Use Case: Earth System Modeling

DataCite Open Hours, 2019-09-04

Martina Stockhause
https://orcid.org/0000-0001-6636-4972
German Climate Computing Center (DKRZ)

Background: DKRZ as DOI Publication Agency

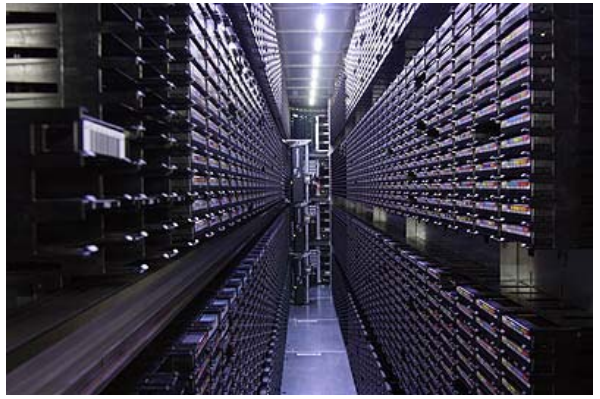
German Climate Computing Center (DKRZ)

The mission of German Climate Computing Center (DKRZ) is to provide high performance computing (HPC) platforms, sophisticated and high capacity data management and services for premium climate science.

- HPC Services
- World Data Center for Climate (WDCC)
- IPCC Data Distribution Centre (IPCC DDC)

DOI Publisher since 2004:

- Partner of STD-DOI project, which led to DataCite
- DataCite DOI #1 - 18.03.2004: http://doi.org/10.1594/WDCC/EH4_OPYC_SRES_A2
- Client/Repository of TIB, becoming a DataCite Member
- DOIs are registered
  - for data in DKRZ's long-term archive and
  - on behalf of ESGF (Earth System Grid Federation)



Background: Partners and Projects

Research Projects in Earth System Modeling (ESM):

- Coupled Model Intercomparison Project Phase 6 – CMIP6:
  WGCM-CMIP6 (science) and WIP (infrastructure)
- Coordinated Regional Climate Downscaling Experiment – CORDEX

Intergovernmental Panel on Climate Change (IPCC) partners:

- IPCC Working Group I – WGI (The Physical Science Basis)
- IPCC Data Distribution Centre – DDC: jointly managed by CEDA, DKRZ, and CIESIN
- IPCC Task Group on Data Support for Climate Change Assessment – TG-Data

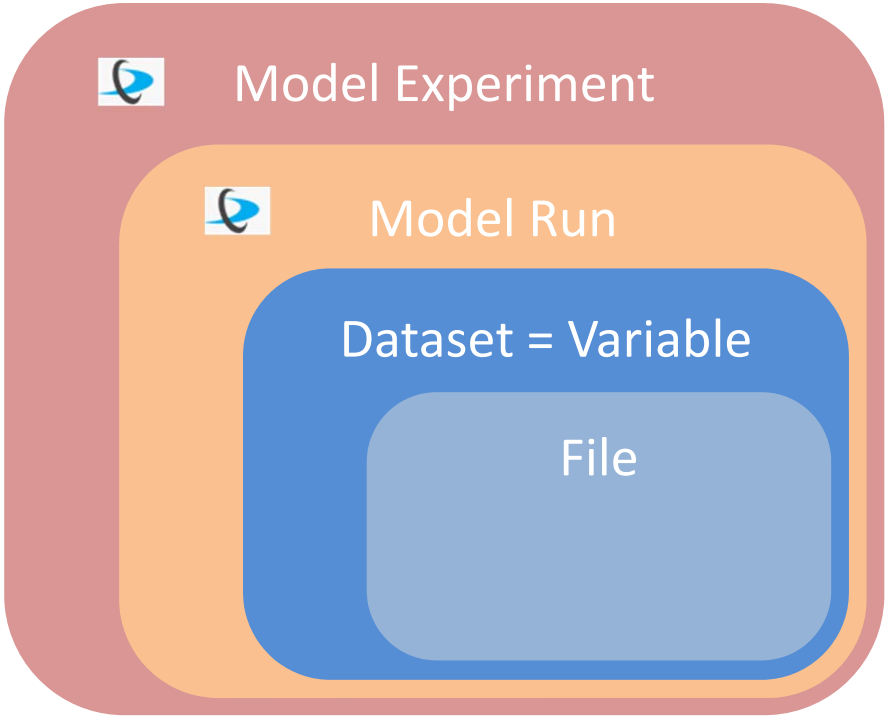
Infrastructure Partners:

