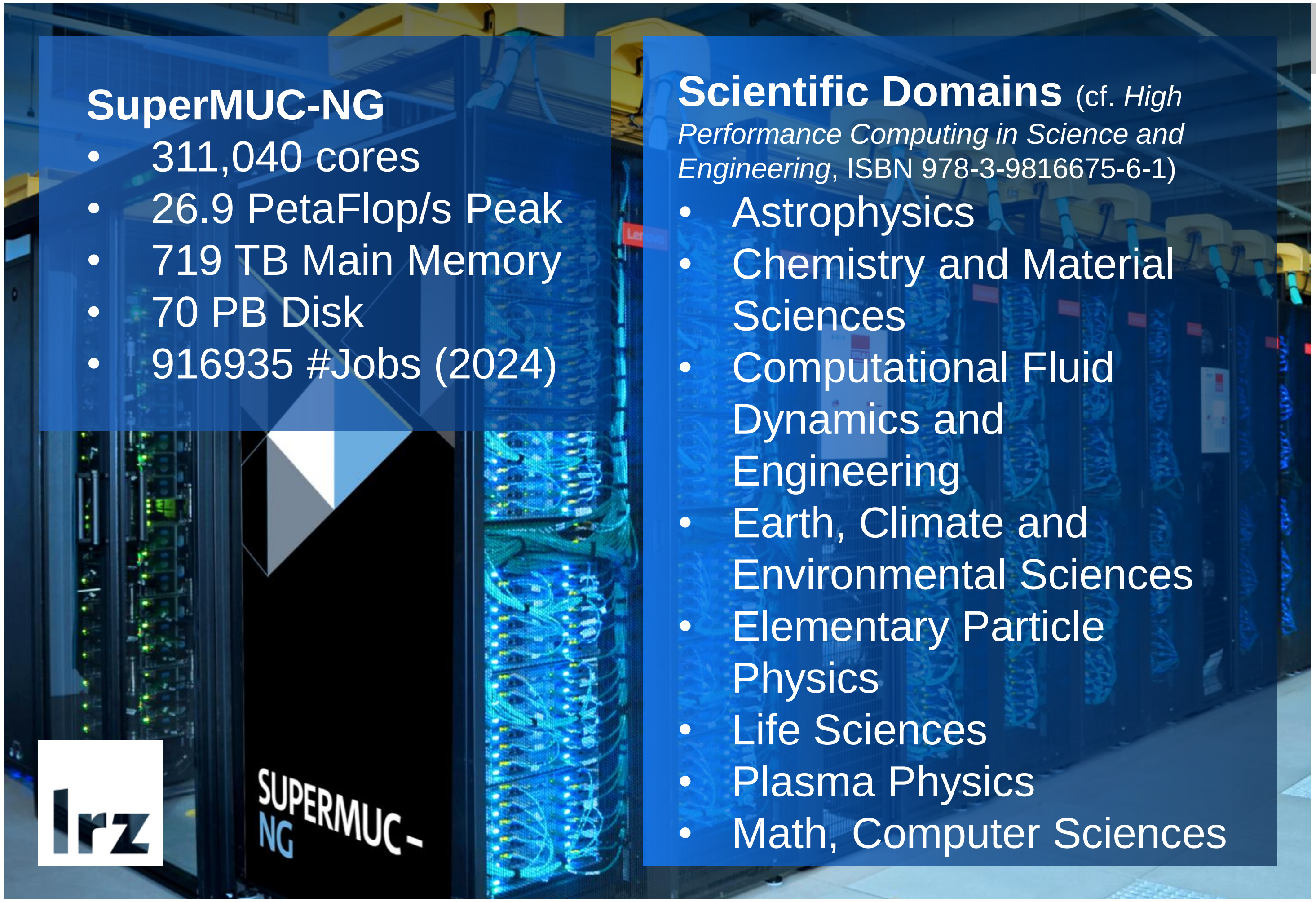
ESS Data Publication at a "General-Purpose" Supercomputing Centre

EGU25 | J. Munke, A. Wellmann, M. Muralidharan, C. Henzen, S. Hachinger

FAIR Data Portal of the Leibniz Supercomputing Centre (LRZ)

