



Improving the Findability of Digital Objects in Climate Science by adopting the FDO Concept

Marco Kulüke, Karsten Peters-von Gehlen, Ivonne Anders

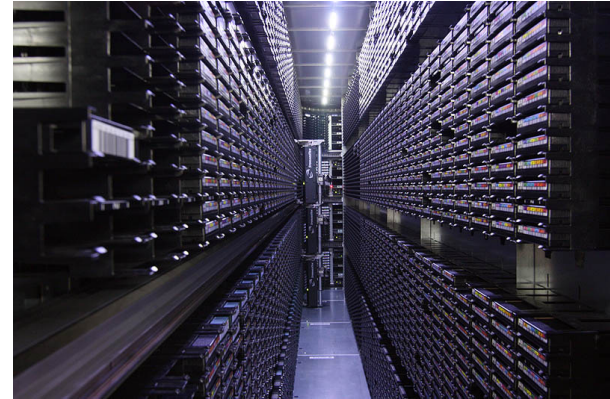
Deutsches Klimarechenzentrum (DKRZ)

Status Quo

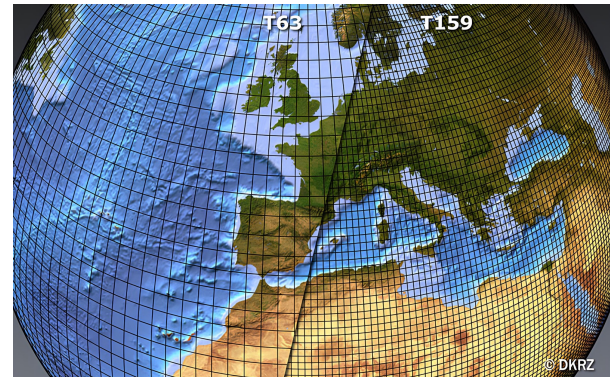
 Large-volume climate model data

 Interdisciplinary data (re)use

 Machine to machine interaction



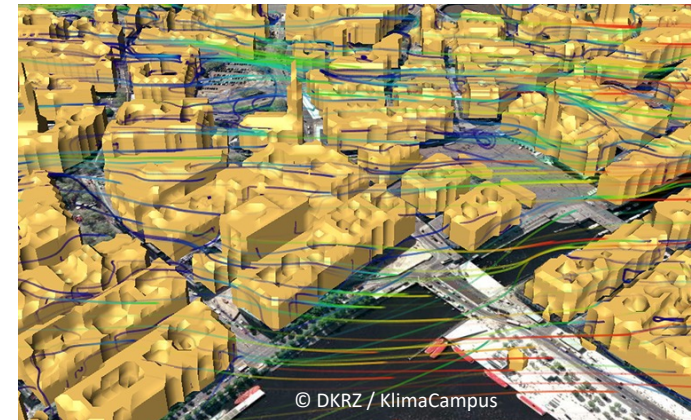
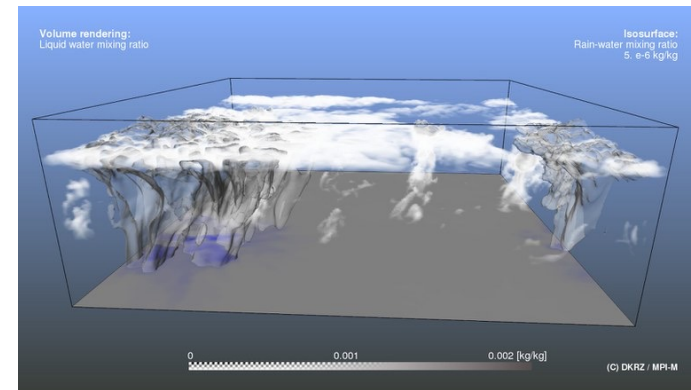
[dkrz.de]



[dkrz.de]

