MAX PLANCK
COMPUTING & DATA FACILITY

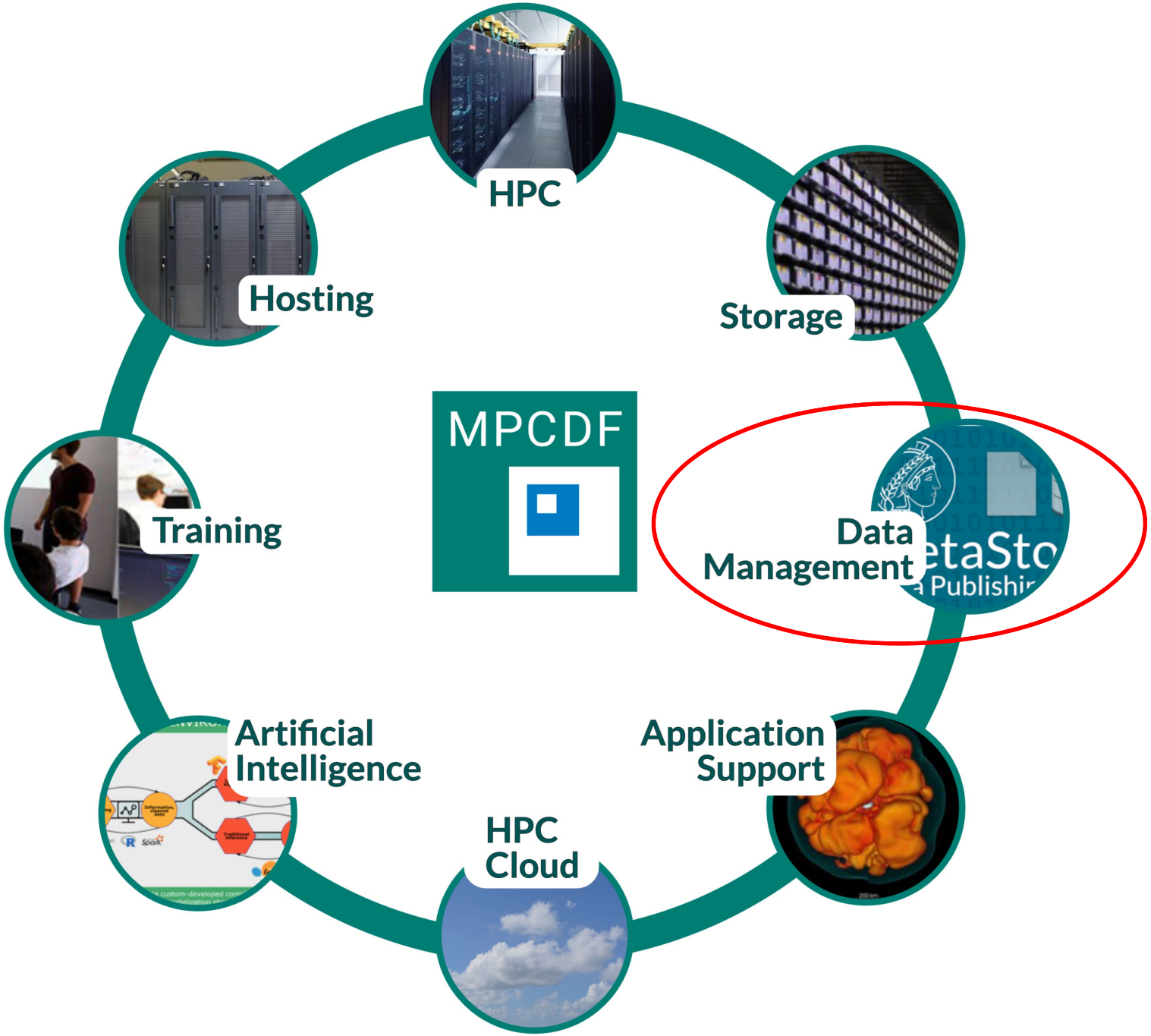
MPCDF

6. FDM Workshop der MPDL

Heidelberg 21.3.2024

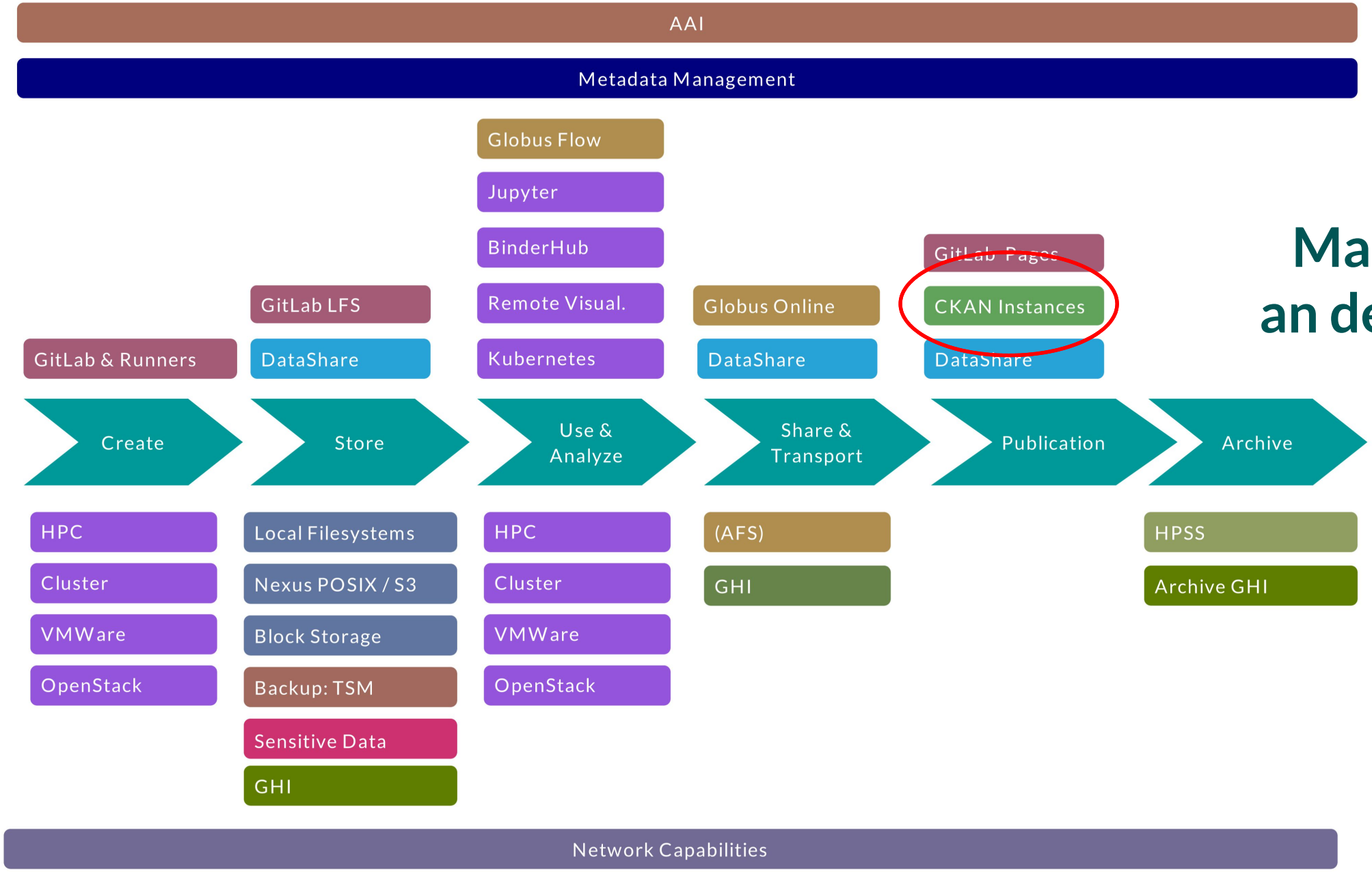
Thomas Zastrow, Nicolas Fabas

MAX PLANCK
COMPUTING & DATA FACILITY



Überblick Tätigkeiten der MPCDF

Data Management an der MPCDF



Forschungsdaten

MAX PLANCK COMPUTING & DATA FACILITY MPCDF

- 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

MAX PLANCK COMPUTING & DATA FACILITY MPCDF

DataShare (Sync&Share)

