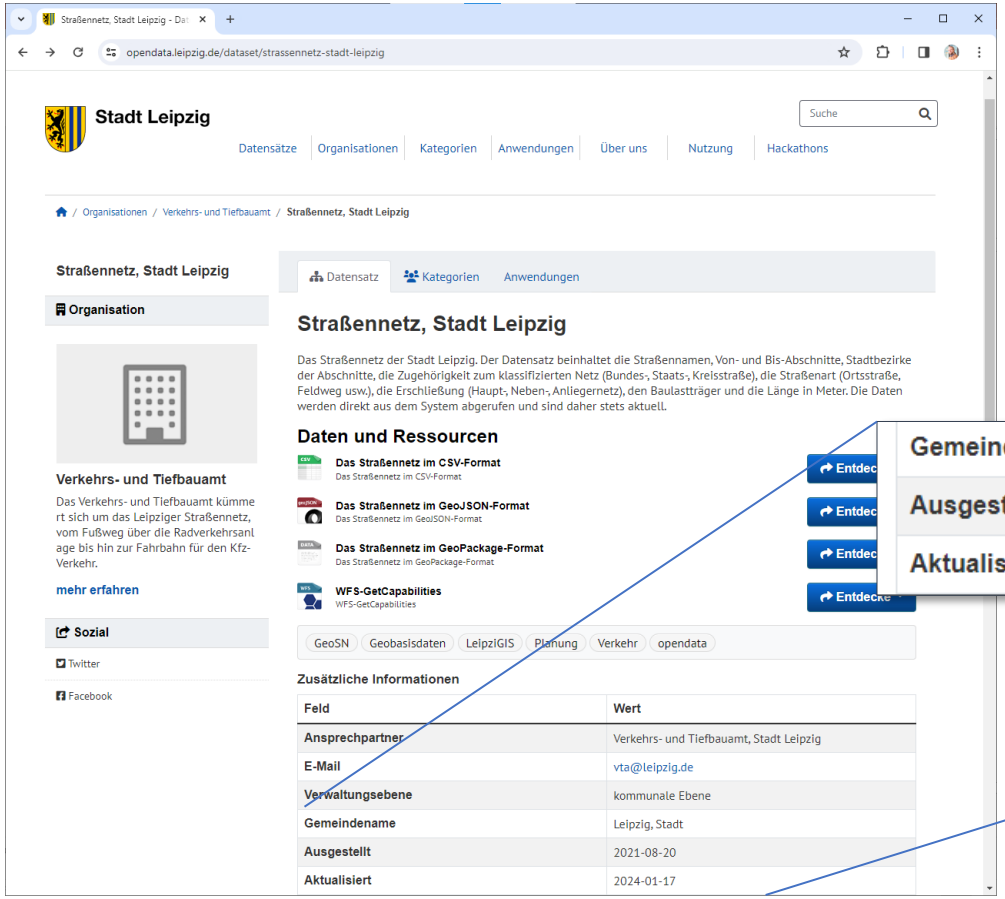
https://doi.org/10.5281/zenodo.28325

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

https://www.leipzig.de/

# Resource Identifiers

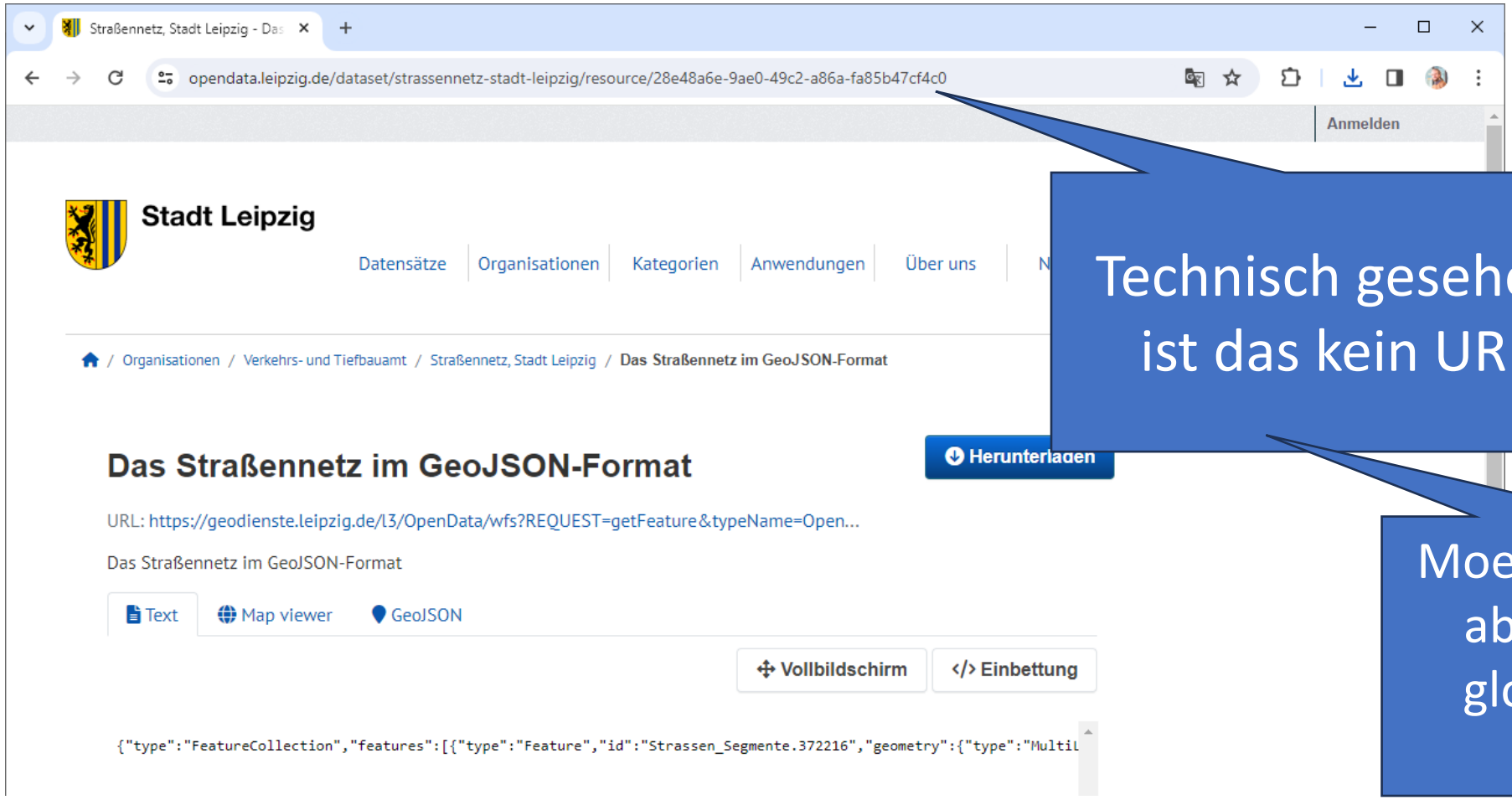
- Unique Identifier zeigen immer auf die gleichen Daten




