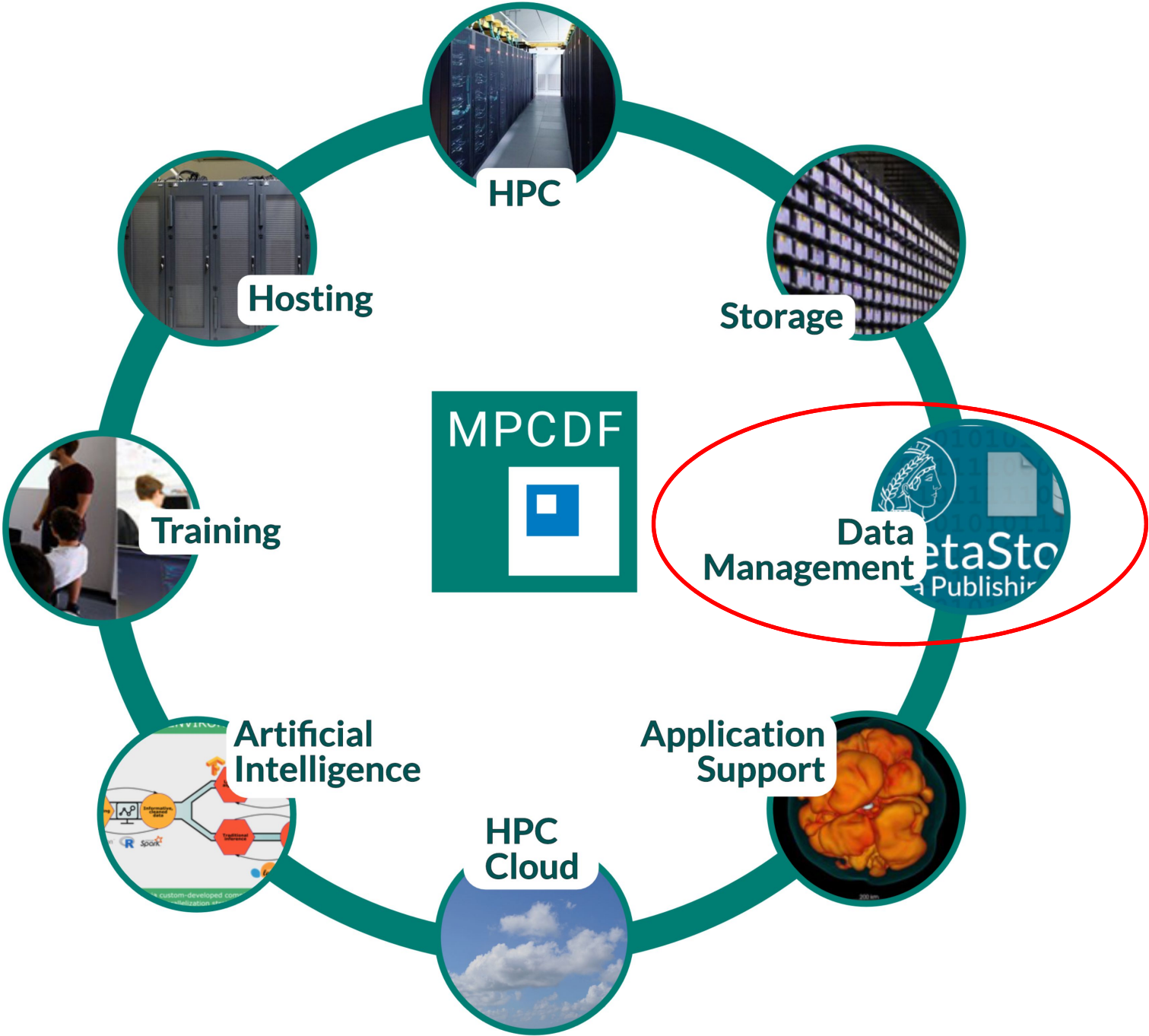




## 6. FDM Workshop der MPDL

Heidelberg 21.3.2024  
Thomas Zastrow, Nicolas Fabas



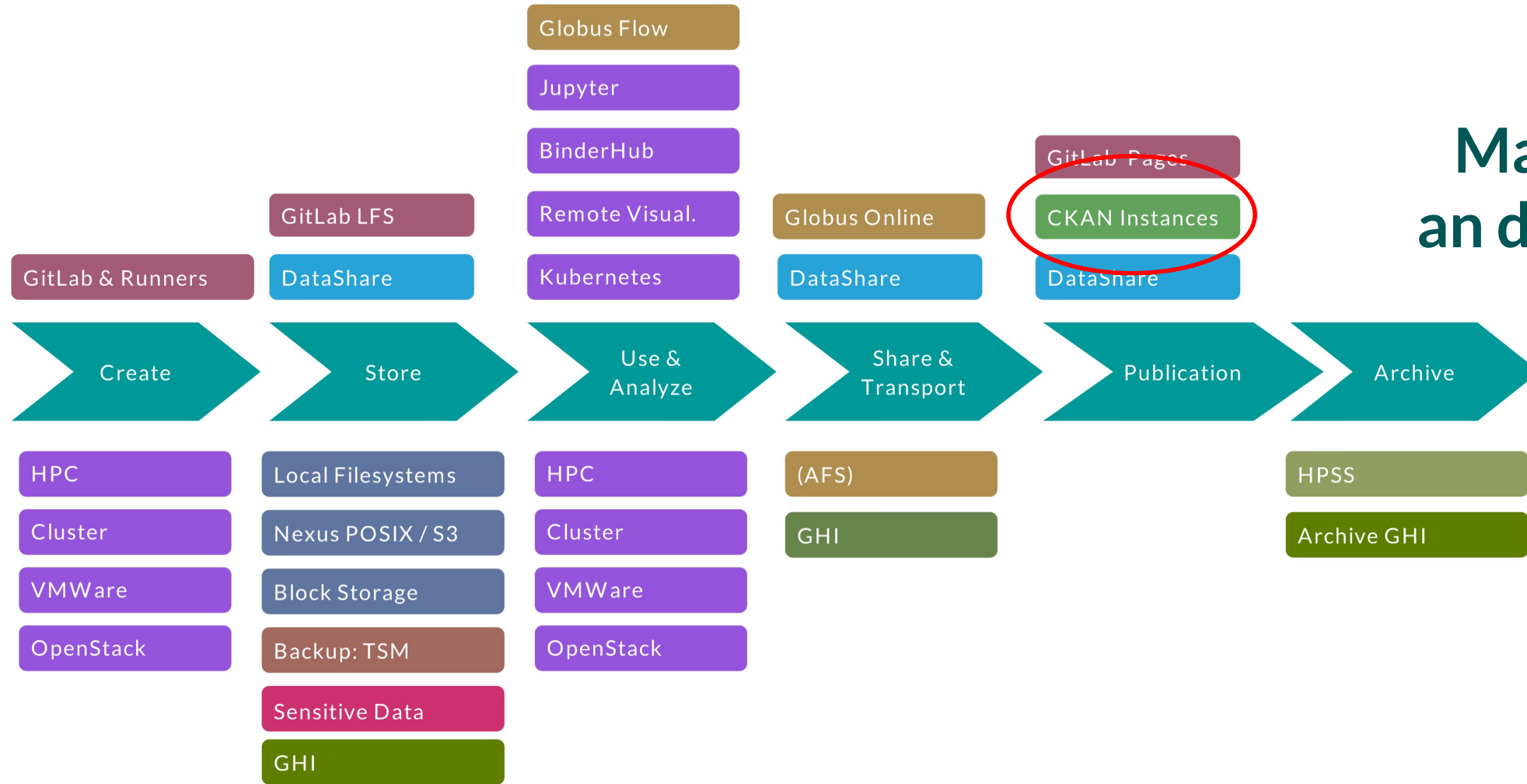
# Überblick Tätigkeiten der MPCDF

AAI

Metadata Management



# Data Management an der MPCDF



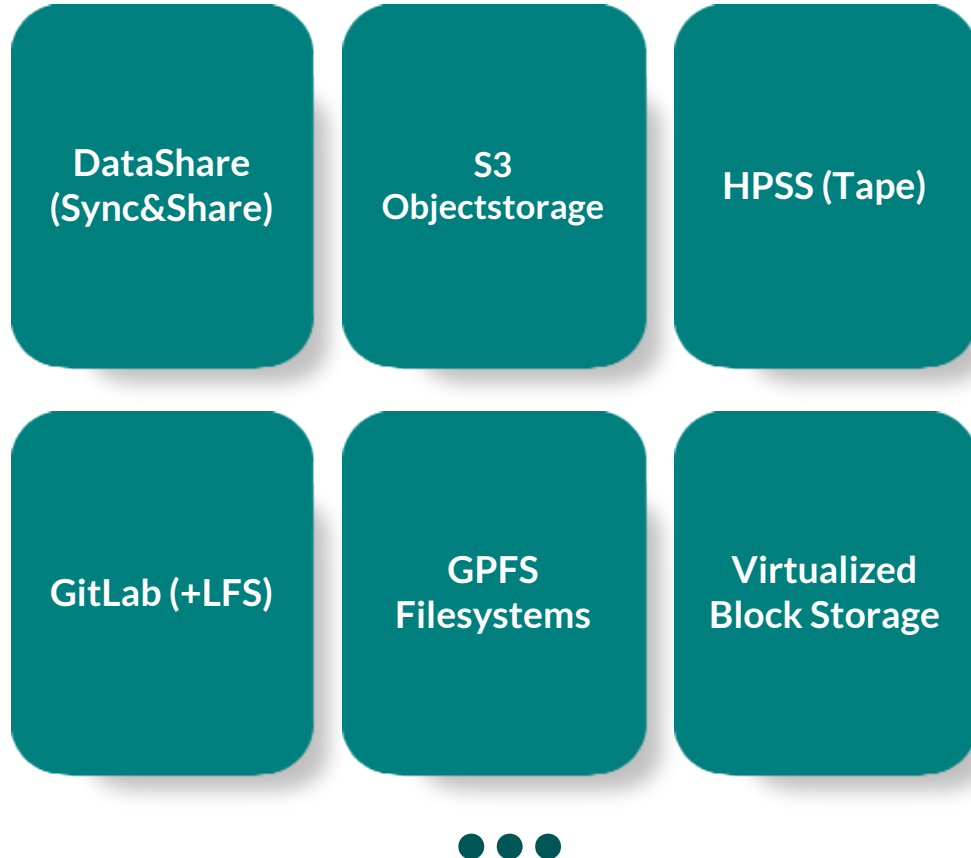
Network Capabilities

# Forschungsdaten



- An der MPCDF werden auf verschiedenste Weisen Forschungsdaten erzeugt
  - Simulationen auf den HPC Systemen
  - Analyse von Beobachtungsdaten
  - ...
- Naturgemäß haben diese Forschungsdaten einen heterogenen Charakter:
  - Unterschiedlichste Datenformate
  - Von klein bis ganz groß
  - ...

# Unterschiedlichste Storage Systeme



**Wie können an der MPCDF  
gespeicherte Forschungsdaten  
publiziert werden?**

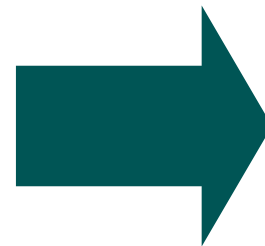




# Lösung: Datenrepository



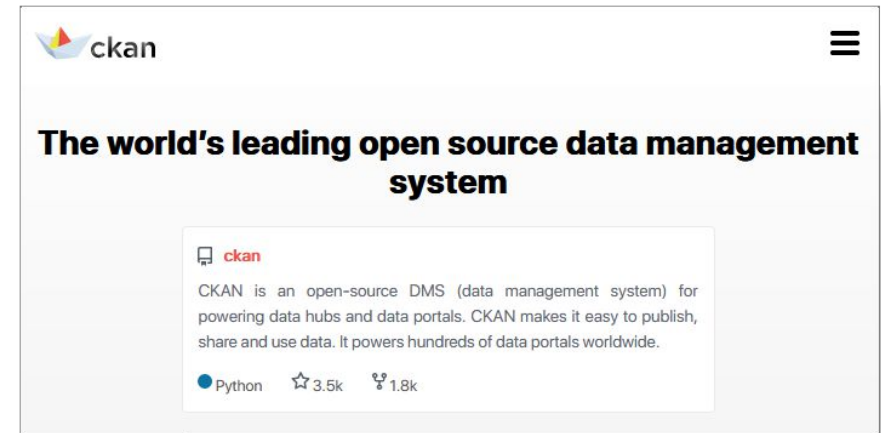
- Ein Datenrepository speichert Daten und die zugehörigen Metadaten
- Ist durchsuchbar
- Regelt den Zugriff auf Daten und Metadaten