S3 Objectstorage

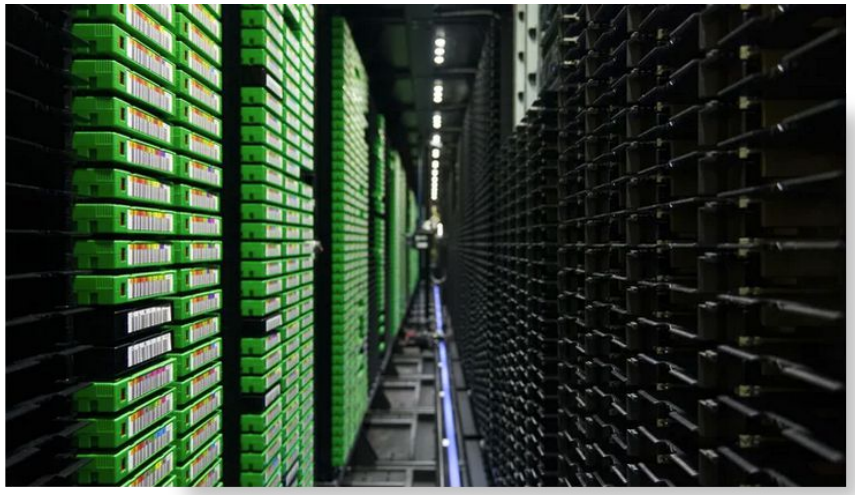
HPSS (Tape)

GitLab (+LFS)

GPFS Filesystems

Virtualized Block Storage

●●●



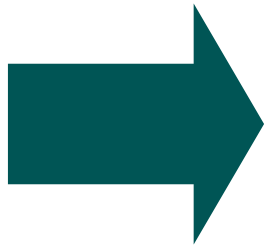
MAX PLANCK
COMPUTING & DATA FACILITY
MPCDF

Wie können an der MPCDF gespeicherten Forschungsdaten publiziert werden?

Lösung: Datenrepository

MAX PLANCK COMPUTING & DATA FACILITY MPCDF

- 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)
- ...

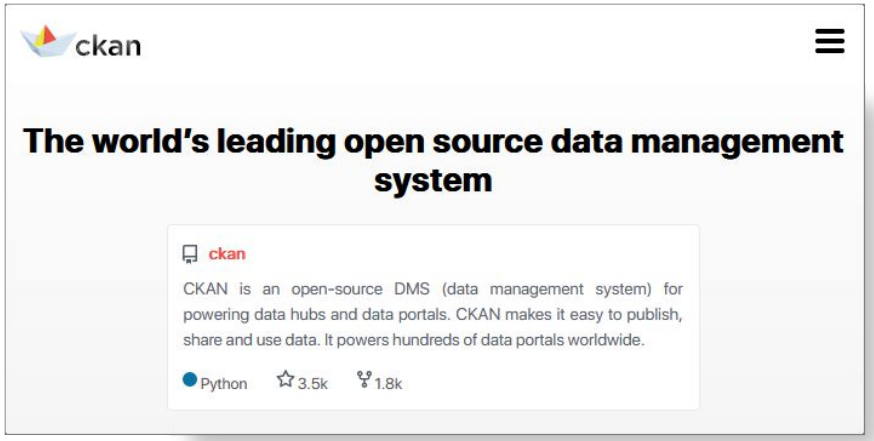


FAIR
Data Principles

Welche Repository Software?