- Earth System Grid Federation – ESGF
- European Network for Earth System Modelling – IS-ENES
- Earth System Documentation – ES-DOC

ESM: Data Characteristics

Earth System Modeling - Data Characteristics:

- High-volume consisting of many datasets created by each run of a climate model
- Highly structured
- Standardized: community and project standards (CMIP6)
- Decentral data with many data nodes and several portals (ESGF)
- Data Replica: copies of data subsets for quick access by local users



Data is commonly **cited as a large collection of datasets created together** in order to keep the balance between data and articles in reference lists.
Data access is provided on File and Dataset granularity.

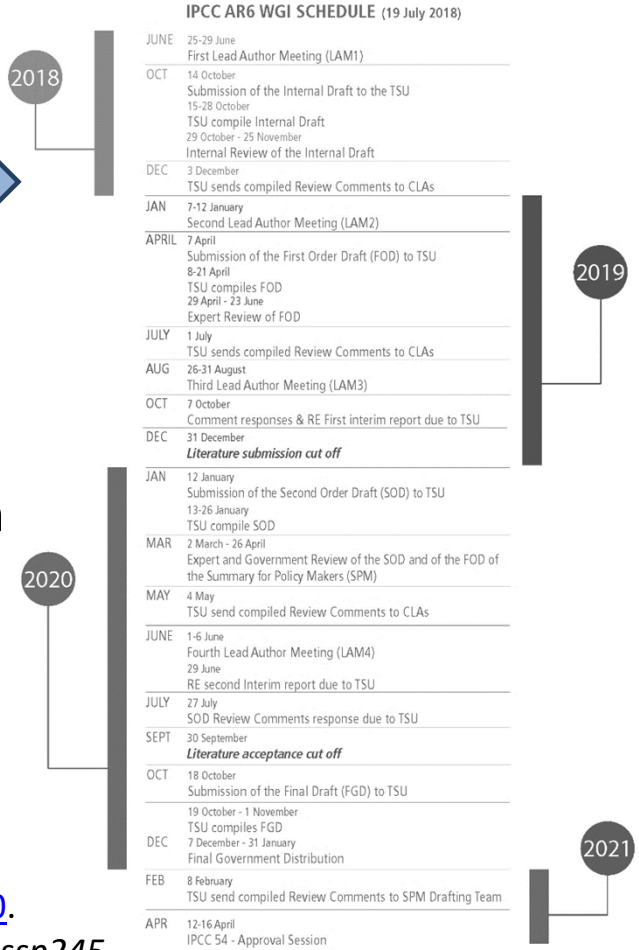
ESM: Workflow Characteristics (Theory)

CMIP6 / ESGF: Data Creation → DDC Support: Data Analysis → IPCC DDC AR6: Data Reuse

Research data with DataCite DOI reference on evolving data collections (since 2018-06-28 and ongoing):

Coupled Model Intercomparison Project Phase 6 (CMIP6) with Earth System Grid Federation (ESGF) as data infrastructure:

- Data added over several years on several data nodes
- Data usage starts when it becomes accessible (prior to creator's paper publication)



(Source: https://www.ipcc.ch/site/assets/uploads/2018/12/AR6_WGI_Schedule.pdf)

Schupfner et al. (2019). *DKRZ MPI-ESM1.2-HR model output prepared for CMIP6 ScenarioMIP*. **Version 20190710.** Earth System Grid Federation. https://doi.org/10.22033/ESGF/CMIP6.2450.