- Stellt Verlinkungen (auch persistent) her
- (Bietet Konvertierungen an)
- ...



**FAIR**  
Data Principles

# Welche Repository Software?

- An der MPCDF: Entscheidung für CKAN
  - Open Source (Python)
  - Breitgefächerter Einsatz in Verwaltungen, Unternehmen etc.
  - Kann Meta- und Objektdaten speichern
  - Erweiterbar via Plugins
  - Vollständige Steuerung mittels REST API möglich
  - ...



<https://ckan.org/>



# CKAN basierte Portale



The screenshot shows the 'muenchen.de' website, which is the official city portal for Munich. It features a navigation bar with 'Übersicht', 'Datensätze', 'Organisationen', and 'Gruppen'. The main content area includes a welcome message 'Willkommen beim Open-Data-Portal München' with a cityscape image, and a search section titled 'Daten durchsuchen' with a search box containing 'z.B. Bierpreise'.

The screenshot shows the 'opendata.swiss' website. It has a navigation bar with 'Data', 'Organizations', 'Showcases', 'Contact', and 'About'. The main banner features a map of Switzerland and the text 'Find Swiss Open Government data' with a large '6,380 Datasets' count. A search box is present with the text 'Search datasets...' and a link to 'Learn more about opendata.swiss'.

The screenshot shows the 'EUDAT' website. It has a navigation bar with 'GO TO EUDAT WEBSITE', 'GUIDELINES', 'COMMUNITIES', 'FACETED SEARCH', 'SEARCH GUIDE', and 'ABOUT'. The main content area features a large search box with the text 'Search data' and a search box containing 'eg. IPCC'. The EUDAT logo is prominently displayed.

The screenshot shows the 'data.gov' website. It has a navigation bar with 'DATA', 'TOPICS', 'RESOURCES', 'STRATEGY', 'DEVELOPERS', and 'CONTACT'. The main content area features the text 'The home of the U.S. Government's open data' and a search box containing 'Health Care Provider Charge Data'. There are also statistics for 'Open-Data-Portal München' showing 173 Datensätze, 10 Organisationen, and 1 Gruppen.