MAX PLANCK COMPUTING & DATA FACILITY MPCDF

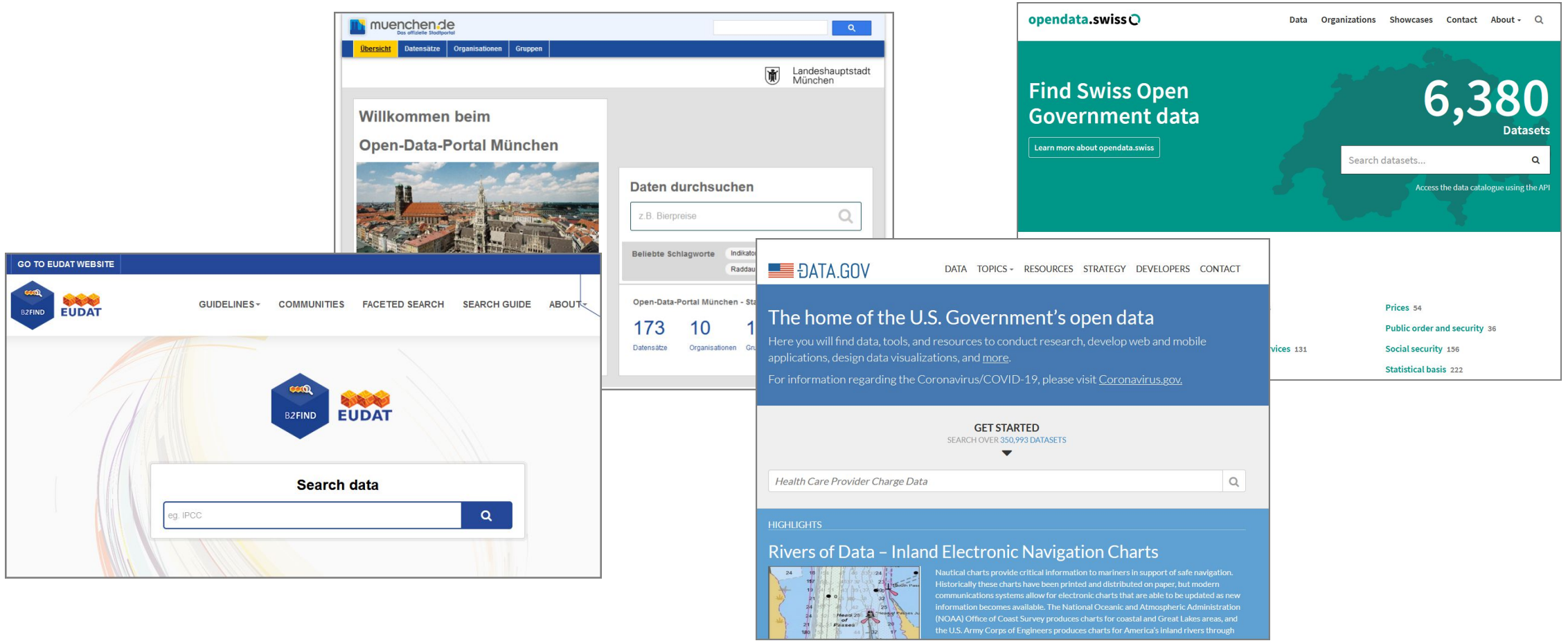
- An der MPCDF: Entscheidung für CKAN
  - Open Source (Python)
  - Breitgefächerter Einsatz in Verwaltungen, Unternehmen etc.
  - Kann Meta- und Objektdaten speichern
  - Erweiterbar via Plugins
  - Vollständige Steuerung mittels REST API möglich
  - …



https://ckan.org/

CKAN basierte Portale

MAX PLANCK
COMPUTING & DATA FACILITY
MPCDF



Konzept: CKAN @ MPCDF

MAX PLANCK COMPUTING & DATA FACILITY

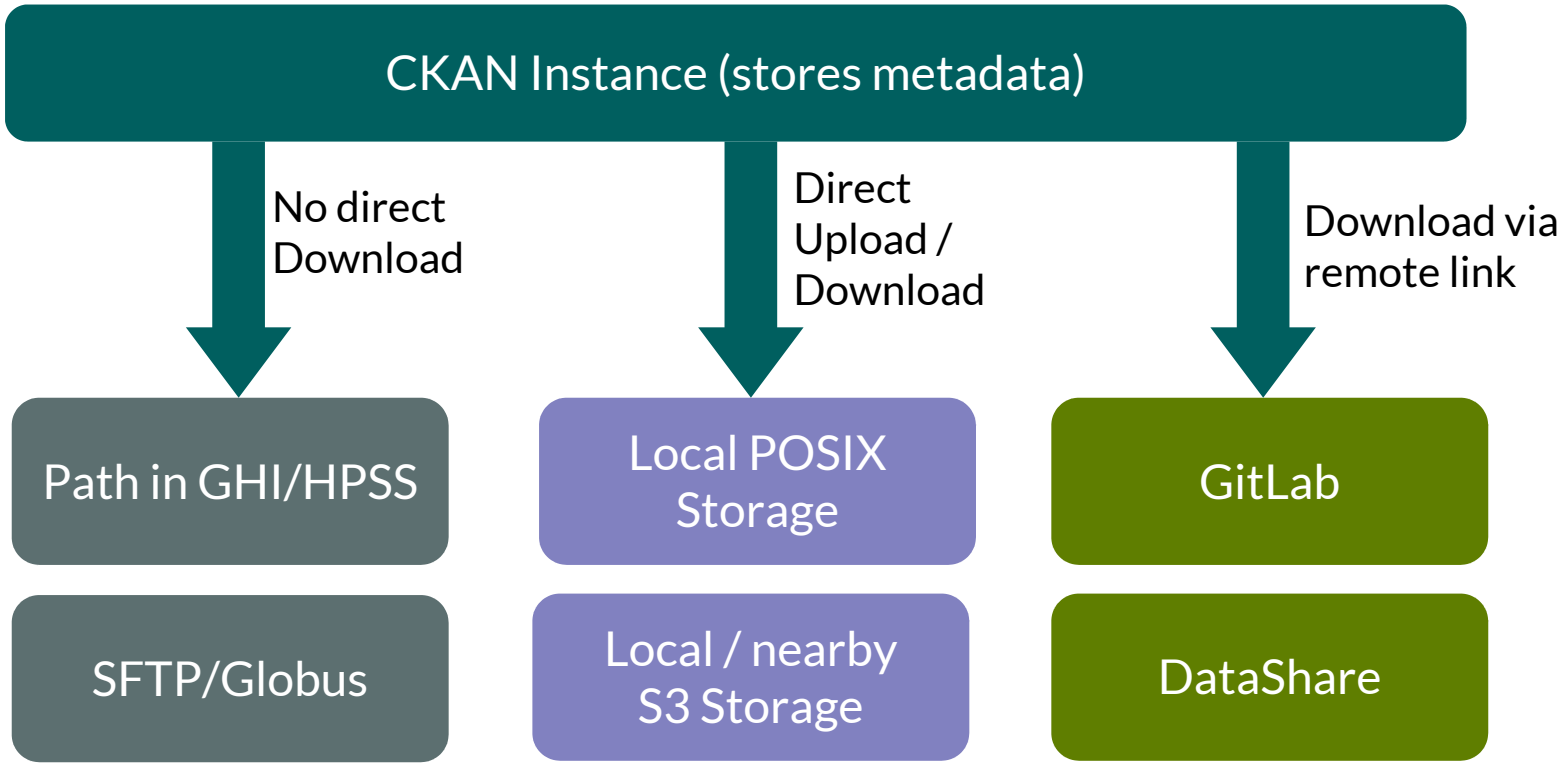
- Sind **nicht** für den einzelnen Nutzer gedacht
- Werden in enger Zusammenarbeit mit einem Institut, Gruppe o.ä. angelegt und betreut
- Bereitstellen einer Test-Instanz
- Betreuung und Weiterentwicklung des DOI-Plugins durch die MPCDF