Large HPC Data Sets

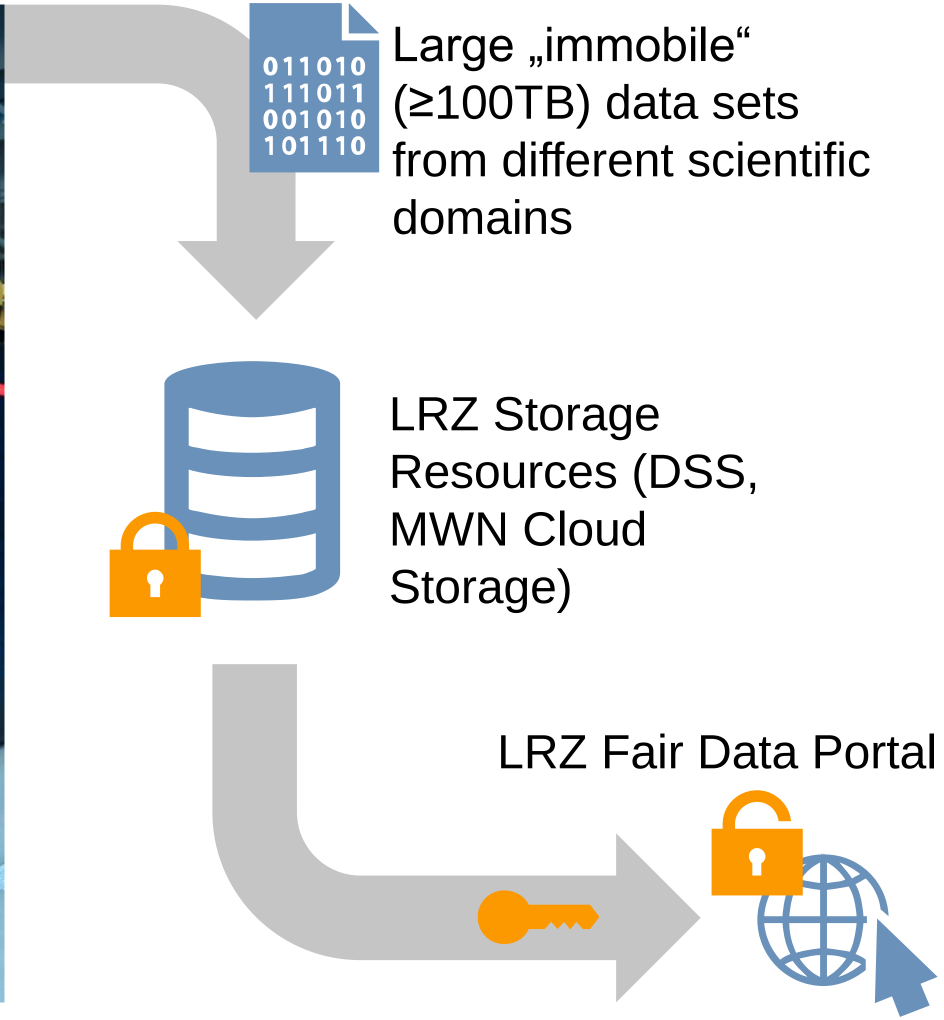


SuperMUC-NG

- 311,040 cores
- 26.9 PetaFlop/s Peak
- 719 TB Main Memory
- 70 PB Disk
- 916935 #Jobs (2024)

Scientific Domains (cf. *High Performance Computing in Science and Engineering*, ISBN 978-3-9816675-6-1)

- Astrophysics
- Chemistry and Material Sciences
- Computational Fluid Dynamics and Engineering
- Earth, Climate and Environmental Sciences
- Elementary Particle Physics
- Life Sciences
- Plasma Physics
- Math, Computer Sciences



Large „immobile“ (≥100TB) data sets from different scientific domains

LRZ Storage Resources (DSS, MWN Cloud Storage)

LRZ Fair Data Portal

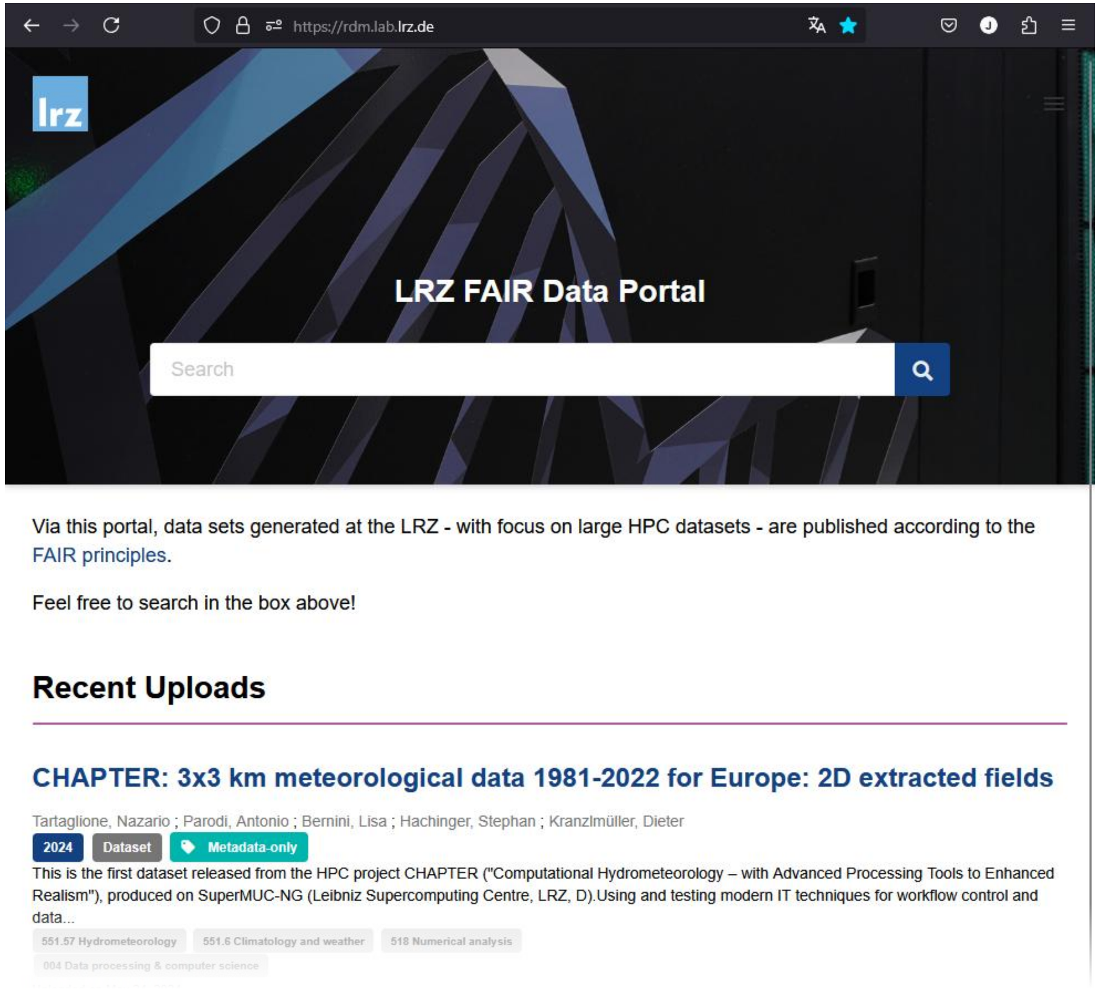
FAIR Data Portal of the Leibniz Supercomputing Centre (LRZ)

Concept / Prototype

- **Datasets** remain in their **original storage location**
- **Metadata** is **automatically extracted** and made **publicly accessible** and **searchable by push into a RDM-portal framework**
- **DOIs** are assigned
- **Data retrieval** is handled separately, via links/methods **put into** the **metadata**

Frontend: Based on InvenioRDM (developed by CERN & partners)
Backend: In-house development (Python, GitLab CI/CD, Docker, VMWare, PostgreSQL, Celery)

rdm.lab.lrz.de



LRZ FAIR Data Portal

Via this portal, data sets generated at the LRZ - with focus on large HPC datasets - are published according to the FAIR principles.

Feel free to search in the box above!