Schupfner et al. (2019). *DKRZ MPI-ESM1.2-HR model output prepared for CMIP6 ScenarioMIP ssp245.* **Version 20190710.** Earth System Grid Federation. https://doi.org/10.22033/ESGF/CMIP6.4398.

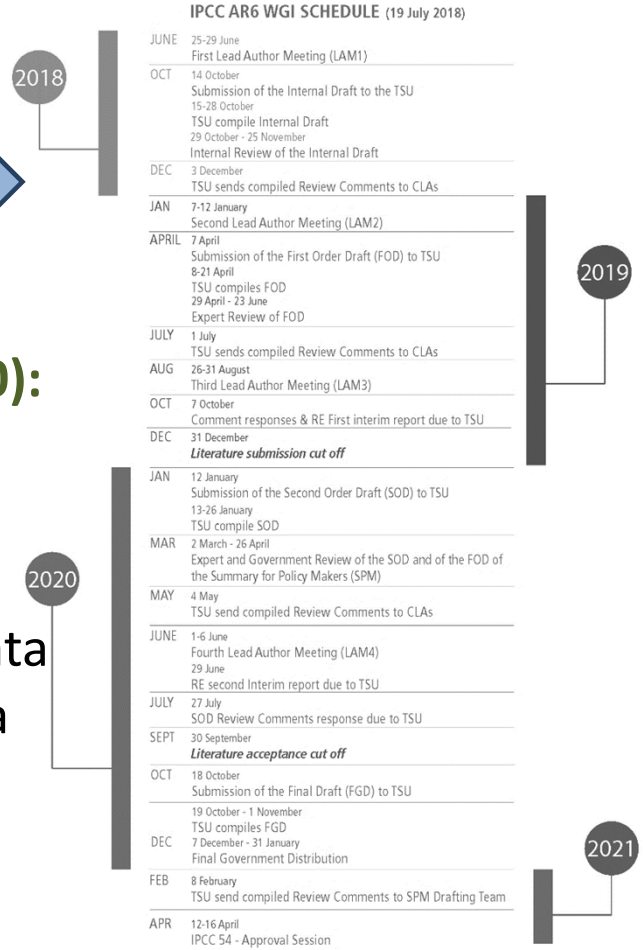
ESM: Workflow Characteristics (Theory)

CMIP6 / ESGF: Data Creation → DDC Support: Data Analysis → IPCC DDC AR6: Data Reuse

CMIP6 data analysis and citation by IPCC authors (2019-2020):

IPCC Data Distribution Centre (DDC) supports IPCC authors in data analyses by providing access to replicated data and a Virtual Workspace:

- CMIP6 data analyzed together with other latest research data
- IPCC AR6 (6th Assessment Report) cites evolving CMIP6 data and stores provenance information on derived IPCC results



(Source: https://www.ipcc.ch/site/assets/uploads/2018/12/AR6_WGI_Schedule.pdf)

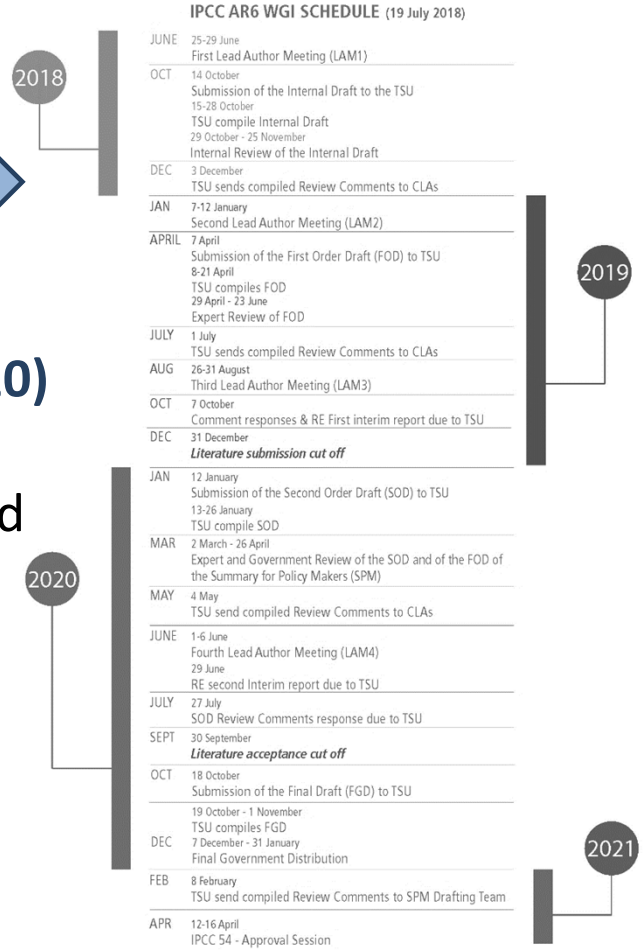
ESM: Workflow Characteristics (Theory)