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

Das ist also kein Unique Identifier

# Resource Identifiers

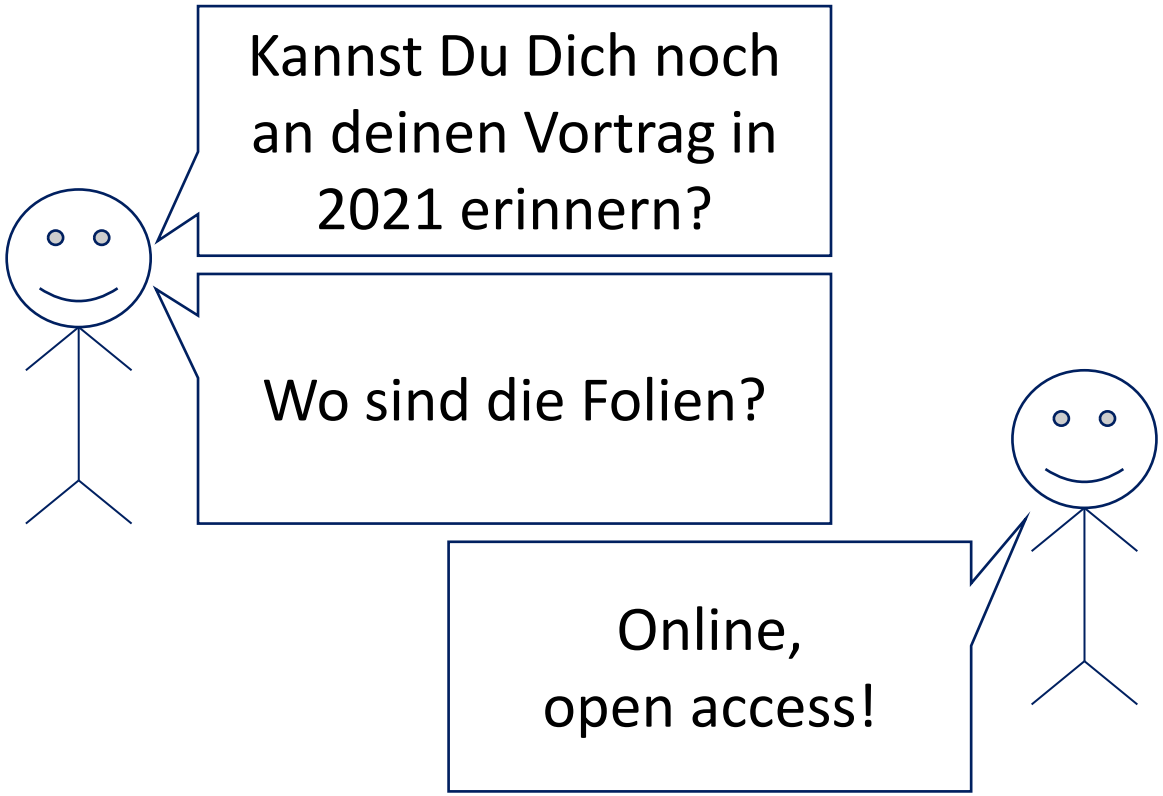


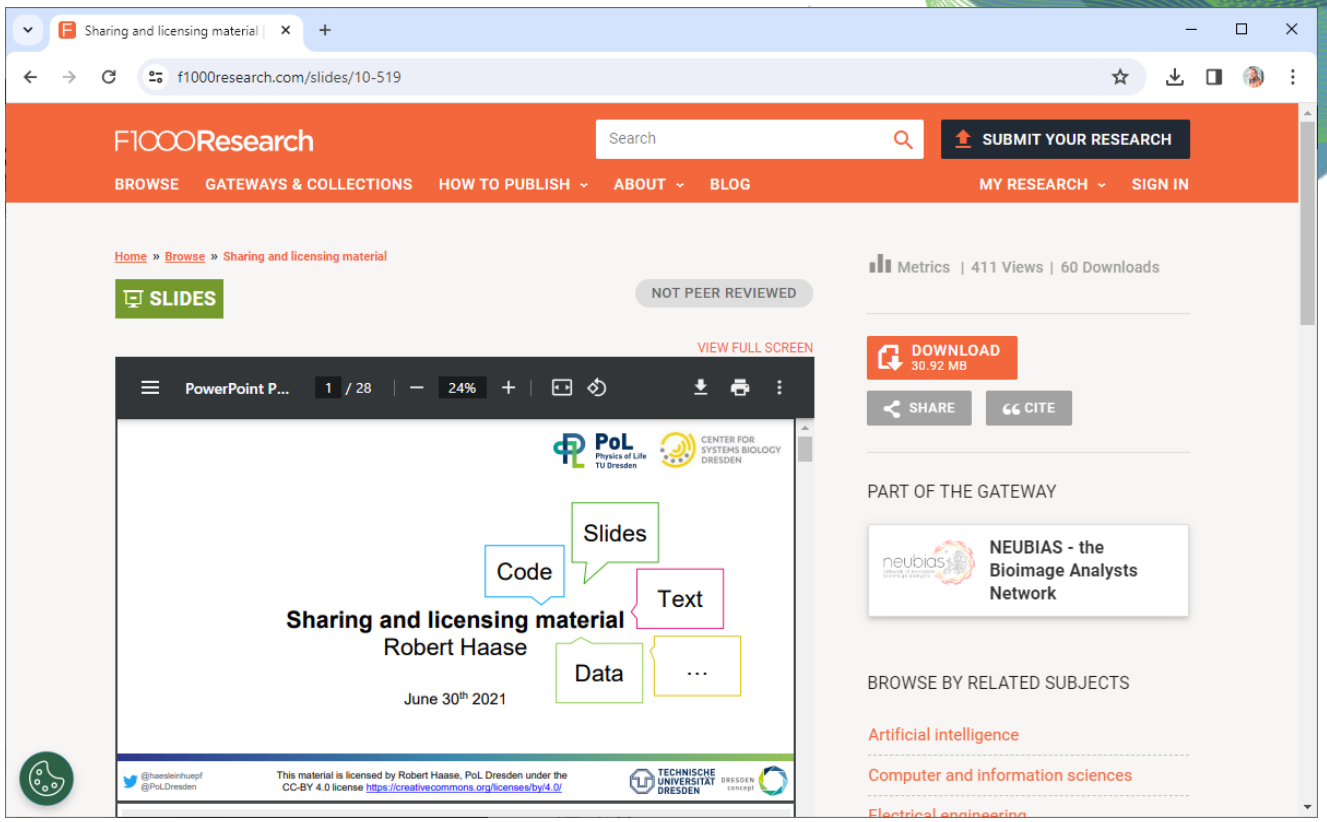

Technisch gesehen ist das kein URI

Moeglicherweise ist er aber nicht in einer globalen Resource registriert.

# Findbarkeit

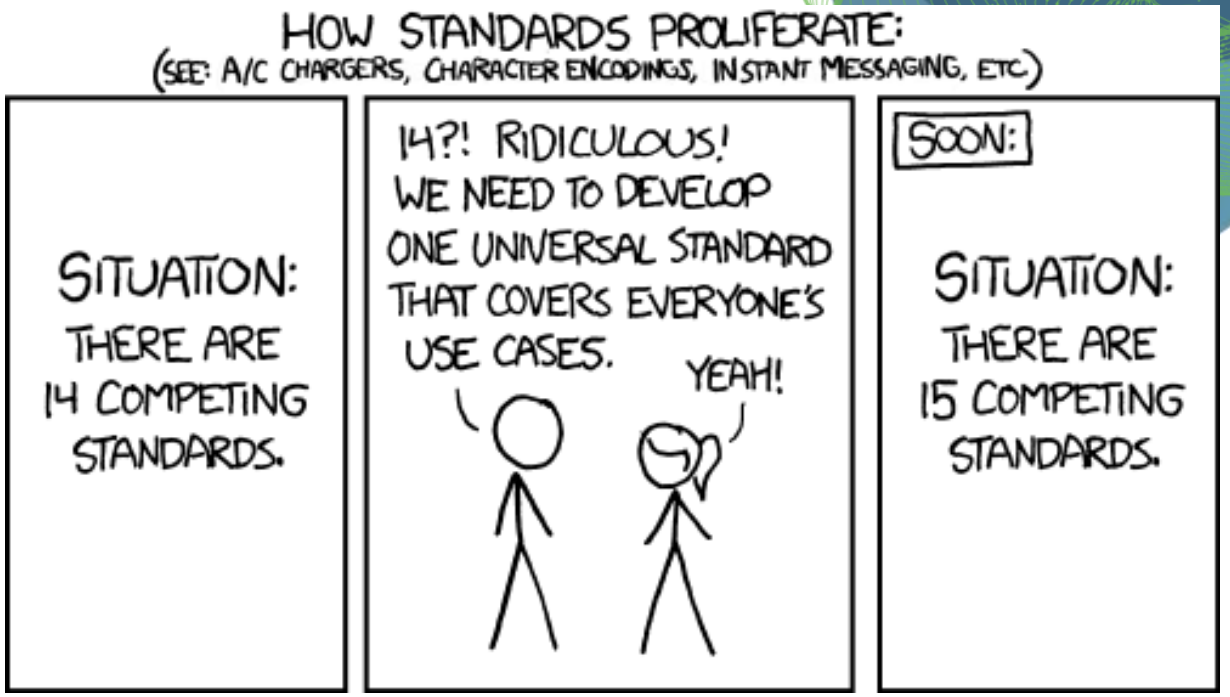
- Unser zukuenftiges Selbst wird uns danken.