Im Detail

MAX PLANCK COMPUTING & DATA FACILITY MPCDF

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

Datenzugriff



Metadaten Schemata

MAX PLANCK COMPUTING & DATA FACILITY MPCDF

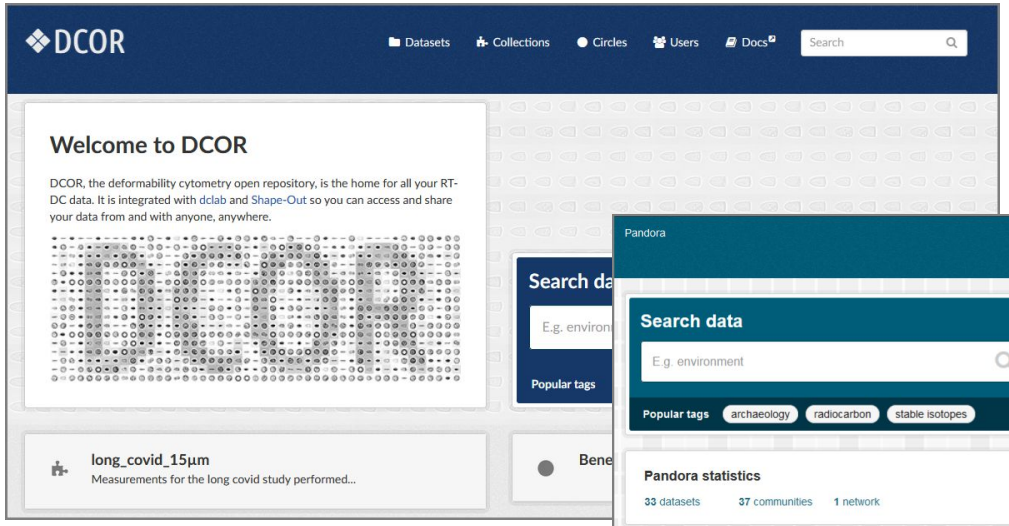
- CKAN unterstützt eine beliebige Anzahl definierter Metadatenschemata (gleichzeitig)
- Sehr komplexe Schemata möglich