The screenshot shows a list of categories on the data.gov website. The categories and their counts are: Prices (54), Public order and security (36), Social security (156), and Statistical basis (222). There is also a partial category 'Services' with a count of 131.

The screenshot shows the 'HIGHLIGHTS' section of the data.gov website. It features a title 'Rivers of Data - Inland Electronic Navigation Charts' and a small map image. The text describes how nautical charts provide critical information to mariners and how modern communications systems allow for electronic charts that can be updated as new information becomes available. It mentions that the NOAA Office of Coast Survey produces charts for coastal and Great Lakes areas, and the U.S. Army Corps of Engineers produces charts for America's inland rivers.

# Konzept: CKAN @ MPCDF

- Sind **nicht** für den einzelnen Nutzer gedacht
- Werden in enger Zusammenarbeit mit einem Institut, Gruppe o.ä. angelegt und betreut
- Bereitstellen einer Test-Instanz
- Betreuung und Weiterentwicklung des DOI-Plugins durch die MPCDF

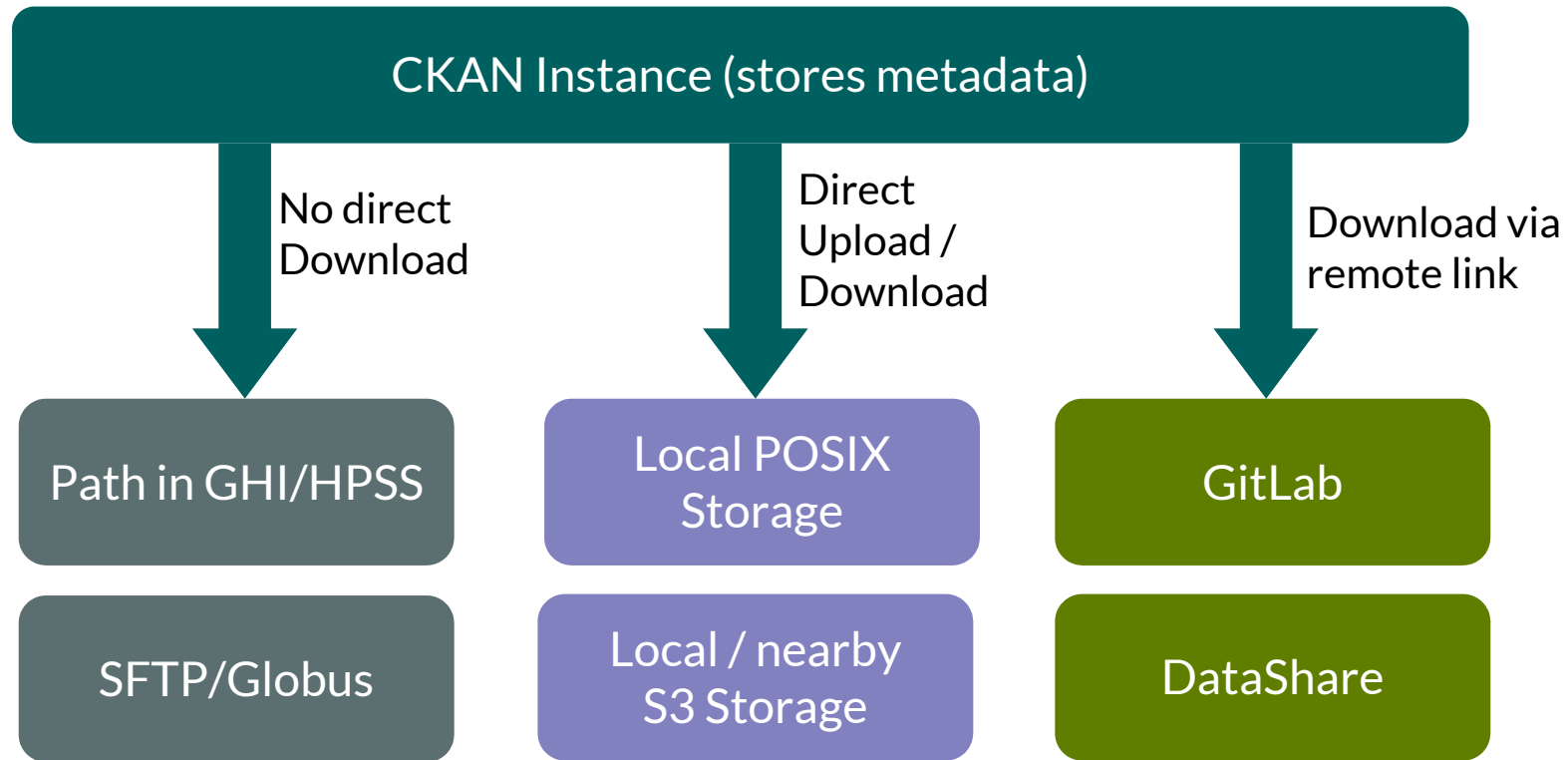


# Im Detail

- Generelle Beratung
- Erstinstallation und -konfiguration auf einer Ubuntu basierten VM
- Plugin Support:
  - Anbindung an den DOI-Service der MPDL
  - Hierarchische Strukturen
  - Anlegen individueller Metadaten-Schemata
- (Maintenance und Updates der (Basis-) Installation)