Recent Uploads

CHAPTER: 3x3 km meteorological data 1981-2022 for Europe: 2D extracted fields

Tartaglione, Nazario ; Parodi, Antonio ; Bernini, Lisa ; Hachinger, Stephan ; Kranzlmüller, Dieter

2024 | Dataset | Metadata-only

This is the first dataset released from the HPC project CHAPTER ("Computational Hydrometeorology – with Advanced Processing Tools to Enhanced Realism"), produced on SuperMUC-NG (Leibniz Supercomputing Centre, LRZ, D).Using and testing modern IT techniques for workflow control and data...

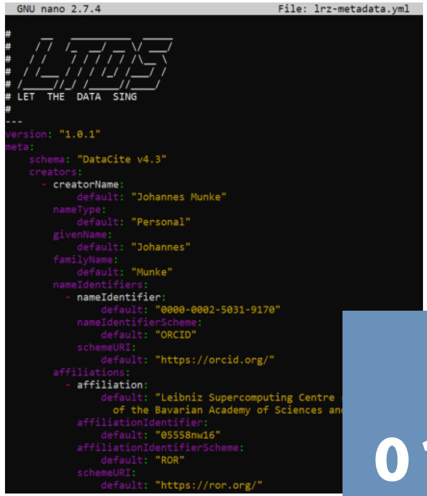
551.57 Hydrometeorology | 551.6 Climatology and weather | 518 Numerical analysis | 004 Data processing & computer science

RDM for Huge Datasets – LRZ Flavour of Harmonised GCS JSC/HLRS/LRZ Workflow

Workflow by InHPC-DE Project / Gauss Centre for Supercomputing

Manual metadata input

Automatized metadata input



```
GNU nano 2.7.4                                    File: lrz-metadata.yml

#
#
#
#
#
# LET  THE  DATA  SING
#
---
version: "1.0.1"
meta:
    schema: "DataCite v4.3"
    creators:
      - creatorName:
            default: "Johannes Munke"
        nameType:
            default: "Personal"
        givenName:
            default: "Johannes"
        familyName:
            default: "Munke"
        nameIdentifiers:
          - nameIdentifier:
                default: "0000-0002-5031-9170"
            nameIdentifierScheme:
                default: "ORCID"
            schemeURI:
                default: "https://orcid.org/"
        affiliations:
          - affiliation:
                default: "Leibniz Supercomputing Centre
                  of the Bavarian Academy of Sciences an
            affiliationIdentifier:
                default: "05558nw16"
            affiliationIdentifierScheme:
                default: "ROR"
            schemeURI:
                default: "https://ror.org/"
```

Default / fixed metadata

Sidecar YAML file

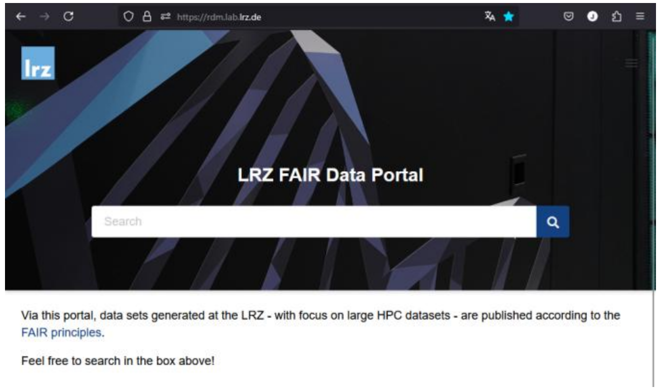
Shared dataset

011010
111011
001010
101110

„Datastore pipeline“ **E**xtract, **T**ransform, **L**oad

[1] DataCite FIND, ACCESS, AND REUSE DATA

DOI minting



Via this portal, data sets generated at the LRZ - with focus on large HPC datasets - are published according to the FAIR principles.

Feel free to search in the box above!

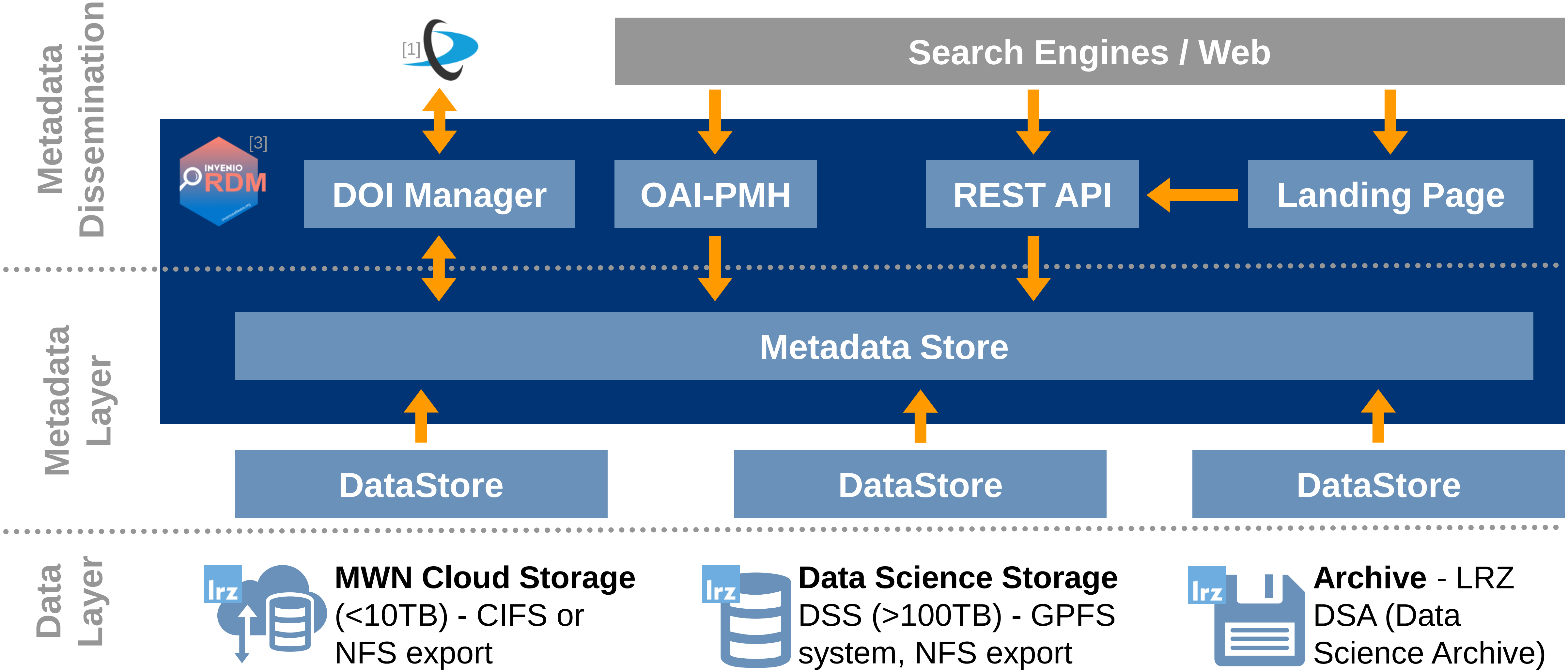
[2] globus

Link to data

[1] https://de.wikipedia.org/wiki/DataCite#/media/Datei:DataCite_logo.png – [2] https://marketing.globuscs.info/production/strapi/uploads/Globus_Logo_88c8b619be.zip