# Die FAIR-Prinzipien

- Accessible / Verfuegbar / Zugreifbar
- A1. (Meta)daten koennen ueber ein Standardisiertes Protokoll empfangen werden
  - A1.1 Das Protokoll ist offen, fei und universell implementierbar
  - A1.2 Das Protokoll erlaubt Authentifizierung und Authorisierung, wenn erforderlich
- A2. Metadaten sind verfuegbar, auch wenn die Daten selbst nicht mehr verfuegbar sind

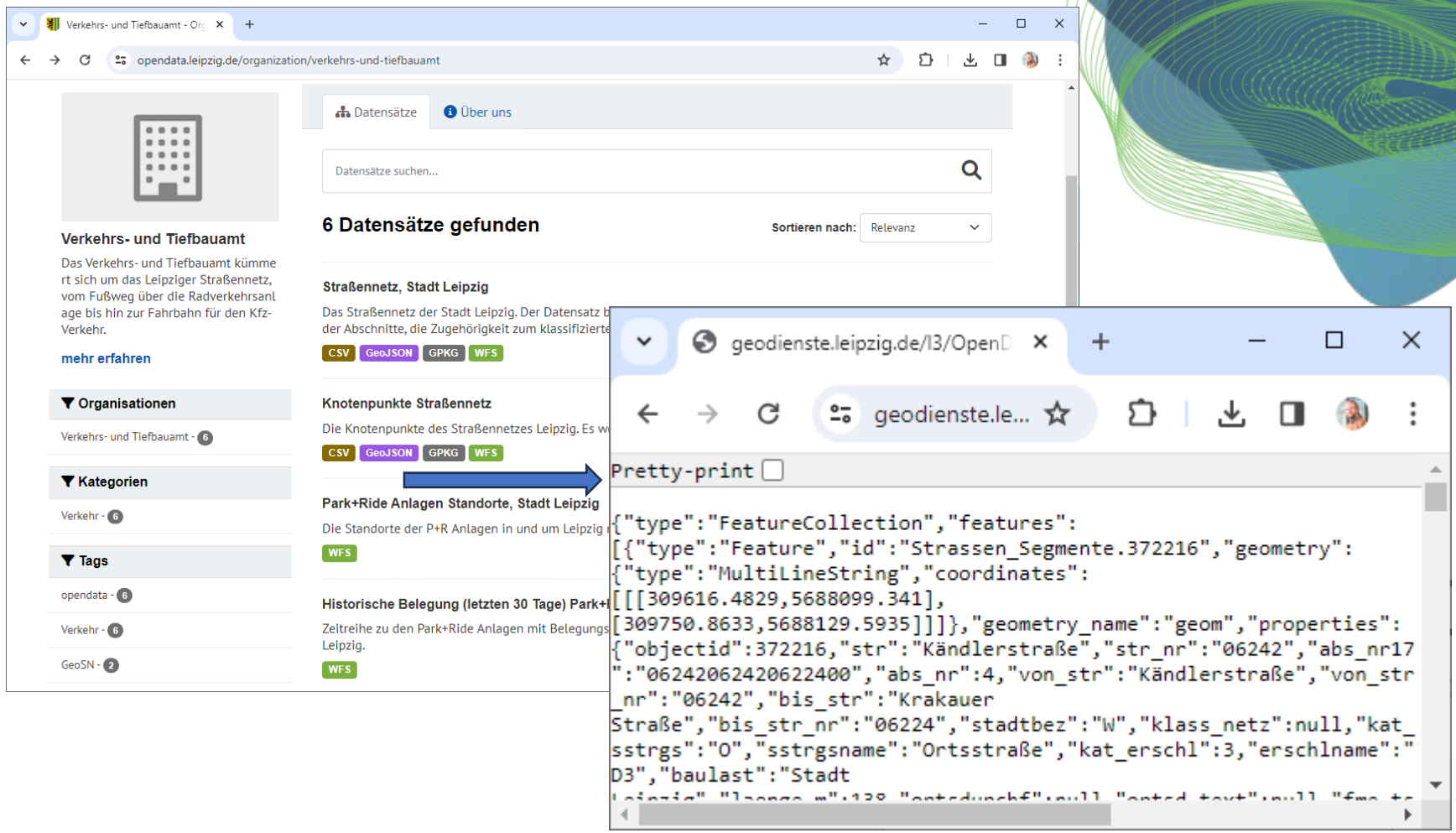





https://www.go-fair.org/fair-principles/

# Accessibility

- Beispiel: Geodaten
  - GeoJSON

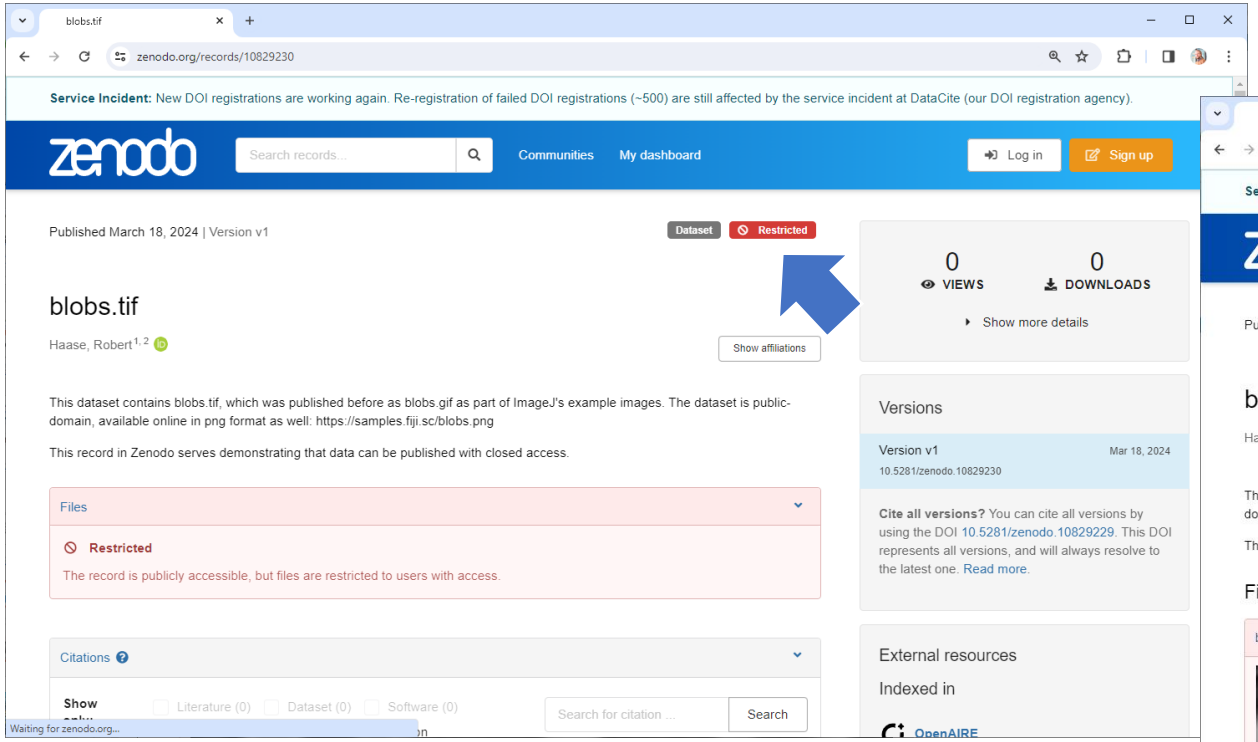


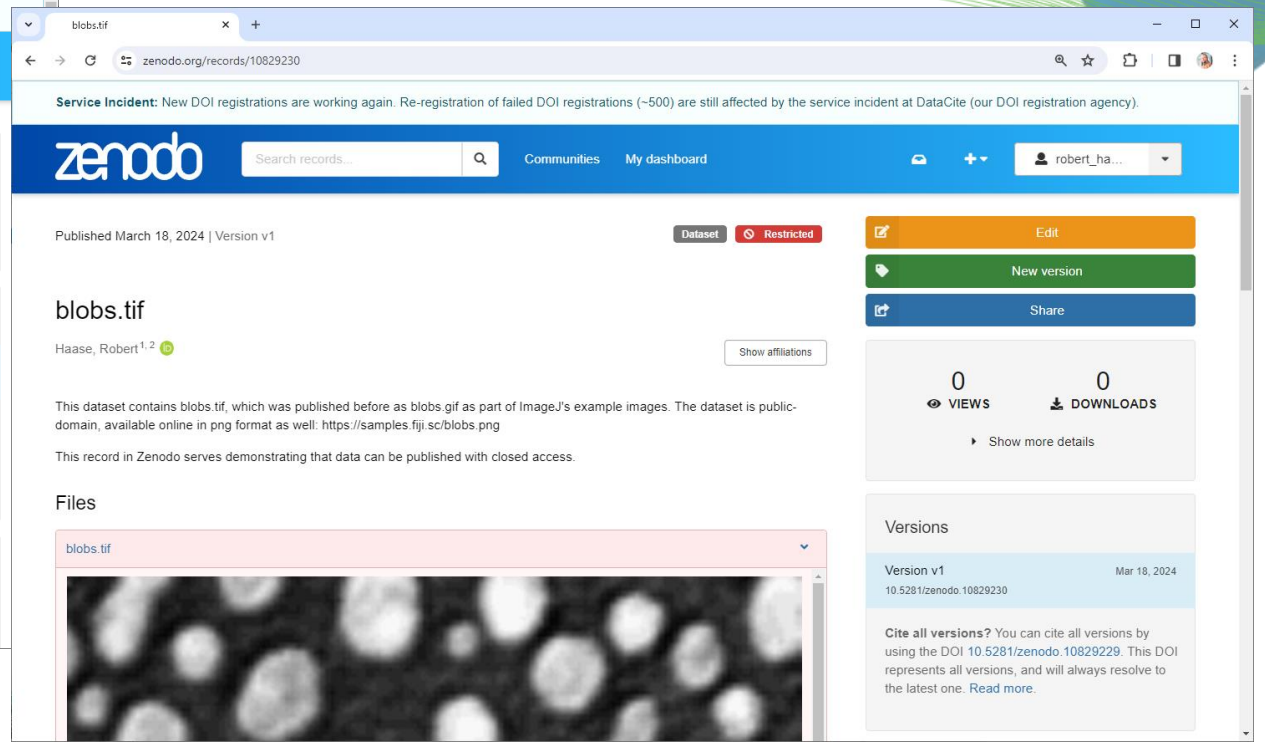