- In Zukunft: Sammlung von Metadatenschemata innerhalb der MPG

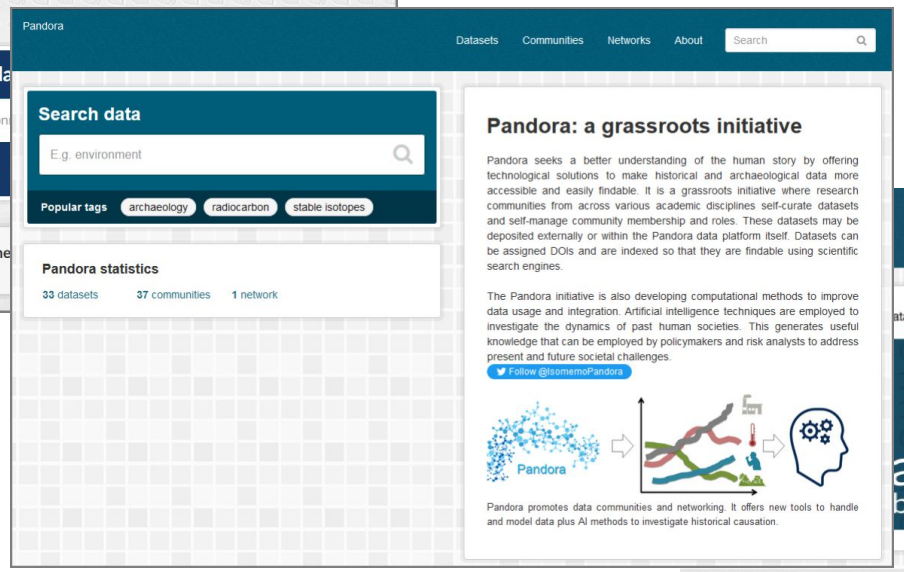
CKAN Instanzen (Stand März 24)

MAX PLANCK COMPUTING & DATA FACILITY

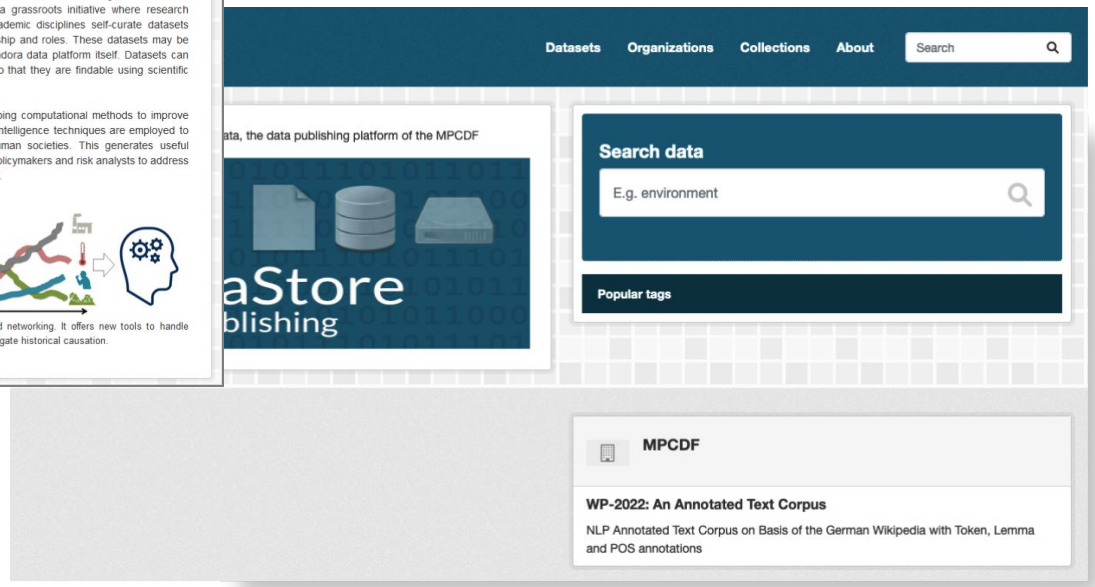
Polymerforschung und MPA: (noch) nicht öffentlich zugänglich