CMIP6 / ESGF: Data Creation → DDC Support: Data Analysis → IPCC DDC AR6: Data Reuse

IPCC AR6 Reference Data (CMIP6 subset snapshot on 10/2020) is archived and gets reused:

CMIP6 data subset used for IPCC AR6 WGI report is transferred into the IPCC DDC long-term archive to build the IPCC DDC AR6 Reference Data Archive. DDC DOIs are registered with relation to CMIP6 evolving data references.

- Recommendation for citation of well-curated and well-documented DDC reference data



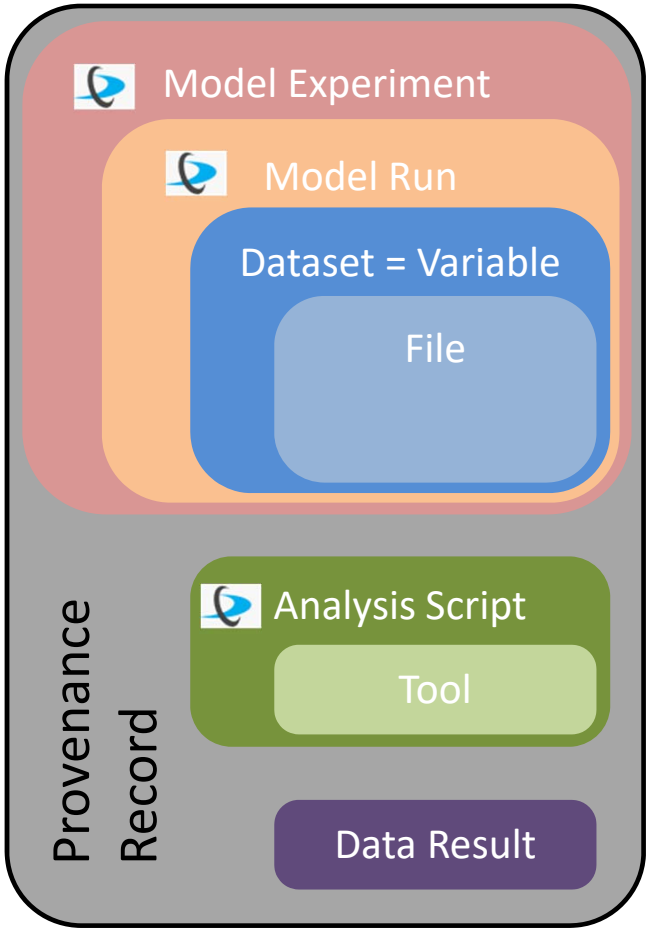
(Source: https://www.ipcc.ch/site/assets/uploads/2018/12/AR6_WGI_Schedule.pdf)

ESM: Improvement of data traceability

DDC Support idea for IPCC AR6:

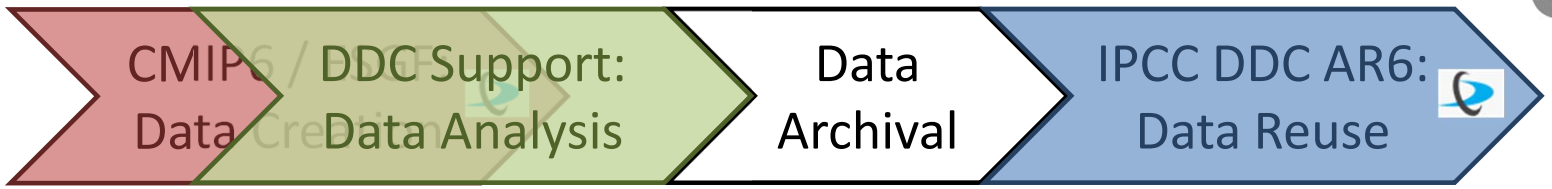
The aim is to improve the traceability of IPCC results like figures, tables, and headline statements.

- **Provenance** (W3C PROV) for traceability, including
  - **Source File IDs** – files
  - **Source data DOIs** – data references
  - **Software IDs** – scripts (GitHub/Zenodo)
  - **Data Result ID** – resulting dataset
- **Formal citation of digital data** for credit
  and as part of Good Scientific Practice (e.g. Commitment Statement in the Earth, Space, and Environmental Sciences)



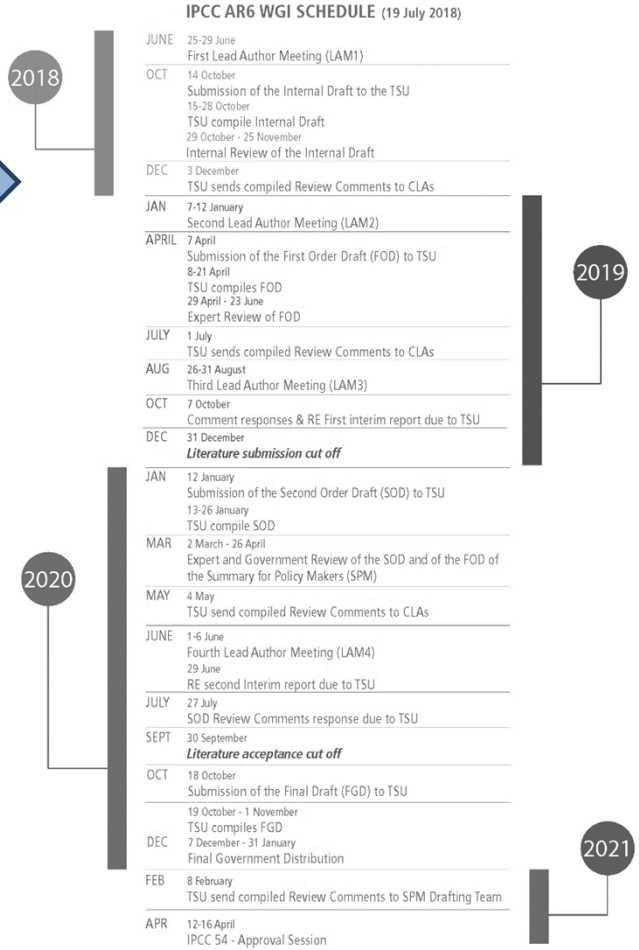
(M. Stockhause et al. (2019): Data Distribution Centre Support for the IPCC Sixth Assessment. http://doi.org/10.5334/dsj-2019-020.)

Challenge 1: Workflow



Workflow phases are not separable:

- Large international research projects tend to be delayed, but IPCC has a fixed schedule
  → data analysis overlaps with data creation
- Long-term archival of high volume data needs time
  → 1 year gap between AR6 submission deadline and IPCC DDC AR6 Reference Data availability



(Source: https://www.ipcc.ch/site/assets/uploads/2018/12/AR6_WGI_Schedule.pdf)