Quellen: Verkehrs und Tiefbauamt Leipzig, https://opendata.leipzig.de/dataset/strassennetz-stadt-leipzig/resource/28e48a6e-9ae0-49c2-a86a-fa85b47cf4c0 https://geodienste.leipzig.de/l3/OpenData/wfs?REQUEST=getFeature&typeName=OpenData%3AStrassen_Segmente&outputFormat=application/json Lizenz: Datenlizenz Deutschland dl-de/by-2-0 www.govdata.de/dl-de/by-2-0

# Restricted Access

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



# Die FAIR-Prinzipien

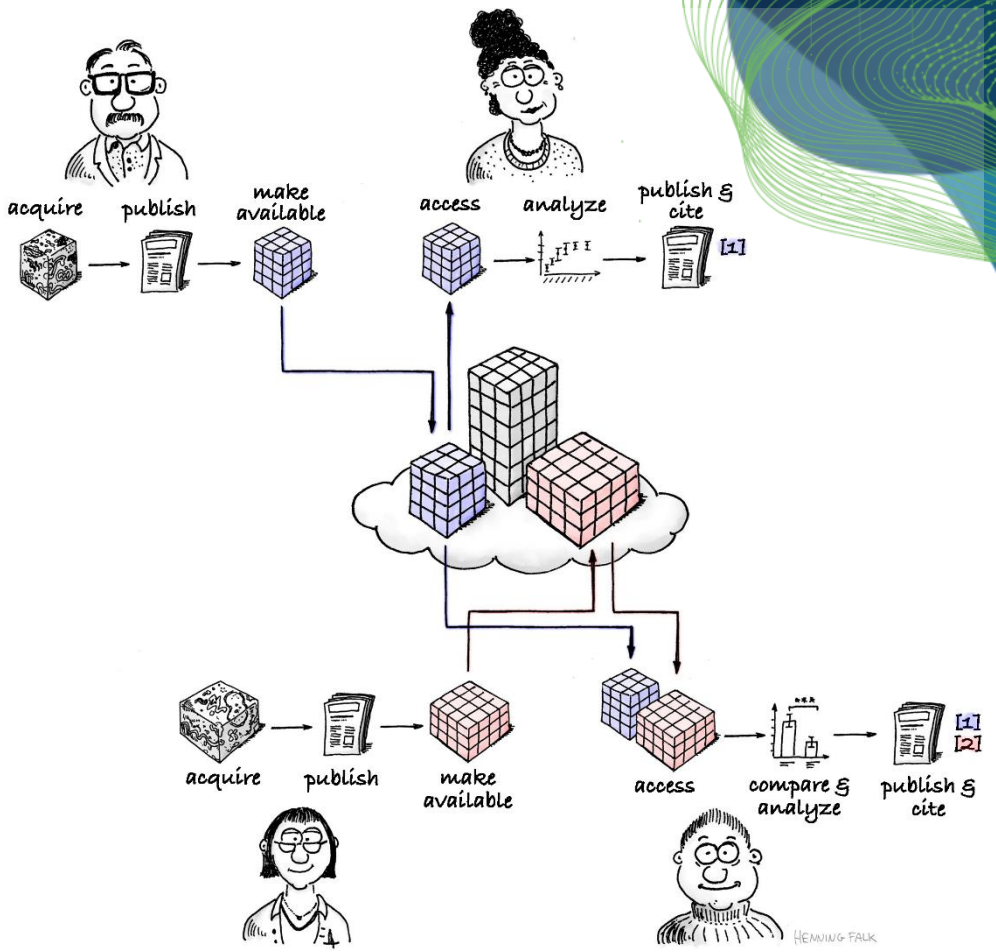
- Interoperable
  - I1. (Meta)daten sind formalisiert in einer zugaenglichen, gemeinsamen, breit angewandten Sprache, geeignet fuer Wissensrepraesentation
  - I2. (Meta)daten nutzen ein Vokabular, dass ebenfalls den FAIR-Prinzipien unterliegt
  - I3. (Meta)daten referenzieren andere qualifizierte (Meta)daten