# Datenzugriff



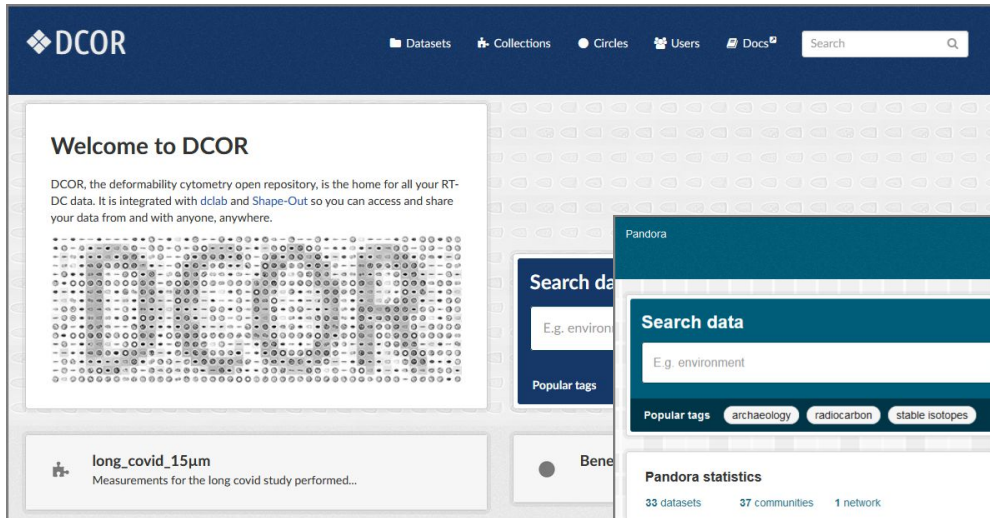
# Metadaten Schemata



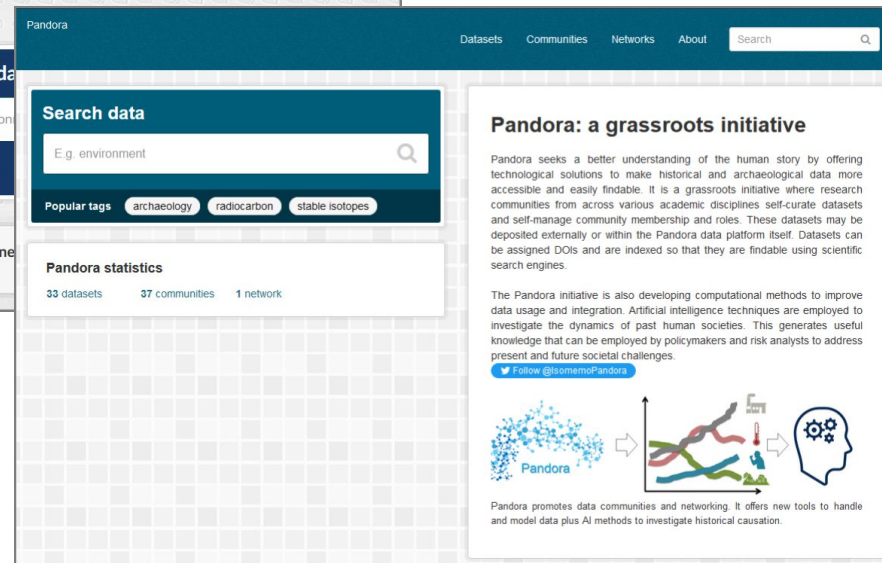
- CKAN unterstützt eine beliebige Anzahl definierter Metadatenschemata (gleichzeitig)
- Sehr komplexe Schemata möglich
- In Zukunft: Sammlung von Metadatenschemata innerhalb der MPG