FAIR Data Portal of the Leibniz Supercomputing Centre (LRZ)

Back-End Functionality



Metadata Dissemination

[1]

Search Engines / Web

[3]

DOI Manager

OAI-PMH

REST API

Landing Page

Metadata Layer

Metadata Store

DataStore

DataStore

DataStore

Data Layer

MWN Cloud Storage (<10TB) - CIFS or NFS export

Data Science Storage DSS (>100TB) - GPFS system, NFS export

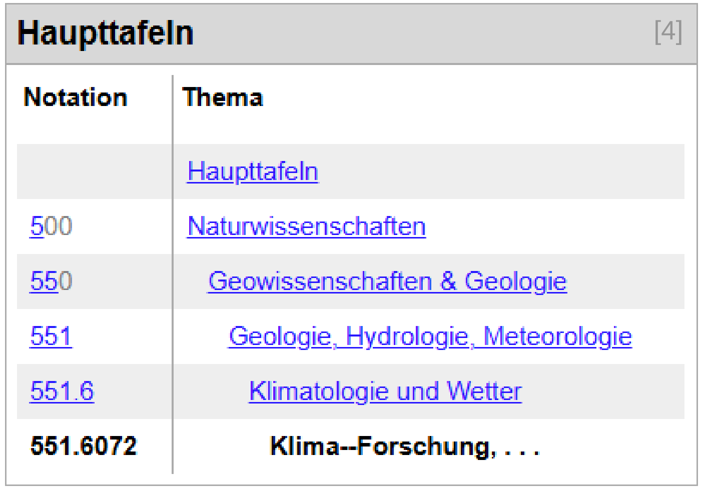
Archive - LRZ DSA (Data Science Archive)

[3] https://inveniosoftware.org/static/img/sticker-inveniordm-hex.svg?h=83bd6b07

FAIR Data Portal of the Leibniz Supercomputing Centre (LRZ)

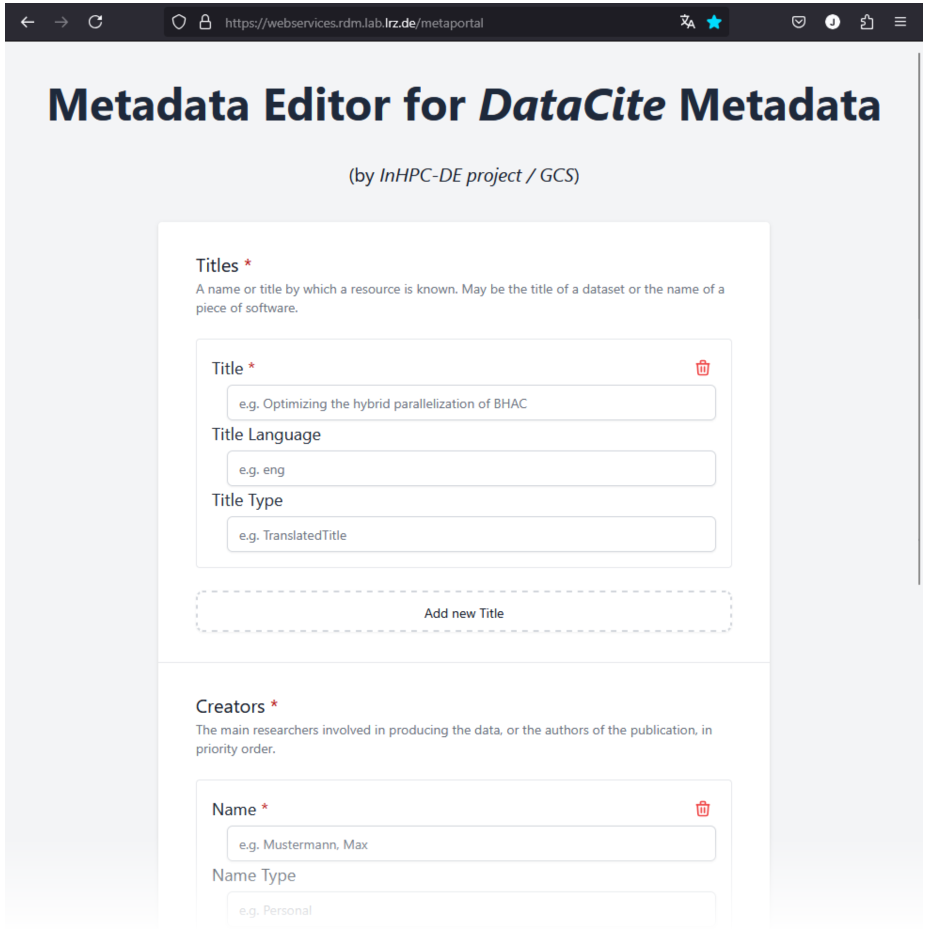
Metadata

- Web interface for easy generation of metadata sidecar files
- Standard: DataCite Metadata Schema v4.x with minor (InvenioRDM) additions and modifications
    - General purpose core metadata properties
- Subject Classification: Dewey Decimal Classification (DDC)



Haupttafeln [4]

Notation	Thema
	Haupttafeln
500	Naturwissenschaften
550	Geowissenschaften & Geologie
551	Geologie, Hydrologie, Meteorologie
551.6	Klimatologie und Wetter
551.6072	**Klima--Forschung, . . .**



https://webservices.rdm.lab.lrz.de/metaportal

Metadata Editor for *DataCite* Metadata

(by *InHPC-DE project / GCS*)

Titles *

A name or title by which a resource is known. May be the title of a dataset or the name of a piece of software.