https://www.go-fair.org/fair-principles

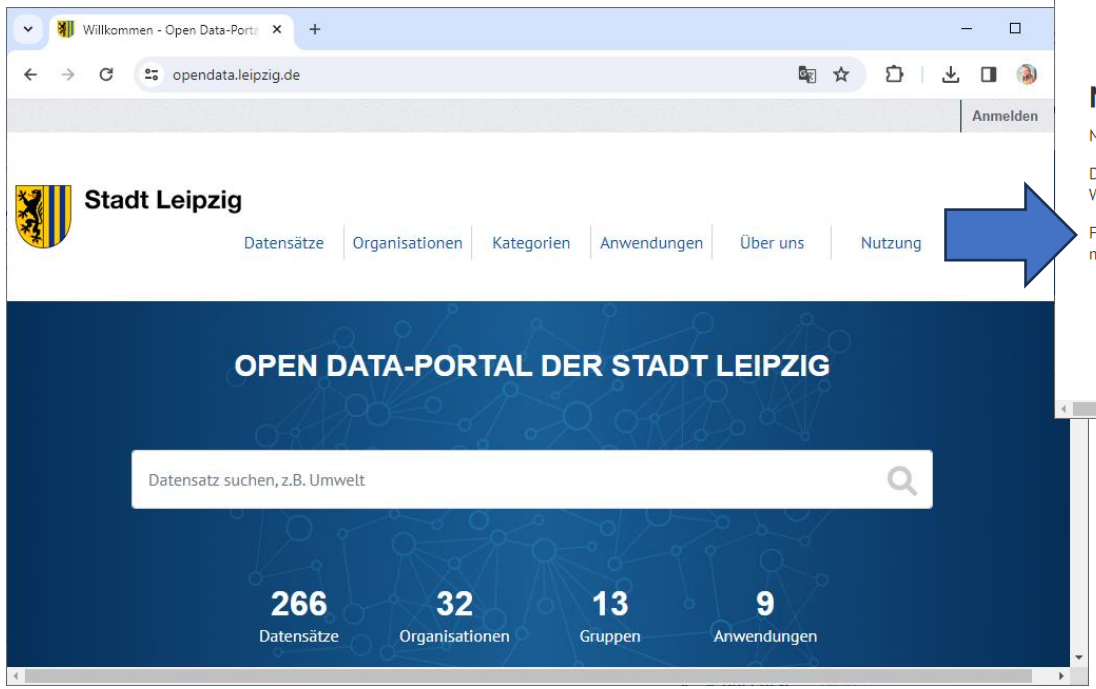
Figure made by Dall-E 3

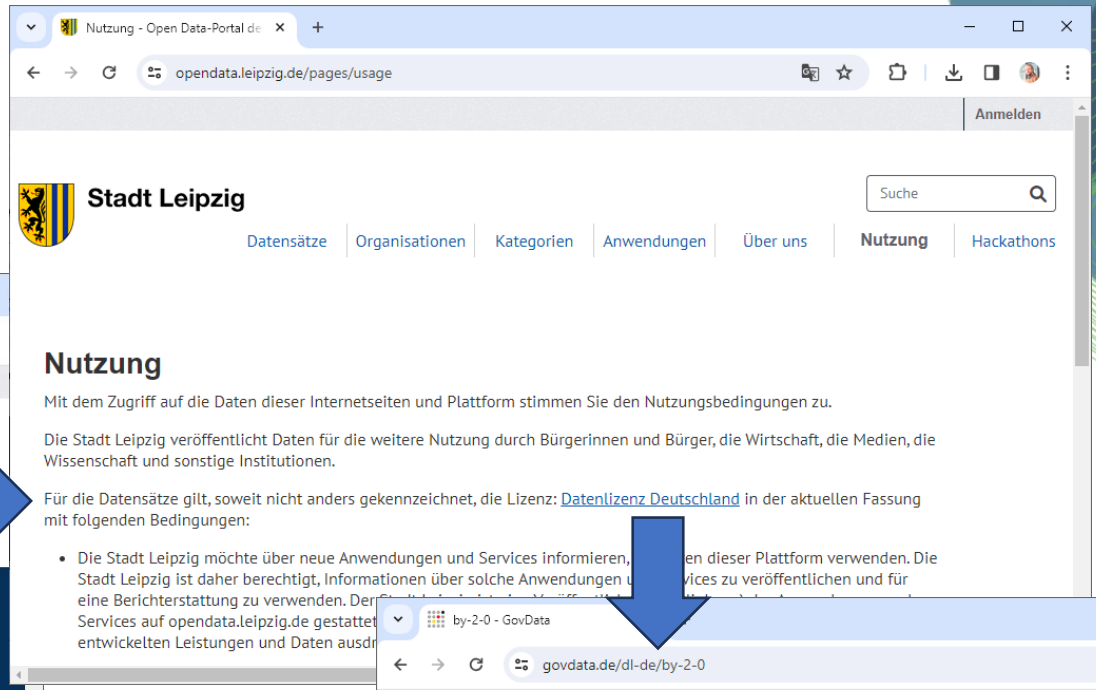
# Die FAIR-Prinzipien

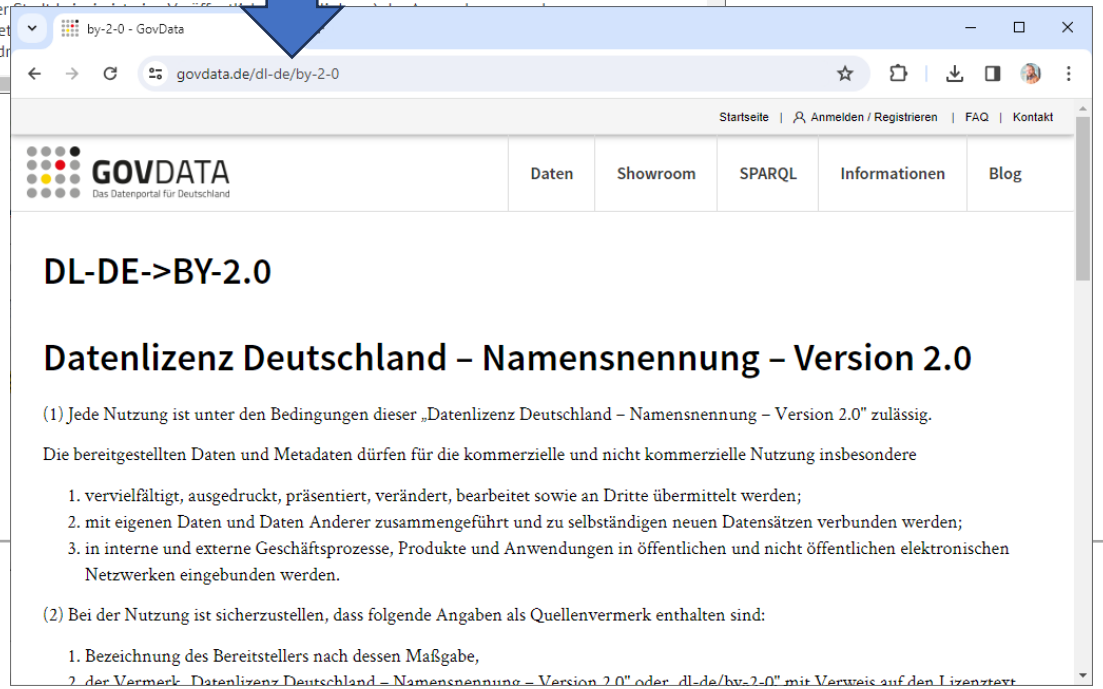
- Reusable / Wiederverwendbar
  - R1. (Meta)daten sind reich an vielfaeltigen, akkuraten und relevanten Attributen
  - R1.1. (Meta)daten werden mit einer klaren und verfuegbaren Nutzungslizenz versehen
  - R1.2. (Meta)daten sind stets mit der detailierten Herkunft erfasst
  - R1.3. (Meta)daten folgen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