# CKAN Instanzen (Stand März 24)

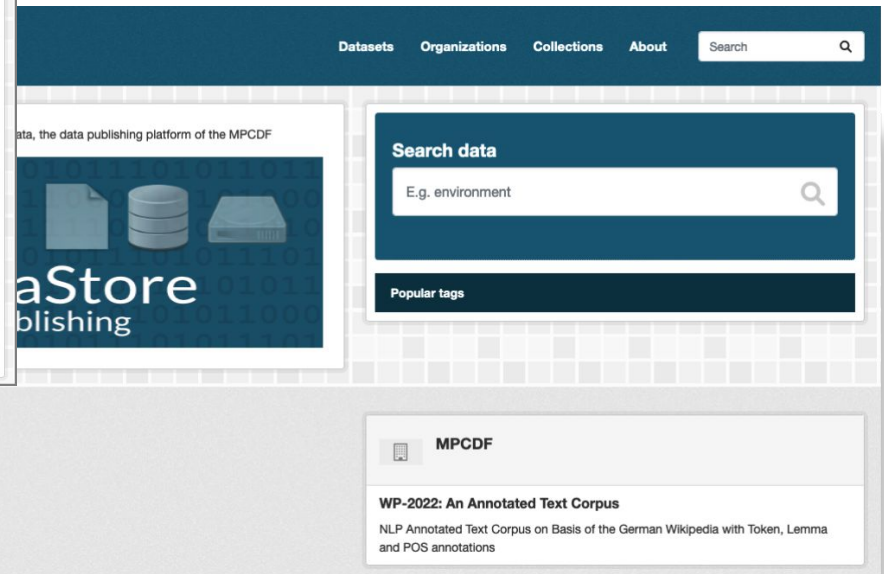
Polymerforschung und MPA: (noch) nicht öffentlich zugänglich



<https://dcor.mpl.mpg.de/>  
(MPI for the Science of Light)



<https://pandoradata.earth/>  
(MPI for Geoanthropology)

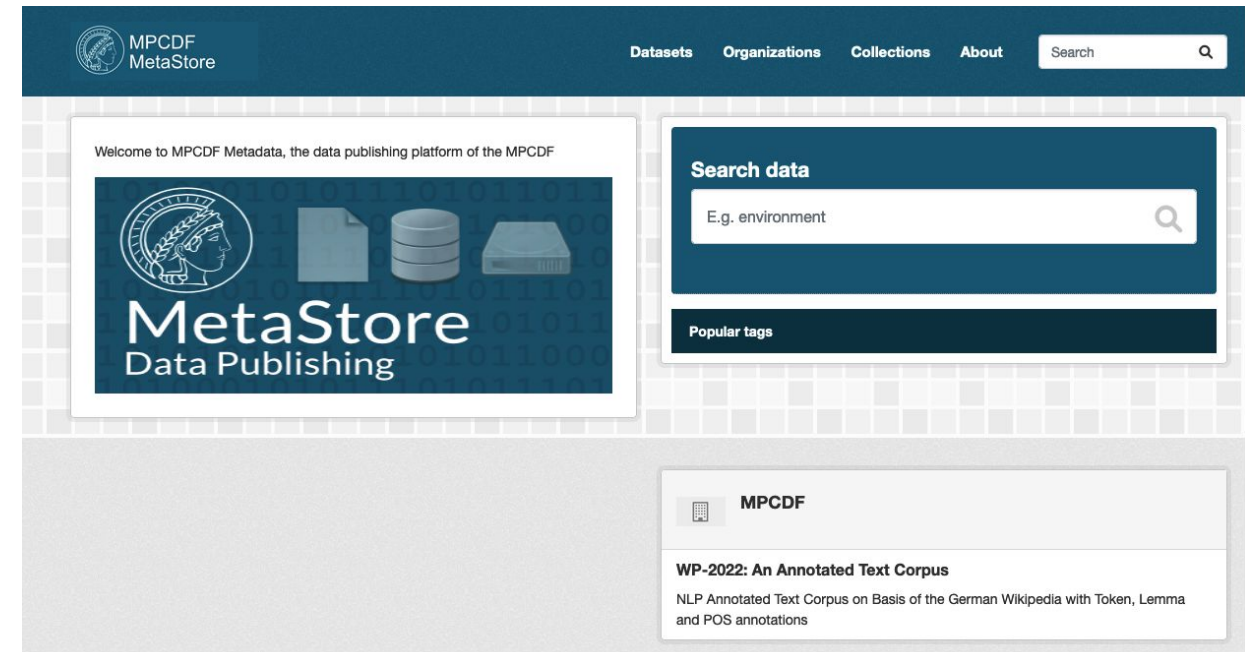


<https://metastore.mpcdf.mpg.de>



# Metastore: Catch All Instanz

- Kann von allen MPis genutzt werden
  - Einrichtung und Verantwortlichkeit für ein Institut (nicht einzelne nutzer)
- Vergabe von DOIs via MPDL
- Upload von Dateien  $\leq 1$  GB



<https://metastore.mpcdf.mpg.de>

# Metastore Metadata Schemata



- DataCite Metadata Schema Version 4.5
- Einfache Version mit 21 Metadaten Feldern („Standard“)
- Komplette Version („Extended“)

DataCite Metadata Schema

The DataCite Metadata Schema is a list of core metadata properties chosen for an accurate and consistent identification of a resource for citation and retrieval purposes, along with recommended use instructions.