Title *

e.g. Optimizing the hybrid parallelization of BHAC

Title Language

e.g. eng

Title Type

e.g. TranslatedTitle

Add new Title

Creators *

The main researchers involved in producing the data, or the authors of the publication, in priority order.

Name *

e.g. Mustermann, Max

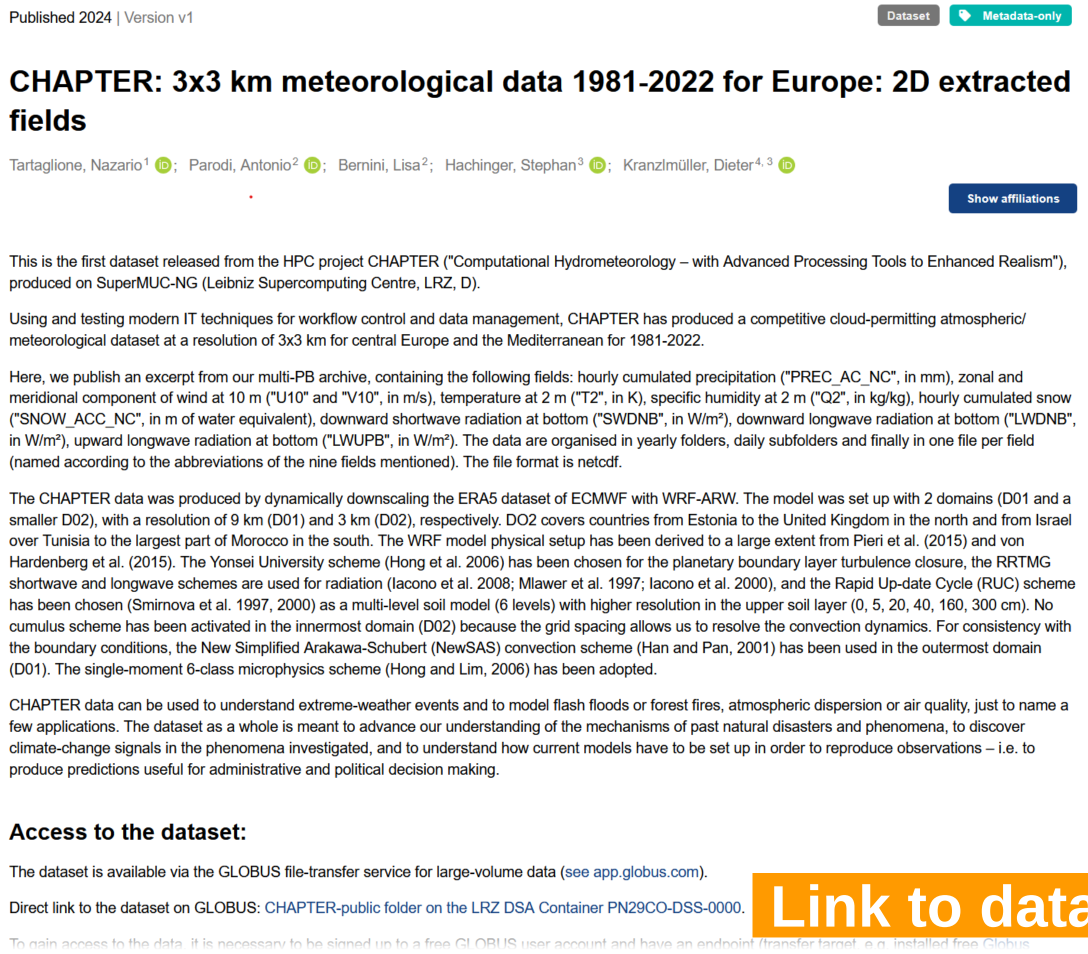
Name Type

e.g. Personal

[4] https://deweysearchde.pansoft.de/

FAIR Data Portal of the Leibniz Supercomputing Centre (LRZ)

Dataset



Published 2024 | Version v1

Dataset   Metadata-only

CHAPTER: 3x3 km meteorological data 1981-2022 for Europe: 2D extracted fields

Tartaglione, Nazario[1]; Parodi, Antonio[2]; Bernini, Lisa[2]; Hachinger, Stephan[3]; Kranzlmüller, Dieter[4, 3]

Show affiliations

This is the first dataset released from the HPC project CHAPTER ("Computational Hydrometeorology – with Advanced Processing Tools to Enhanced Realism"), produced on SuperMUC-NG (Leibniz Supercomputing Centre, LRZ, D).

Using and testing modern IT techniques for workflow control and data management, CHAPTER has produced a competitive cloud-permitting atmospheric/meteorological dataset at a resolution of 3x3 km for central Europe and the Mediterranean for 1981-2022.

Here, we publish an excerpt from our multi-PB archive, containing the following fields: hourly cumulated precipitation ("PREC_AC_NC", in mm), zonal and meridional component of wind at 10 m ("U10" and "V10", in m/s), temperature at 2 m ("T2", in K), specific humidity at 2 m ("Q2", in kg/kg), hourly cumulated snow ("SNOW_ACC_NC", in m of water equivalent), downward shortwave radiation at bottom ("SWDNB", in W/m²), downward longwave radiation at bottom ("LWDNB", in W/m²), upward longwave radiation at bottom ("LWUPB", in W/m²). The data are organised in yearly folders, daily subfolders and finally in one file per field (named according to the abbreviations of the nine fields mentioned). The file format is netcdf.

The CHAPTER data was produced by dynamically downscaling the ERA5 dataset of ECMWF with WRF-ARW. The model was set up with 2 domains (D01 and a smaller D02), with a resolution of 9 km (D01) and 3 km (D02), respectively. DO2 covers countries from Estonia to the United Kingdom in the north and from Israel over Tunisia to the largest part of Morocco in the south. The WRF model physical setup has been derived to a large extent from Pieri et al. (2015) and von Hardenberg et al. (2015). The Yonsei University scheme (Hong et al. 2006) has been chosen for the planetary boundary layer turbulence closure, the RRTMG shortwave and longwave schemes are used for radiation (Iacono et al. 2008; Mlawer et al. 1997; Iacono et al. 2000), and the Rapid Up-date Cycle (RUC) scheme has been chosen (Smirnova et al. 1997, 2000) as a multi-level soil model (6 levels) with higher resolution in the upper soil layer (0, 5, 20, 40, 160, 300 cm). No cumulus scheme has been activated in the innermost domain (D02) because the grid spacing allows us to resolve the convection dynamics. For consistency with the boundary conditions, the New Simplified Arakawa-Schubert (NewSAS) convection scheme (Han and Pan, 2001) has been used in the outermost domain (D01). The single-moment 6-class microphysics scheme (Hong and Lim, 2006) has been adopted.