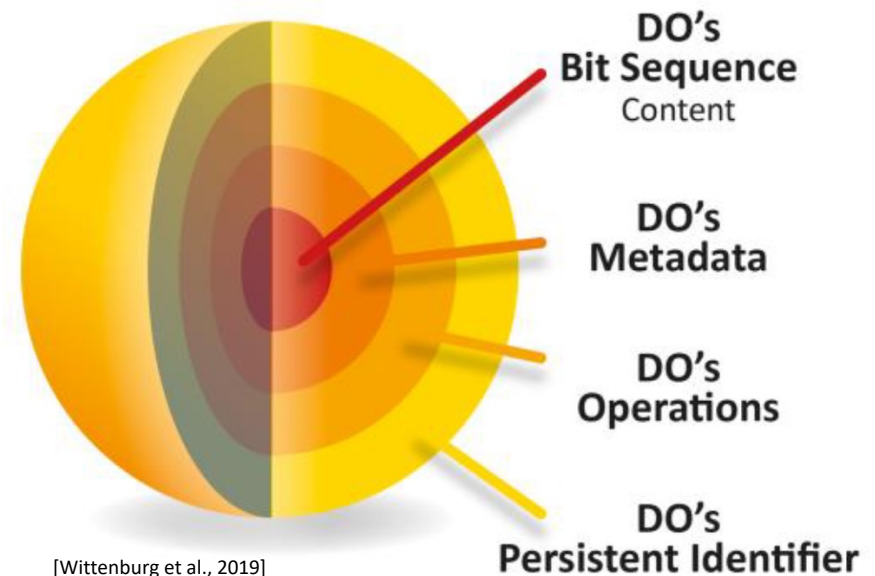
Challenges

- 🔍 Hard to find the right data
- 📈 Increasing workflow complexity
- 📈 Increasing data amounts



What are FAIR Digital Objects?

- Digital Objects (DOs) optimized for M2M actionability
- Includes data, metadata, documents, software
- *Concept of specifications* driven by international & interdisciplinary working groups



FAIR DIGITAL OBJECTS  FORUM

<https://fairdo.org/>

Imagine FDO as DO with a "Passport"



Standardized format

Machine and human readable



Provenance information

Access rights

Author attribution

Location

Verifiable



Unique persistent identifier

Metadata about content

License

FDO-readiness of CMIP6 Data*

Systematic review from researcher perspective:

- M2M actionable?
- Encourage Cross-disciplinary research?

The screenshot shows a Zenodo project page. At the top, the Zenodo logo is on the left, and search, communities, and dashboard links are on the right. Below the header, the project title is 'Strategy and Requirements for FDO-Standard in Simulation-based Climate Science - Milestone M3.1'. The authors listed are Kulüke, Marco; Peters-von Gehlen, Karsten; and Anders, Ivonne. The project is published on December 22, 2023, and is version 1.1. A file named 'ESIWACE3_Milestone_3.1.docx.pdf' is listed with a size of 1.7 MB. The preview shows the 'Introduction' section of the document, which discusses the need for a framework to ensure FAIRness in climate research and mentions machine-to-machine (M2M) actionability.

[Kulüke et al., 2023]

* 6th Phase of the Coupled Model Intercomparison Project [Eyring et al., 2016; Petrie et al., 2021]



Findable

- **Obstacle 1:** no mapping between meaningful names and model variable names
- **Obstacle 2:** PID profile does not contain variable info

The screenshot shows the 'CMIP6 Data Information View' for the dataset `pr_day_MPI-ESM1-2-LR_historical_r11i1p1f1_gn_18700101-18891231.nc`. It includes a table of 'General Information' with fields like Dataset Id, Persistent Identifier, Filesize, and Checksum. Below that is a 'Data access' section with links to the dataset on various servers (e.g., esgf3.dkrz.de, esgf-data1.jinl.gov). At the bottom, it states 'The file is part of the following aggregation(s)' and lists the aggregation: 'CMIP6_CMPMIP-M-MPI-ESM1-2-LR_historical_r11i1p1f1_day_pr_gn (version 20210901) (085890)'. A note at the bottom indicates 'This PID landing page service is provided by (German Climate Computing Centre)'.

PID profile of CMIP dataset

```
[3]: xarray.Dataset
```

► Dimensions: (time: 7305, bnds: 2, lat: 96, lon: 192)

▼ Coordinates:

time	(time)	datetime64[ns]	1870-01-01T12:00:00 ... 1889-12-...	📄 📄
lat	(lat)	float64	-88.57 -86.72 ... 86.72 88.57	📄 📄
lon	(lon)	float64	0.0 1.875 3.75 ... 356.2 358.1	📄 📄

▼ Data variables:

time_bnds	(time, bnds)	datetime64[ns]	...	📄 📄
lat_bnds	(lat, bnds)	float64	...	📄 📄
lon_bnds	(lon, bnds)	float64	...	📄 📄
pr	(time, lat, lon)	float32	...	📄 📄

▼ Attributes:

- CDO : Climate Data Operators version 2.0.0rc2 (<https://mpimet.mpg.de/cdo>)
- Conventions : CF-1.7 CMIP-6.2
- activity_id : CMIP