# Lizensierung



# 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

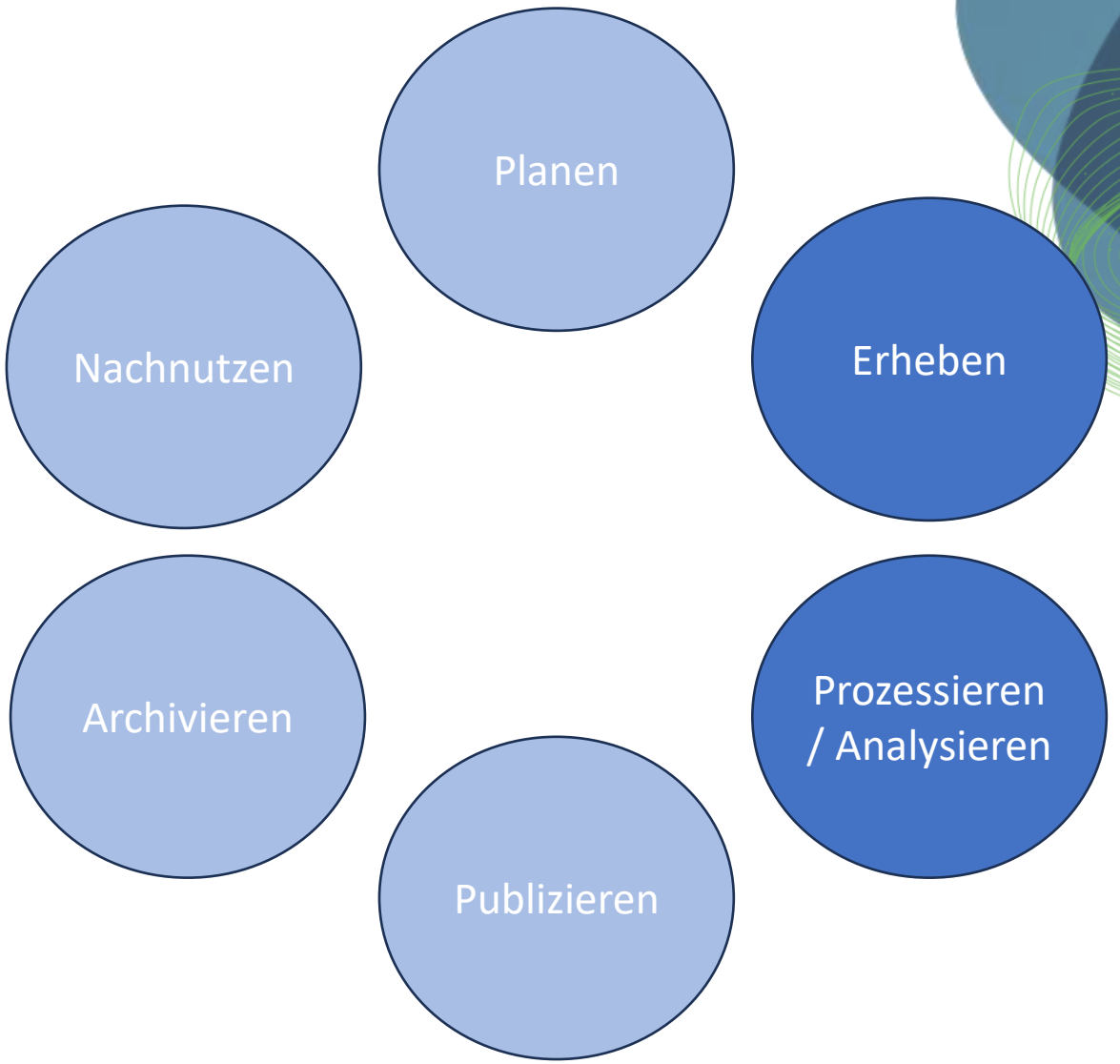
CC-BY

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

Figure made by Dall-E 3

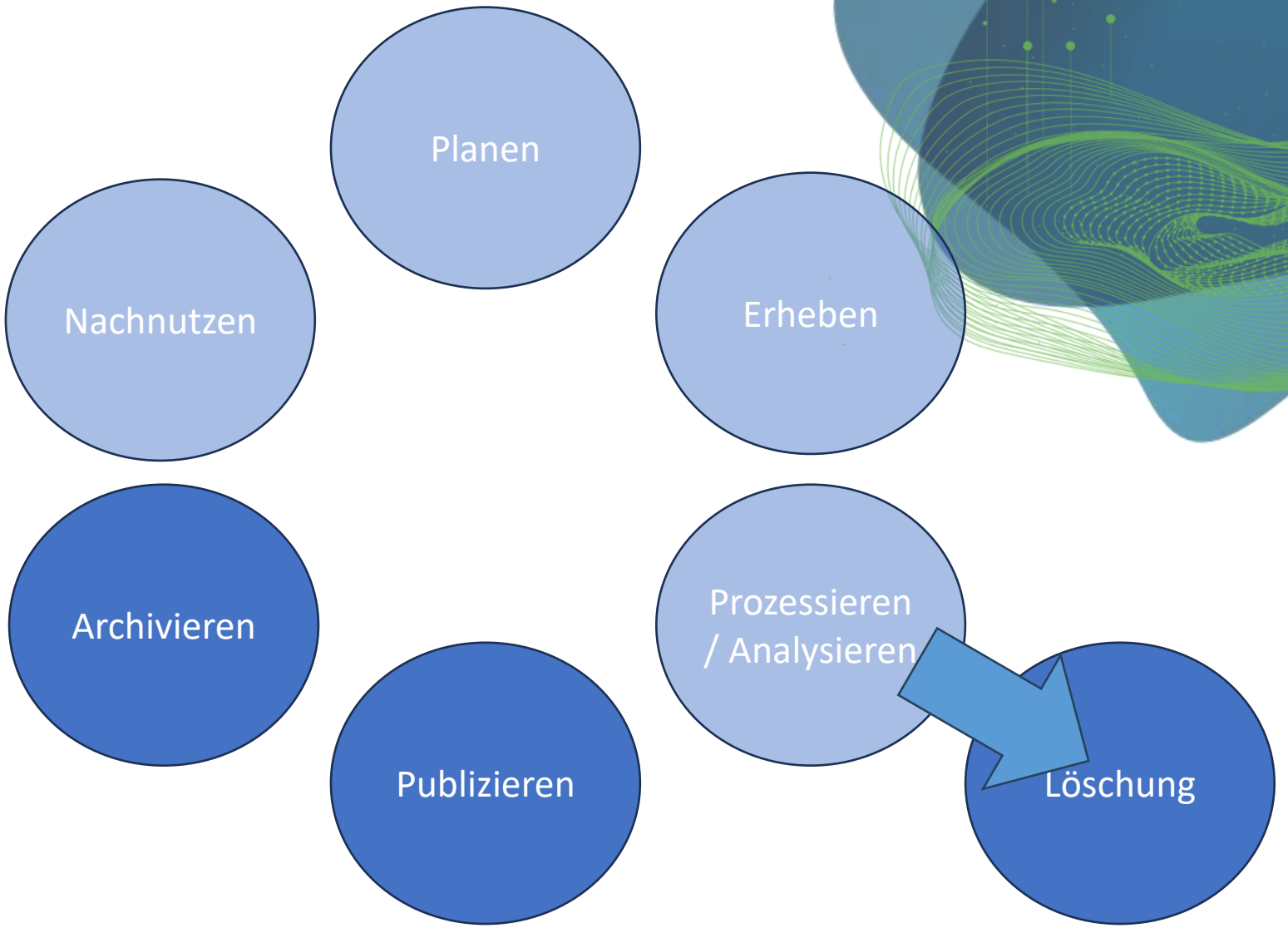
# Arten von Daten

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

Figure made by Dall-E 3

# Lebenszyklus

- Veröffentlichsungs-rechte
- Gesetzliche Vorschriften
- Autorenschaften
- Registrierung (-> Findbarkeit)