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



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

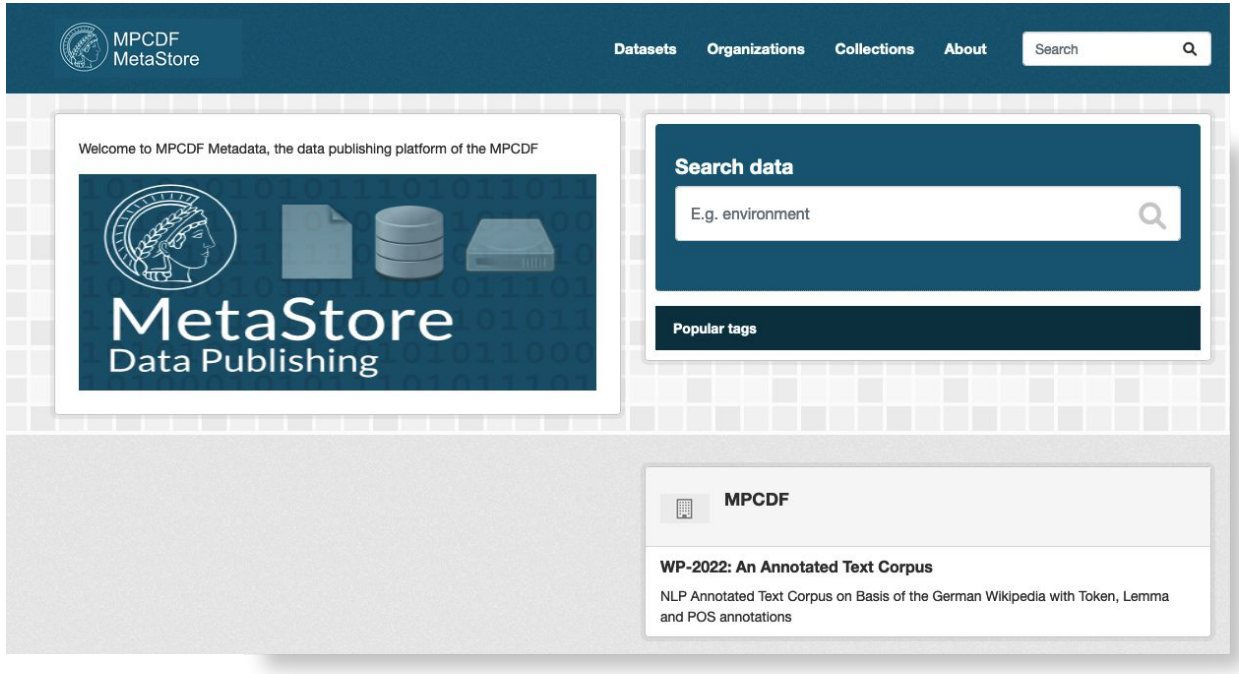


https://metastore.mpcdf.mpg.de

Metastore: Catch All Instanz

MAX PLANCK COMPUTING & DATA FACILITY

- Kann von allen MPIs genutzt werden
  - Einrichtung und Verantwortlichkeit für ein Institut (nicht einzelne nutzer)
- Vergabe von DOIs via MPDL
- Upload von Dateien <= 1 GB

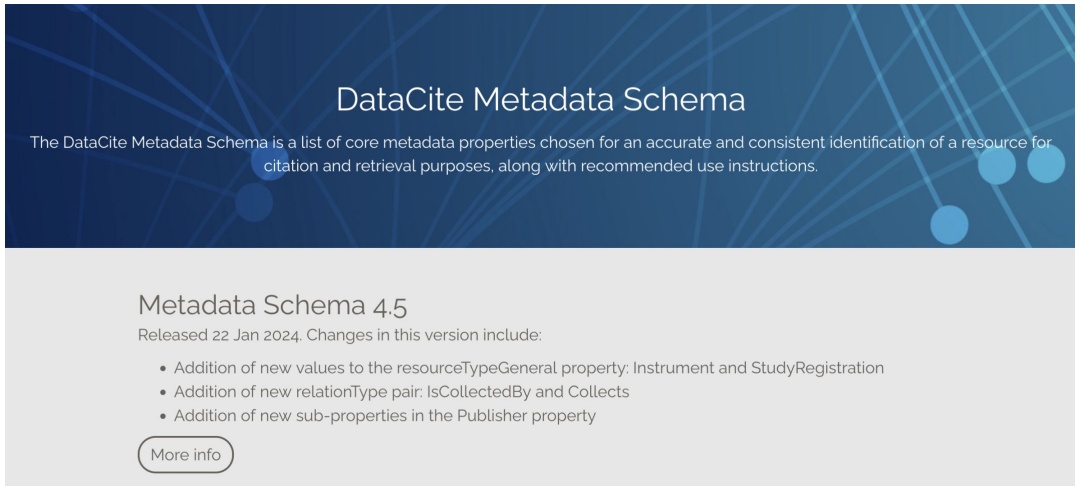


https://metastore.mpcdf.mpg.de

Metastore Metadata Schemata

MAX PLANCK COMPUTING & DATA FACILITY MPCDF

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



MPCDF Metadata Tool Suite

MAX PLANCK COMPUTING & DATA FACILITY

1

MMD Tool Suite (CLI)