CHAPTER data can be used to understand extreme-weather events and to model flash floods or forest fires, atmospheric dispersion or air quality, just to name a few applications. The dataset as a whole is meant to advance our understanding of the mechanisms of past natural disasters and phenomena, to discover climate-change signals in the phenomena investigated, and to understand how current models have to be set up in order to reproduce observations – i.e. to produce predictions useful for administrative and political decision making.

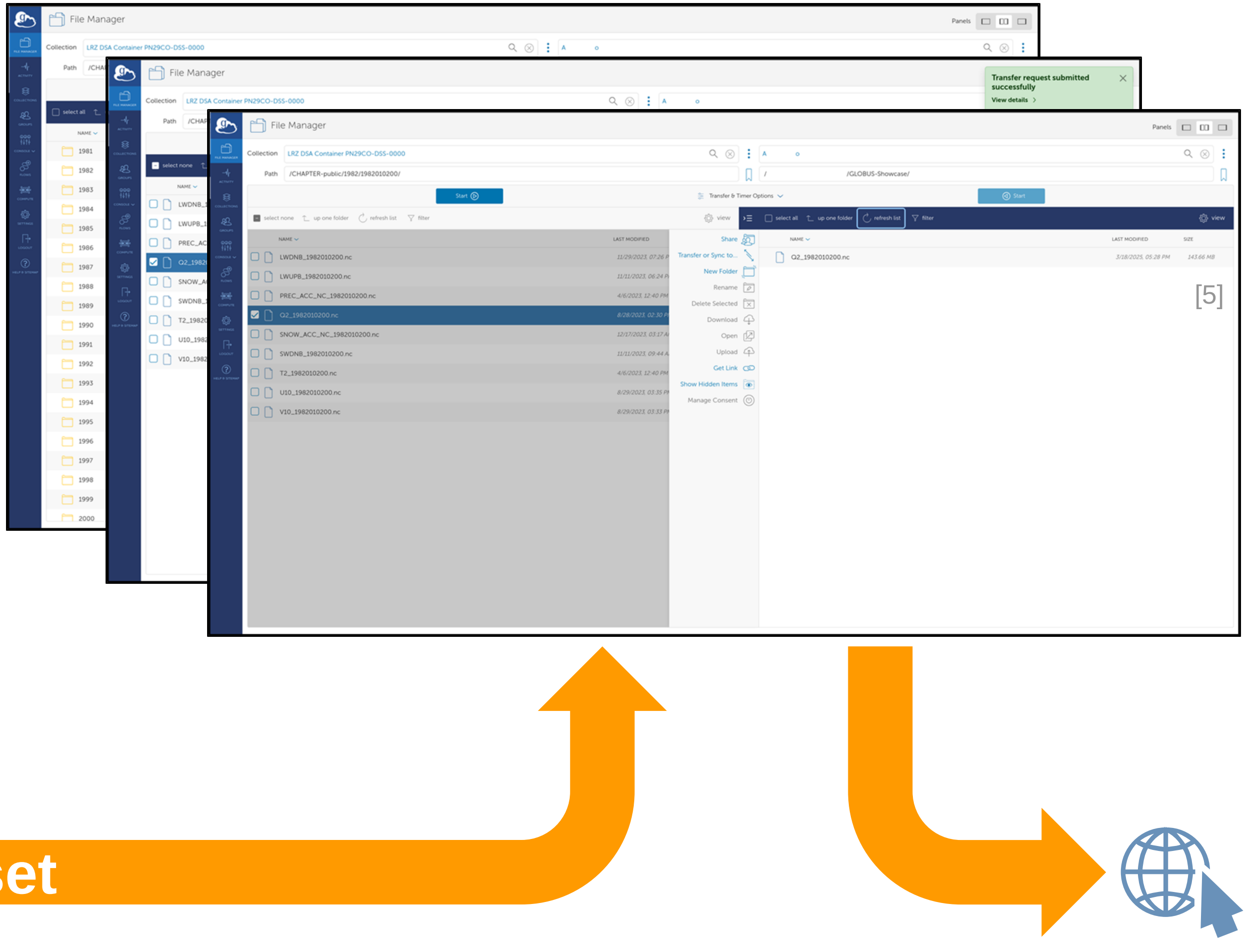
Access to the dataset:

The dataset is available via the GLOBUS file-transfer service for large-volume data (see app.globus.com).

Direct link to the dataset on GLOBUS: CHAPTER-public folder on the LRZ DSA Container PN29CO-DSS-0000.

To gain access to the data, it is necessary to be signed up to a free GLOBUS user account and have an endpoint (transfer target, e.g. installed free Globus