Metadata Schema 4.5

Released 22 Jan 2024. Changes in this version include:

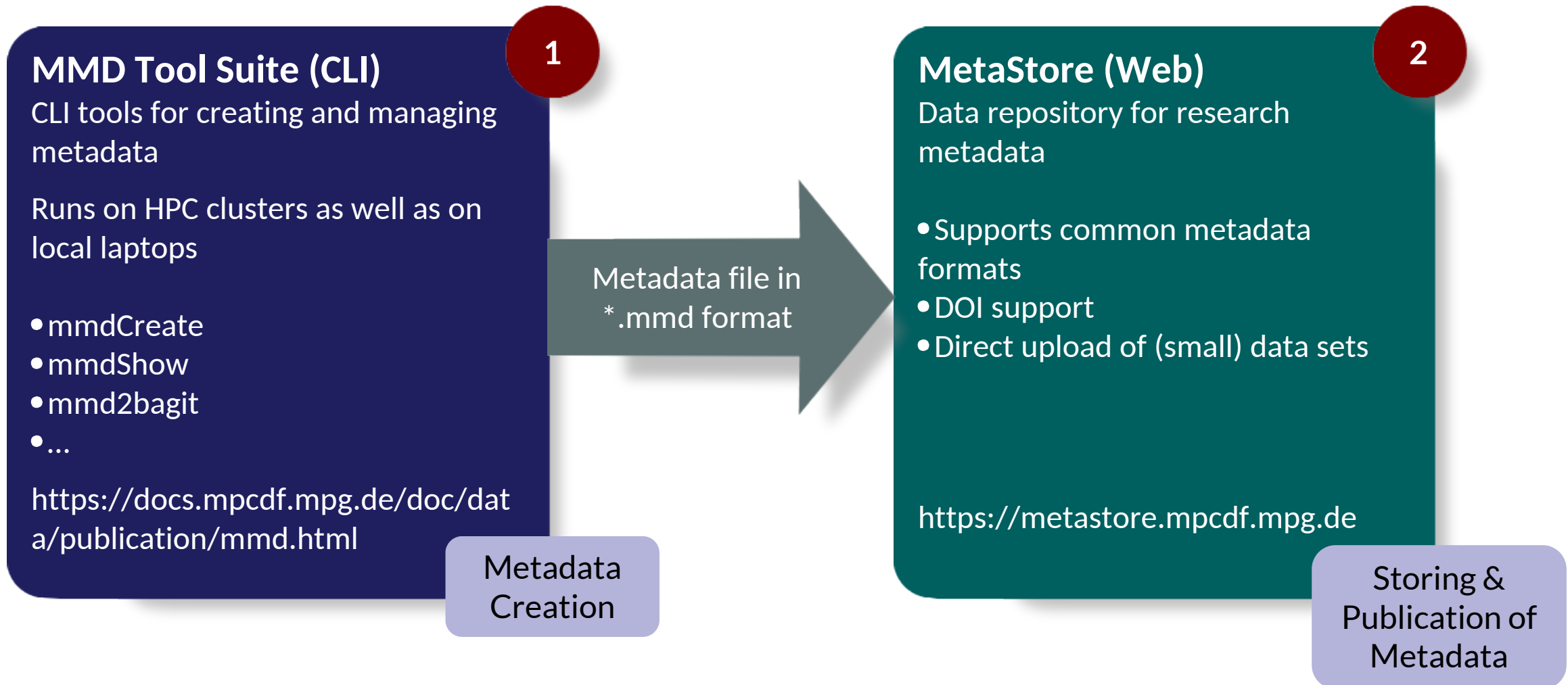
- Addition of new values to the resourceTypeGeneral property: Instrument and StudyRegistration
- Addition of new relationType pair: IsCollectedBy and Collects
- Addition of new sub-properties in the Publisher property

[More info](#)