CLI tools for creating and managing metadata

Runs on HPC clusters as well as on local laptops

- mmdCreate
- mmdShow
- mmd2bagit
- …

https://docs.mpcdf.mpg.de/doc/data/publication/mmd.html

Metadata Creation

Metadata file in *.mmd format

2

MetaStore (Web)

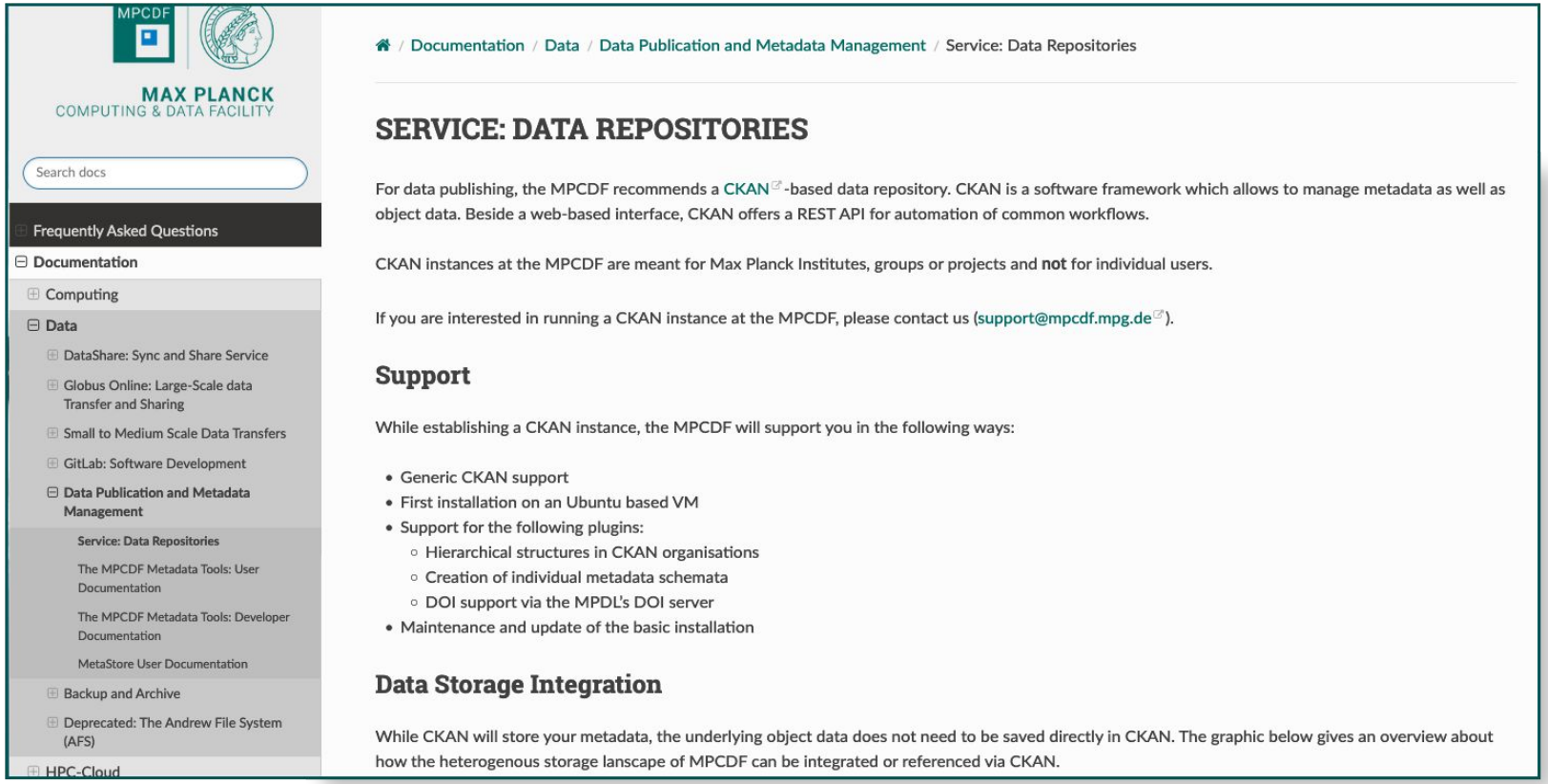
Data repository for research metadata

- Supports common metadata formats
- DOI support
- Direct upload of (small) data sets

https://metastore.mpcdf.mpg.de

Storing & Publication of Metadata

Weitere Informationen



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

MAX PLANCK
COMPUTING & DATA FACILITY
MPCDF

Vielen Dank!

thomas.zastrow@mpcdf.mpg.de
nicolas.fabas@mpcdf.mpg.de