Link to dataset



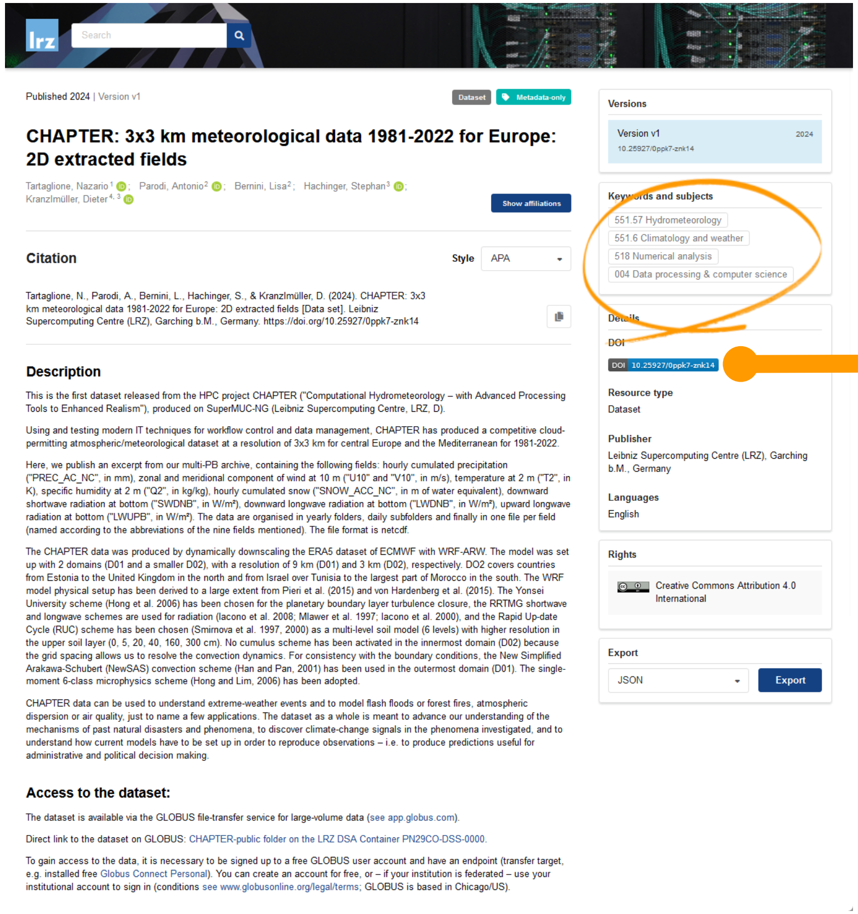
File Manager
Collection LRZ DSA Container PN29CO-DSS-0000
Path /CHAPTER-public/1982/1982010200/
Transfer request submitted successfully
LWDNB_1982010200.nc
LWUPB_1982010200.nc
PREC_ACC_NC_1982010200.nc
Q2_1982010200.nc
SNOW_ACC_NC_1982010200.nc
SWDNB_1982010200.nc
T2_1982010200.nc
U10_1982010200.nc
V10_1982010200.nc
/GLOBUS-Showcase/
Start
[5]

[5] https://app.globus.org/file-manager?origin_id=c1faeaa3-4751-4358-ae61-34e45e99afcc&origin_path=%2FCHAPTER-public%2F&two_pane=false

FAIR Data Portal of the Leibniz Supercomputing Centre (LRZ)

Harvesting

rdm.lab.lrz.de



Published 2024 | Version v1

CHAPTER: 3x3 km meteorological data 1981-2022 for Europe: 2D extracted fields

Tartaglione, Nazario[1]; Parodi, Antonio[2]; Bernini, Lisa[2]; Hachinger, Stephan[3]; Kranzlmüller, Dieter[4, 3]

Citation

Style: APA

Tartaglione, N., Parodi, A., Bernini, L., Hachinger, S., & Kranzlmüller, D. (2024). CHAPTER: 3x3 km meteorological data 1981-2022 for Europe: 2D extracted fields [Data set]. Leibniz Supercomputing Centre (LRZ), Garching b.M., Germany. https://doi.org/10.25927/0ppk7-znk14

Description

This is the first dataset released from the HPC project CHAPTER ("Computational Hydrometeorology – with Advanced Processing Tools to Enhanced Realism"), produced on SuperMUC-NG (Leibniz Supercomputing Centre, LRZ, D).

Using and testing modern IT techniques for workflow control and data management, CHAPTER has produced a competitive cloud-permitting atmospheric/meteorological dataset at a resolution of 3x3 km for central Europe and the Mediterranean for 1981-2022.

Here, we publish an excerpt from our multi-PB archive, containing the following fields: hourly cumulated precipitation ("PREC_AC_NC", in mm), zonal and meridional component of wind at 10 m ("U10" and "V10", in m/s), temperature at 2 m ("T2", in K), specific humidity at 2 m ("Q2", in kg/kg), hourly cumulated snow ("SNOW_ACC_NC", in m of water equivalent), downward shortwave radiation at bottom ("SWDNB", in W/m²), downward longwave radiation at bottom ("LWDNB", in W/m²), upward longwave radiation at bottom ("LWUPB", in W/m²). The data are organised in yearly folders, daily subfolders and finally in one file per field (named according to the abbreviations of the nine fields mentioned). The file format is netcdf.

The CHAPTER data was produced by dynamically downscaling the ERA5 dataset of ECMWF with WRF-ARW. The model was set up with 2 domains (D01 and a smaller D02), with a resolution of 9 km (D01) and 3 km (D02), respectively. DO2 covers countries from Estonia to the United Kingdom in the north and from Israel over Tunisia to the largest part of Morocco in the south. The WRF model physical setup has been derived to a large extent from Pieri et al. (2015) and von Hardenberg et al. (2015). The Yonsei University scheme (Hong et al. 2006) has been chosen for the planetary boundary layer turbulence closure, the RRTMG shortwave and longwave schemes are used for radiation (Iacono et al. 2008; Mlawer et al. 1997; Iacono et al. 2000), and the Rapid Up-date Cycle (RUC) scheme has been chosen (Smirnova et al. 1997, 2000) as a multi-level soil model (6 levels) with higher resolution in the upper soil layer (0, 5, 20, 40, 160, 300 cm). No cumulus scheme has been activated in the innermost domain (D02) because the grid spacing allows us to resolve the convection dynamics. For consistency with the boundary conditions, the New Simplified Arakawa-Schubert (NewSAS) convection scheme (Han and Pan, 2001) has been used in the outermost domain (D01). The single-moment 6-class microphysics scheme (Hong and Lim, 2006) has been adopted.

CHAPTER data can be used to understand extreme-weather events and to model flash floods or forest fires, atmospheric dispersion or air quality, just to name a few applications. The dataset as a whole is meant to advance our understanding of the mechanisms of past natural disasters and phenomena, to discover climate-change signals in the phenomena investigated, and to understand how current models have to be set up in order to reproduce observations – i.e. to produce predictions useful for administrative and political decision making.

Access to the dataset:

The dataset is available via the GLOBUS file-transfer service for large-volume data (see app.globus.com).

Direct link to the dataset on GLOBUS: CHAPTER-public folder on the LRZ DSA Container PN29CO-DSS-0000.