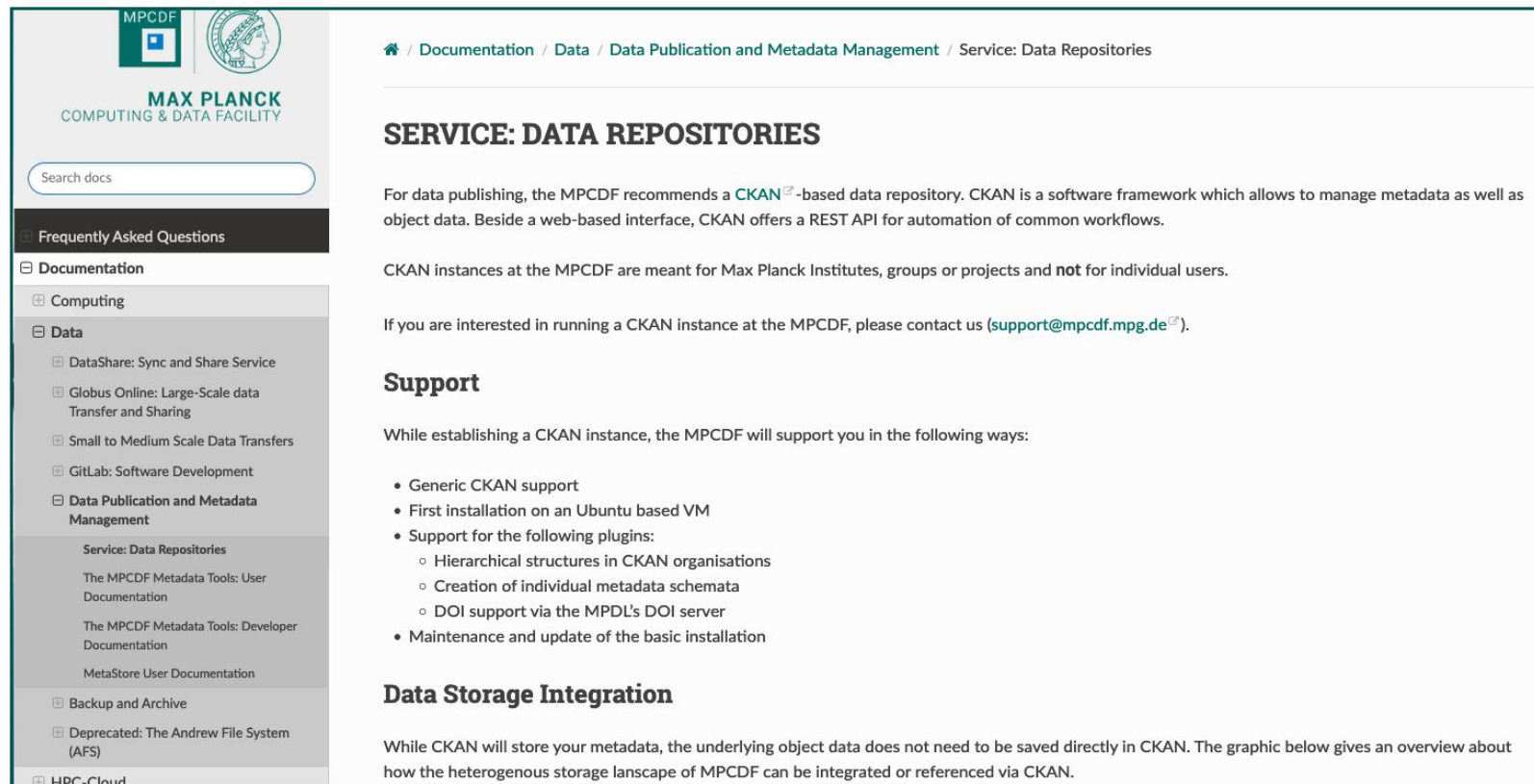
 Add standard Dataset

 Add extended Dataset

# MPCDF Metadata Tool Suite



# Weitere Informationen



The screenshot shows a web page from the MPCDF documentation. The header includes the MPCDF logo and the text 'MAX PLANCK COMPUTING & DATA FACILITY'. Below the header is a search bar labeled 'Search docs'. A navigation menu on the left lists various categories: 'Documentation', 'Computing', 'Data', 'Data Publication and Metadata Management', 'Service: Data Repositories', 'The MPCDF Metadata Tools: User Documentation', 'The MPCDF Metadata Tools: Developer Documentation', 'MetaStore User Documentation', 'Backup and Archive', 'Deprecated: The Andrew File System (AFS)', and 'HPC-Cloud'. The main content area has a breadcrumb trail: 'Home / Documentation / Data / Data Publication and Metadata Management / Service: Data Repositories'. The title is 'SERVICE: DATA REPOSITORIES'. The text explains that the MPCDF recommends a CKAN-based data repository and provides contact information for support. A 'Support' section lists the types of assistance provided, and a 'Data Storage Integration' section explains how CKAN integrates with the MPCDF's storage infrastructure.

🏠 / Documentation / Data / Data Publication and Metadata Management / Service: Data Repositories

## SERVICE: DATA REPOSITORIES

For data publishing, the MPCDF recommends a [CKAN](#)-based data repository. CKAN is a software framework which allows to manage metadata as well as object data. Beside a web-based interface, CKAN offers a REST API for automation of common workflows.

CKAN instances at the MPCDF are meant for Max Planck Institutes, groups or projects and **not** for individual users.

If you are interested in running a CKAN instance at the MPCDF, please contact us ([support@mpcdf.mpg.de](mailto:support@mpcdf.mpg.de)).

### Support

While establishing a CKAN instance, the MPCDF will support you in the following ways:

- Generic CKAN support
- First installation on an Ubuntu based VM
- Support for the following plugins:
  - Hierarchical structures in CKAN organisations
  - Creation of individual metadata schemata
  - DOI support via the MPDL's DOI server
- Maintenance and update of the basic installation

### Data Storage Integration

While CKAN will store your metadata, the underlying object data does not need to be saved directly in CKAN. The graphic below gives an overview about how the heterogenous storage lanscape of MPCDF can be integrated or referenced via CKAN.

<https://docs.mpcdf.mpg.de/doc/data/publication/datapublishing.html>



# Vielen Dank!

[thomas.zastrow@mpcdf.mpg.de](mailto:thomas.zastrow@mpcdf.mpg.de)  
[nicolas.fabas@mpcdf.mpg.de](mailto:nicolas.fabas@mpcdf.mpg.de)