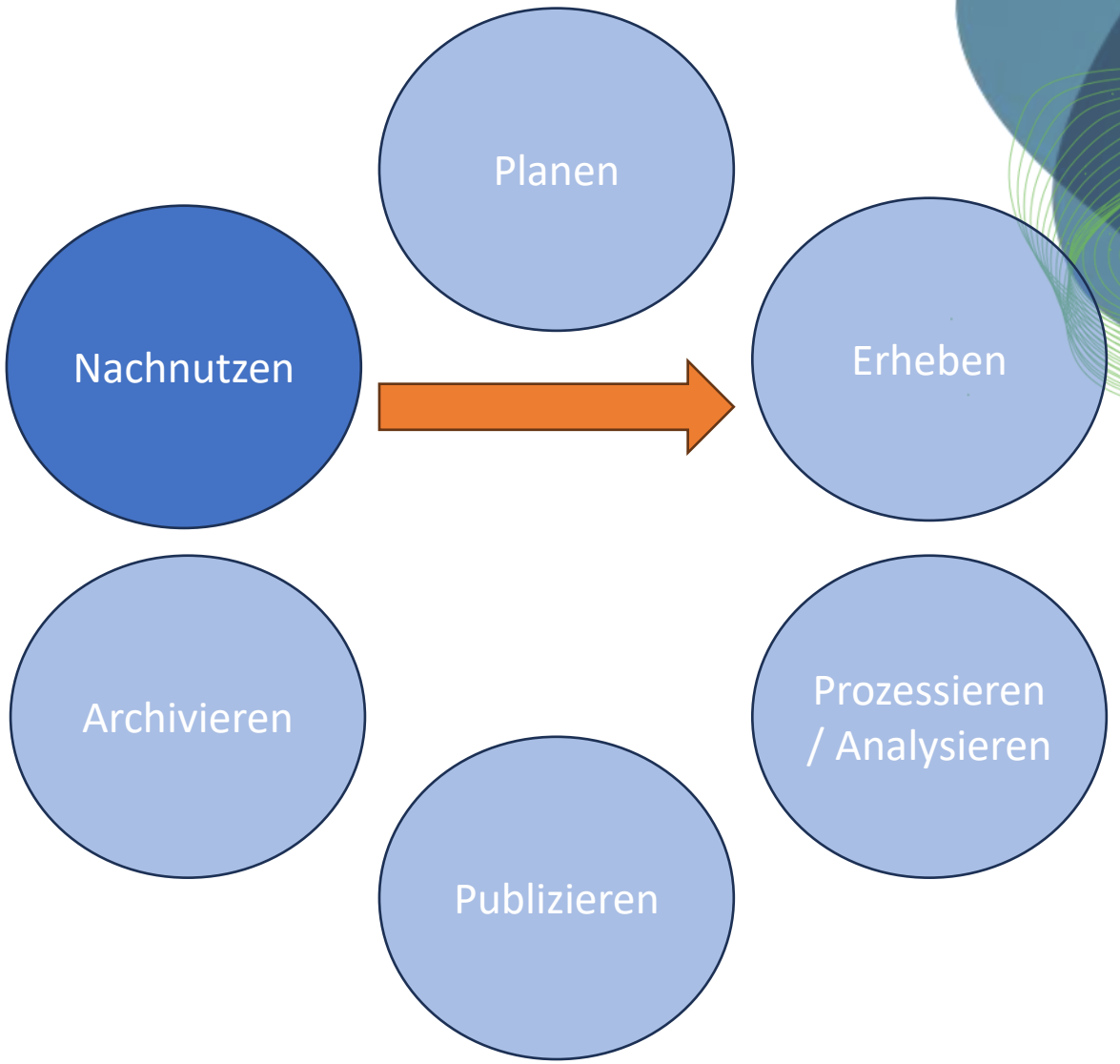





Adaptiert von: https://forschungsdaten.info/themen/informieren-und-planen/datenlebenszyklus/