Challenge 2: Granularities

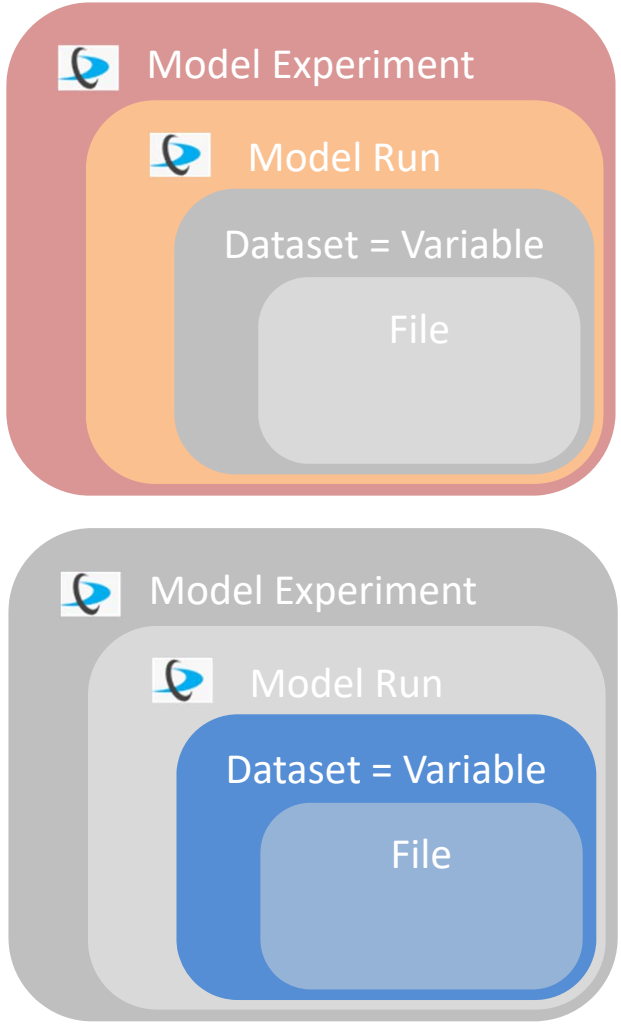
Citation or DOI granularity:

- Data citation in papers or by digital data to give credit
- Data usage or impact statistics (Scholix use case)

COUNTER Code download and provenance records should include information from both granularities.
→ How to integrate DOIs in COUNTER Code reports?

Data access granularity:

- Data access and data provenance
- Download statistics and specific data download analyses (COUNTER Code of Practice for Research Data)



Challenge 3: Another Metadata Format…

Change to DataCite REST API:

- Metadata formats currently supported:
  - DataCite XML: OAI-Server for harvesting (OpenAire)
  - ESGF JSON: JSON representation of DataCite metadata for infrastructure partners and providers of citation information via API
- DataCite JSON is an additional format to support and maintain a template for

→ possibly base64-encoded XML within JSON to be implemented
→ XMLs can be validated against XSD before registration

Used DataCite Services

Used DataCite Services:

1. DataCite MDS API:
2. DataCite Search:
3. Statistic / Status / Fabrica support pages
4. Test environment

- DOI / Metadata Registration
- DOI Registration Status

Plans / Investigation to use DataCite Services:

1. DataCite REST API:
2. DataCite Event Data:
3. Usage Report API:

- DOI / Metadata Registration…
- Planned to use as Scholix hub (currently OpenAire hub used)
- Ongoing granularity discussion

Further Information:

- DKRZ Long-Term Archive: http://cera-www.dkrz.de
- Citation Service (ESGF): http://cmip6cite.wdc-climate.de
- WGCM-CMIP6: https://www.wcrp-climate.org/wgcm-cmip/wgcm-cmip6
- IPCC DDC: http://ipcc-data.org
- DDC Support: https://cedadev.github.io/ipcc_ddc/
- ESGF: http://esgf.llnl.gov
- Enabling FAIR Data Project: http://www.copdess.org/enabling-fair-data-project/

Stockhause, M., Juckes, M., Chen, R., Moufouma Okia, W., Pirani, A., Waterfield, T., Xing, X. and Edmunds, R., 2019. Data Distribution Centre Support for the IPCC Sixth Assessment. Data Science Journal, 18(1), p.20. doi: http://doi.org/10.5334/dsj-2019-020.