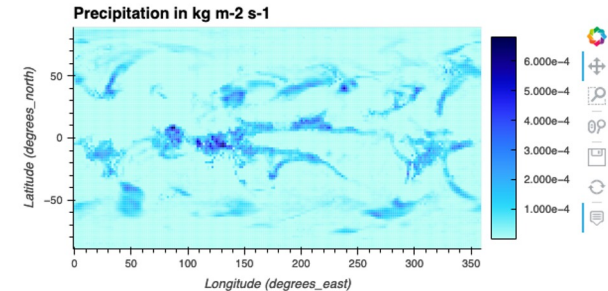
NetCDF header of same dataset

Accessible & Interoperable

hd1:21.14100/0163da97-35d1-387c-a51b-0d12fc4d3b24



CMIP6 DOs are machine accessible
and to a large extent interoperable

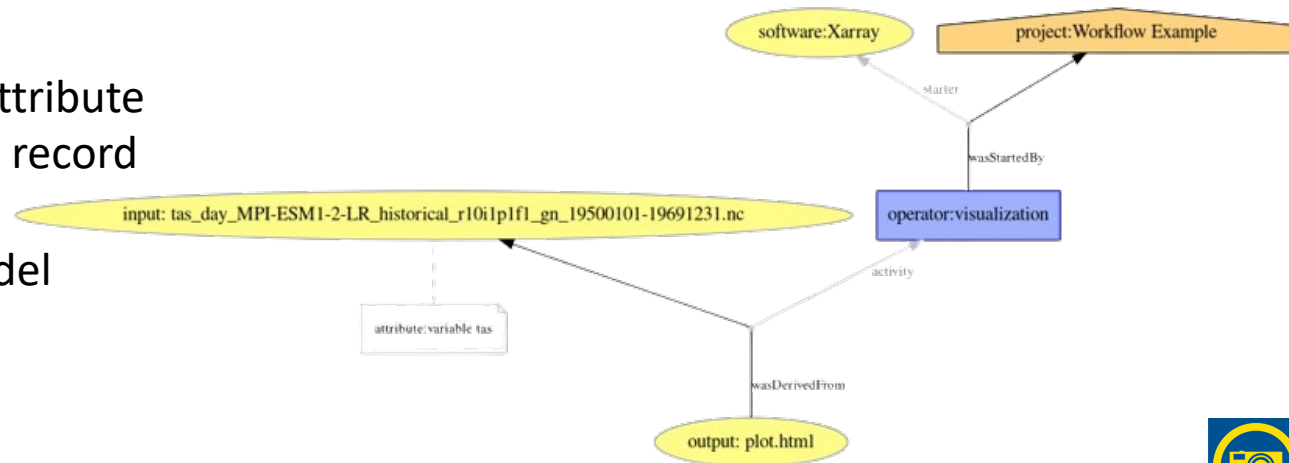


Reusable




CMIP6 DOs have *history* attribute
with sporadic provenance record

→ Adopt PROV data model

[Belhajjame et al., 2013]



How can FDO Concept help?

-  Enrich data PID profiles
-  Data catalogs as FDOs
-  Break down jargon barriers



Use Case: Apply FDO Concept to ESM Data

- 1 PB high-res ESM data derived from nextGEMS project¹
- Stored on HEALPix grid² → allows hierarchical output and horizontal chunking



Publish Intake / STAC catalog at WDCC³ with persistent identifier



¹ <https://nextgems-h2020.eu/>

² <https://healpix.sourceforge.io/>

³ <https://www.wdc-climate.de/>

Thank You!

References

Wittenburg, P.; Strawn, G.; Mons, B.; Bonino, L.; Schultes, E. *Digital Objects as Drivers towards Convergence in Data Infrastructures; EUDAT: Helsinki, Finland, 2019.* <https://doi.org/10.23728/b2share.b605d85809ca45679b110719b6c6cb11>

Eyring, V., G. A. Meehl, C. A. Senior, B. Stevens, R. J. Stouffer, and K. E. Taylor. 2016. "Overview of the Coupled Model Intercomparison Project Phase 6 (CMIP6) experimental design and organization." *Geoscientific Model Development* 9:1937--1958. <https://doi.org/10.5194/gmd-9-1937-2016>

Petrie, Ruth, Sebastien Denvil, Sasha Ames, Guillaume Levavasseur, Sandro Fiore, Chris Allen, Fabrizio Antonio, et al. 2021. "Coordinating an operational data distribution network for CMIP6 data." *Geoscientific Model Development* 14:629–644. <https://doi.org/10.5194/gmd-14-629-2021>

Kulüke, M., Peters-von Gehlen, K., & Anders, I. (2023). *Strategy and Requirements for FDO-Standard in Simulation-based Climate Science - Milestone M3.1.* Zenodo. <https://doi.org/10.5281/zenodo.10423391>

Belhajjame, K., R. B'Far, J. Cheney, S. Coppens, S. Cresswell, Y. Gil, P. Groth, et al. 2013. "PROV-DM: The PROV Data Model." Edited by L. Moreau and P. Missier. W3C. <https://www.w3.org/TR/2013/REC-prov-dm-20130430/>



Funded by the European Union. This work has received funding from the European High Performance Computing Joint Undertaking (JU) under grant agreement No 101093054.



Funded by Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) project number: 460036893