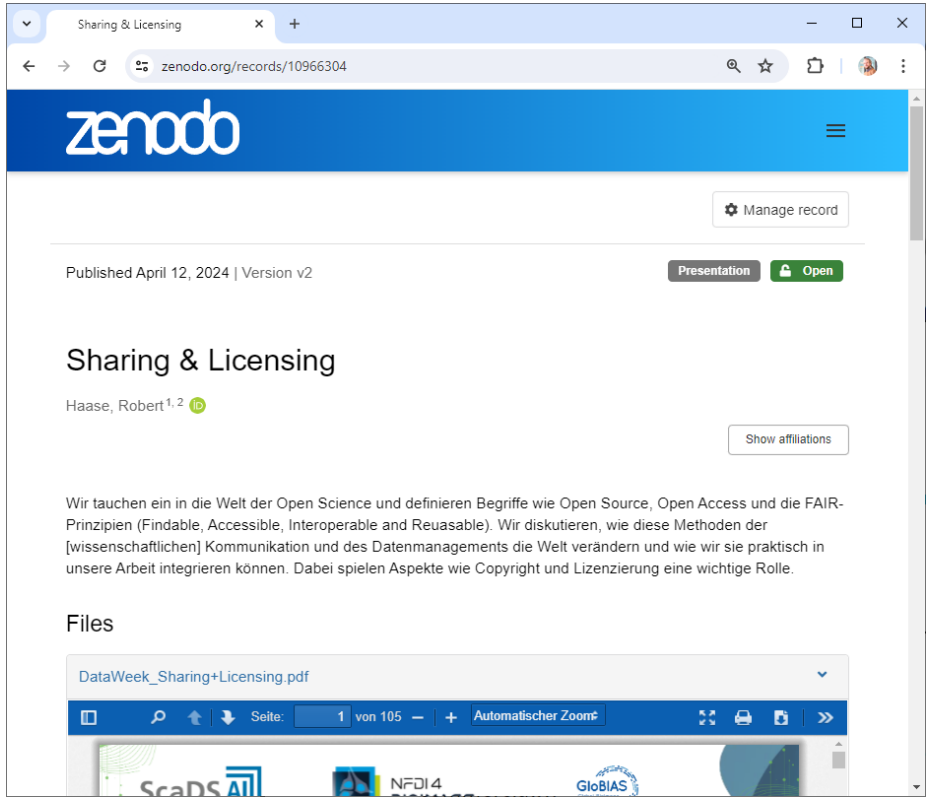
# Lebenszyklus

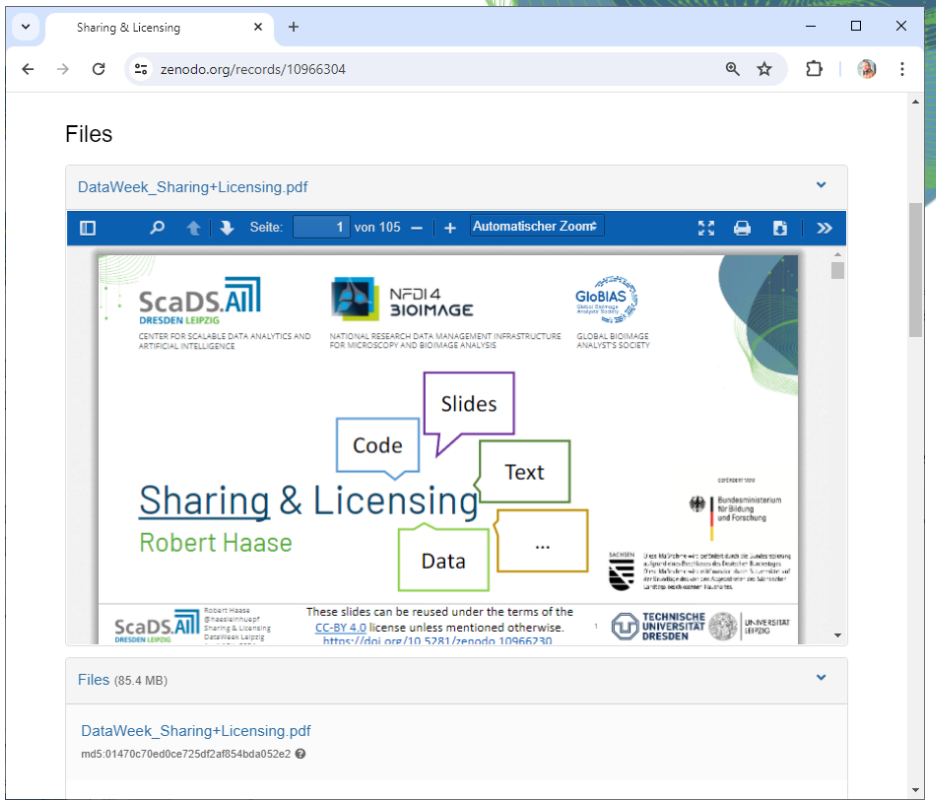
- Potentieller Mehrgewinn
- Nachhaltigkeit
- Wichtig: Lizensierung



# Lizensierung

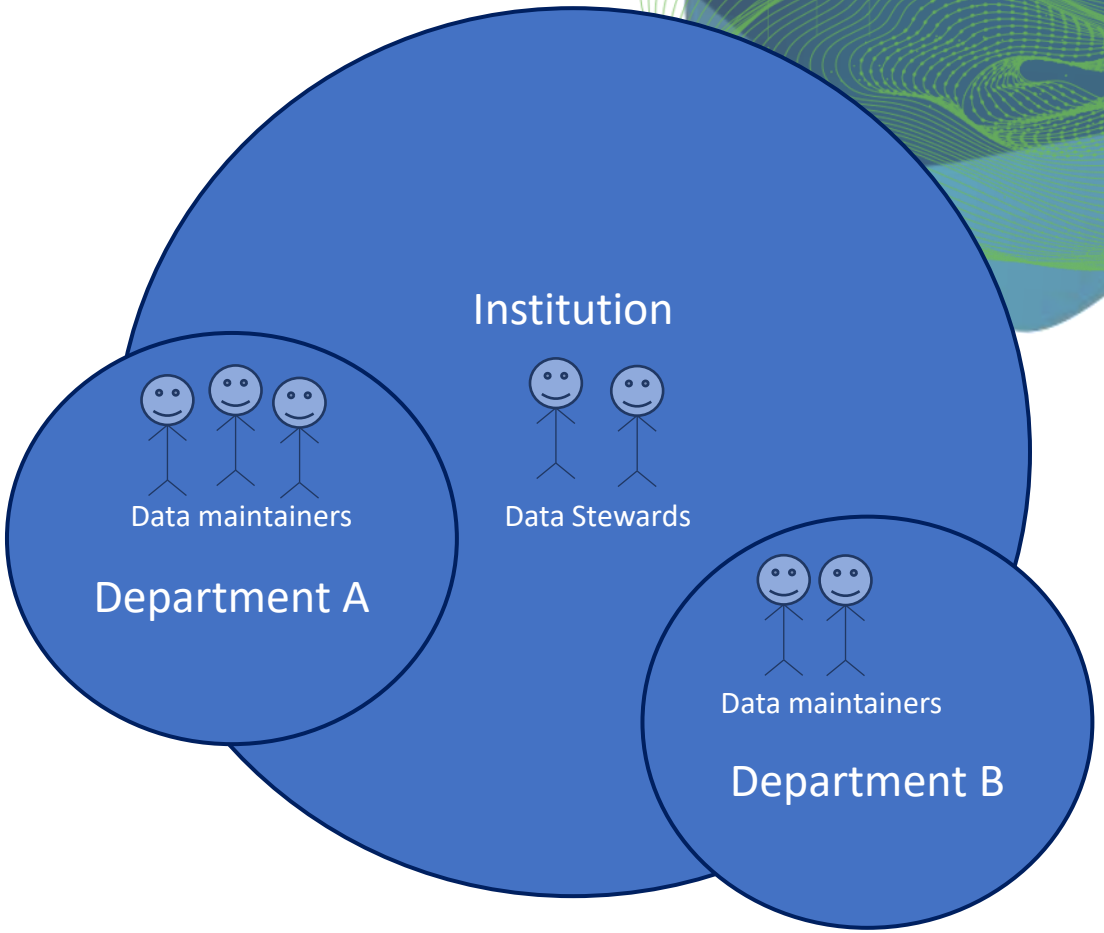
- Session gestern @ DataWeek Leipzig



# 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

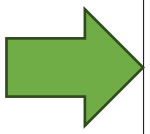
# Datenmanagmentplan (DMP)

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

# Datenmanagmentplan (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 acquistion, 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 manuscipt using overleaf, an online service for collaborative manuscipt 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, garantees 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 ≠ Job-Profile
- Datenmanagementplaene dienen dazu Projekte vorab zu durchdenken und nachvollziehbar zu dokumentieren.

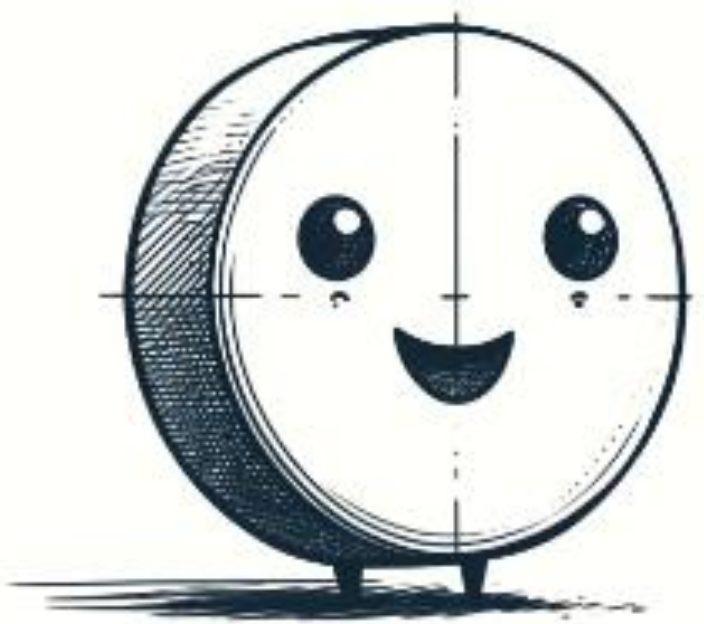


Figure made by Dall-E 3

# Acknowledgements


Communities & platforms

image.sc

NFDI4 BIOIMAGE

GloBIAS Global BioImage Analysts' Society

neubias network of european bioimage analysts

FocalPlane Where biology meets microscopy

SaxFDM

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

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.

Chan Zuckerberg Initiative

DFG Deutsche Forschungsgemeinschaft

Some Figures were generated using DALL·E