To gain access to the data, it is necessary to be signed up to a free GLOBUS user account and have an endpoint (transfer target, e.g. installed free Globus Connect Personal). You can create an account for free, or – if your institution is federated – use your institutional account to sign in (conditions see www.globusonline.org/legal/terms; GLOBUS is based in Chicago/US).

Versions

Version v1 2024
10.25927/0ppk7-znk14

Keywords and subjects

- 551.57 Hydrometeorology
- 551.6 Climatology and weather
- 518 Numerical analysis
- 004 Data processing & computer science

Details

DOI
DOI 10.25927/0ppk7-znk14

Resource type
Dataset

Publisher
Leibniz Supercomputing Centre (LRZ), Garching b.M., Germany

Languages
English

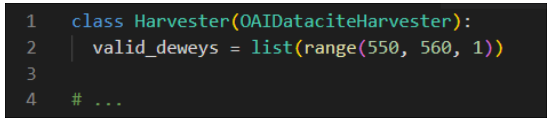
Rights

Creative Commons Attribution 4.0 International

Export

JSON

This dataset was harvested from the general-purpose LRZ Fair Data Portal by the domain specific AWI Earth Data Portal

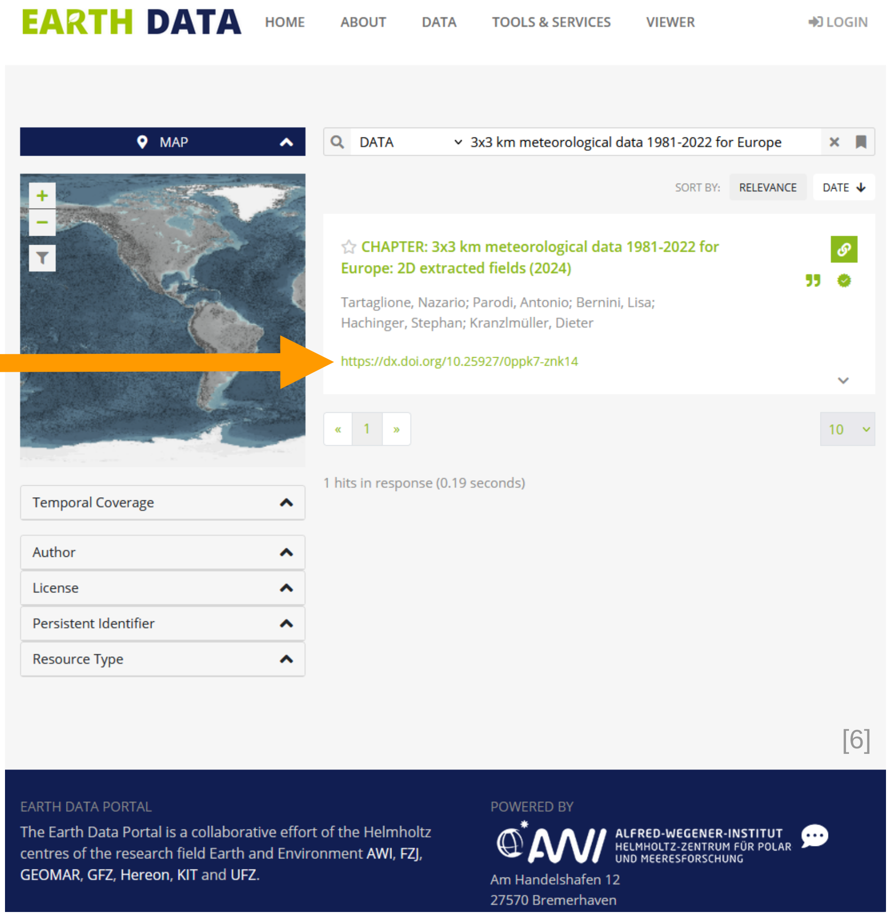


```
class Harvester(OAIDataciteHarvester):
  valid_deweys = list(range(550, 560, 1))

  # ...
```

The datasets to be harvested are identified based on their Dewey classification.

earth-data.de/data



DATA: 3x3 km meteorological data 1981-2022 for Europe

CHAPTER: 3x3 km meteorological data 1981-2022 for Europe: 2D extracted fields (2024)

Tartaglione, Nazario; Parodi, Antonio; Bernini, Lisa; Hachinger, Stephan; Kranzlmüller, Dieter

https://dx.doi.org/10.25927/0ppk7-znk14

1 hits in response (0.19 seconds)

EARTH DATA PORTAL

The Earth Data Portal is a collaborative effort of the Helmholtz centres of the research field Earth and Environment **AWI, FZJ, GEOMAR, GFZ, Hereon, KIT** and **UFZ**.

POWERED BY

AWI ALFRED-WEGENER-INSTITUT HELMHOLTZ-ZENTRUM FÜR POLAR UND MEERESFORSCHUNG

Am Handelshafen 12
27570 Bremerhaven

[6]

[6] https://earth-data.de/data?offset=0&q=3x3+km+meteorological+data+1981-2022+for+Europe

Discussion

Challenges

- Large „immobile“ data sets (file size limit of e.g. zenodo.org)
- Different scientific domains
- Different storage resources

Status

- Frontend is live (manual publication)
- Backend is work in progress (automated scanning logic, DSA access)
- LRZ DSS/DSA and MWN Cloud Storage are supported

Solutions

- Metadata publication only
- Meaningful metadata (incl. DDC)
- Harvesting interfaces (e.g. OAI-PMH)
- Modularised architecture for easy support of versatile storage

Outlook

- Towards live backend
- Planning official LRZ service
- Support for different storage resources

FAIR Data Portal of the Leibniz Supercomputing Centre (LRZ)

LRZ FDM/RDM Team and Projects

You’re welcome to visit our booths 18-19-21 (DKRZ, UHH, NFDI4Earth, DLR, LRZ)!

Thanks for your attention from LRZ FDM/RDM Team

Stephan hachinger@lrz.de (lead)   Johannes munke@lrz.de (tech lead)

Parts of this work have been supported by the project InHPC-DE, funded by the German Federal Ministry of Education and Research (Förderkennzeichen 16HPC02).

rdm@lists.lrz.de

lrz.de/technologien/daten

Leibniz Supercomputing Centre (LRZ),
Boltzmannstr. 1, 85748 Garching bei
München, Germany

LRZ Fair Data Portal:
rdm.lab.lrz.de