Science & Technology Facilities Council | UK Research and Innovation

NERC SCIENCE OF THE ENVIRONMENT

Collaborative frameworks and best practices

Provenance and documentation, storing data underlying figures and tables, data referencing

Martin Juckes, CEDA.
martin.juckes@stfc.ac.uk

Findable **A**ccessible **I**nteroperable **R**eusable

Discoverable **A**ccessible **R**e-usable **K**nowledgable

is-enes INFRASTRUCTURE FOR THE EUROPEAN NETWORK FOR EARTH SYSTEM MODELLING

Work supported by IS-ENES3 with funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 824084

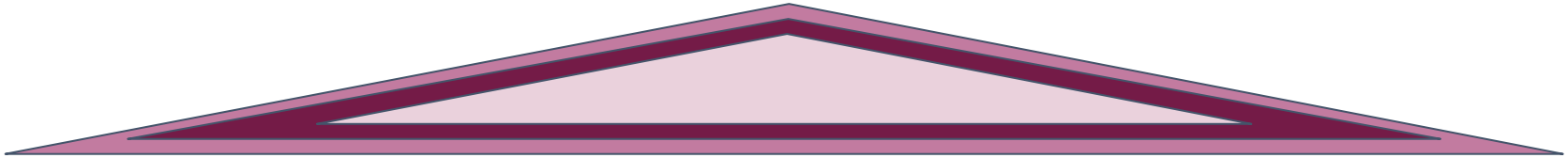
europa.eu

FAIR data in action: Objectives

How do we get beyond the platitudes of wanting to be “FAIR”?

- Transparent and citable chain of documentation, code and data, giving accessibility for users and credit to contributors;

- Robust and efficient exchange among networks of science and IT teams;



JASMIN and DKRZ servers

Shared computing resources;
Access to shared libraries and large datasets;

- GitHub
- Zenodo

Collaborative platforms supporting extensive user communities;

WDCC and CEDA Archive Services

Data citation services;
Cataloguing datasets;

Sponsors and Credits

BEIS fund CEDA and BMBF funds DKRZ to provide support for IPCC, under governance of TG-DATA

Department for Business, Energy & Industrial Strategy

Bundesministerium für Bildung und Forschung

With H2020 funding, IS-ENES supports many services for Earth System Modelling and related data services.

europa.eu

is-enes
INFRASTRUCTURE FOR THE EUROPEAN NETWORK FOR EARTH SYSTEM MODELLING

Through the Working Group on Coupled Models (WGCM) and CMIP Panel, WCRP members deliver a suite of climate projections, predictions and reconstructions(CMIP6)

WCRP
World Climate Research Programme

ESGF
Earth System Grid Federation

ESGF enables the efficient global dissemination of model output

Centre for Environmental Data Analysis
SCIENCE AND TECHNOLOGY FACILITIES COUNCIL
NATURAL ENVIRONMENT RESEARCH COUNCIL

National Centre for Atmospheric Science
NATURAL ENVIRONMENT RESEARCH COUNCIL

National Centre for Earth Observation
NATURAL ENVIRONMENT RESEARCH COUNCIL

Github: one slide introduction

Software repositories:

+ forking and merging branches;
+ support for documentation;
+ issue tracking and task management;

Wikis;

Export to Zenodo;

Plus many more tools for collaboration.

github.com

Zenodo: one slide introduction

Built and developed by researchers, to ensure that everyone can join in Open Science.

zenodo.org

A catch-all repository for digital artefacts, including data, software and other artefacts in support of publications, also conference presentations, posters, etc;

Easy generation of DOIs; Integration with GitHub; DOI reservation

Data plotted vs. data used: an example

Example: Fig. 5.5 of FOD: Global anthropogenic CO2 emissions.

Data plotted: 9 annual time series (with uncertainty).

Data used in the plot: national emissions inventories and model output.

The **data plotted** is not accessible to most readers if we do not save this IPCC product.

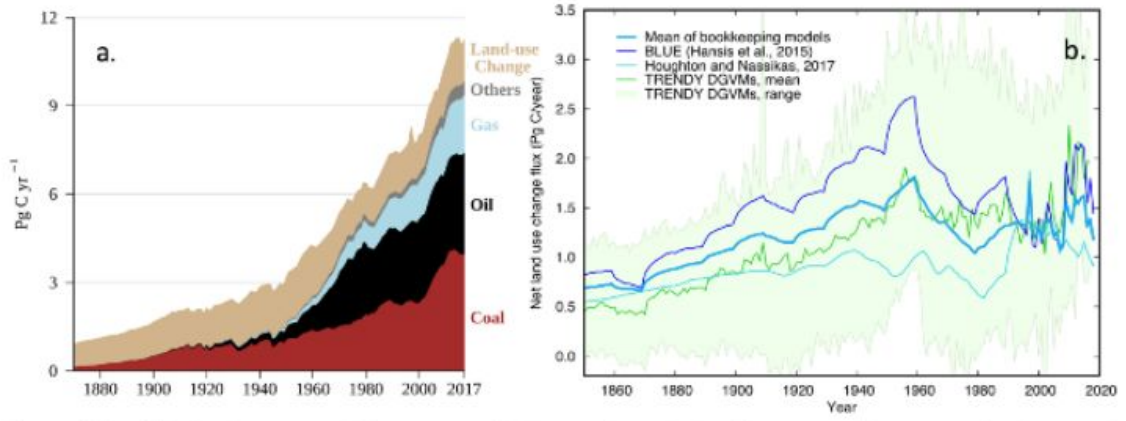


Figure 5.5: Global anthropogenic CO_2 emissions: A) Historical trends of anthropogenic CO_2 emission for the period 1870 to 2016. Data sources: (Andrew, 2018; BP, 2018; IEA, 2017; Marland et al.; Quéré et al., 2018). B) The net land use change CO_2 flux (Pg yr^{-1}) from two bookkeeping and 16 dynamic global vegetation models (Le Quéré et al., 2018a). Bookkeeping models are BLUE (Hansis et al., 2015; Houghton and Nassikas, 2017) both updated as described in (Le Quéré et al., 2018a). All estimates are unsmoothed annual data. Note that the estimates differ in process comprehensiveness of the models and in definition of flux components included in the net land use change flux.

Data and Software Citations

Enabling FAIR Data Project developed author guidelines on data and software citation in scholarly publications. The commitment is signed by publishers, repositories, organizations (www.copdess.org/enabling-fair-data-project).

Author Guidelines:

- Deposit data in a trusted repository (int. standard CoreTrustSeal): IPCC DDC
- ***Cite data in text body with a corresponding reference in the reference list***
- ***Similar instructions for software***
- Provide unrestricted access to all data and materials underlying reported findings

CMIP6 Data Citation Guidance example: http://bit.ly/2gBCuqM

Data Citation: Motivation

Why cite data?

- Give credit to data providers
- Improve traceability of research findings

CMIP6 Data: Three Steps for Data Citation

I. Find CMIP6 Data References

input4MIPs example:

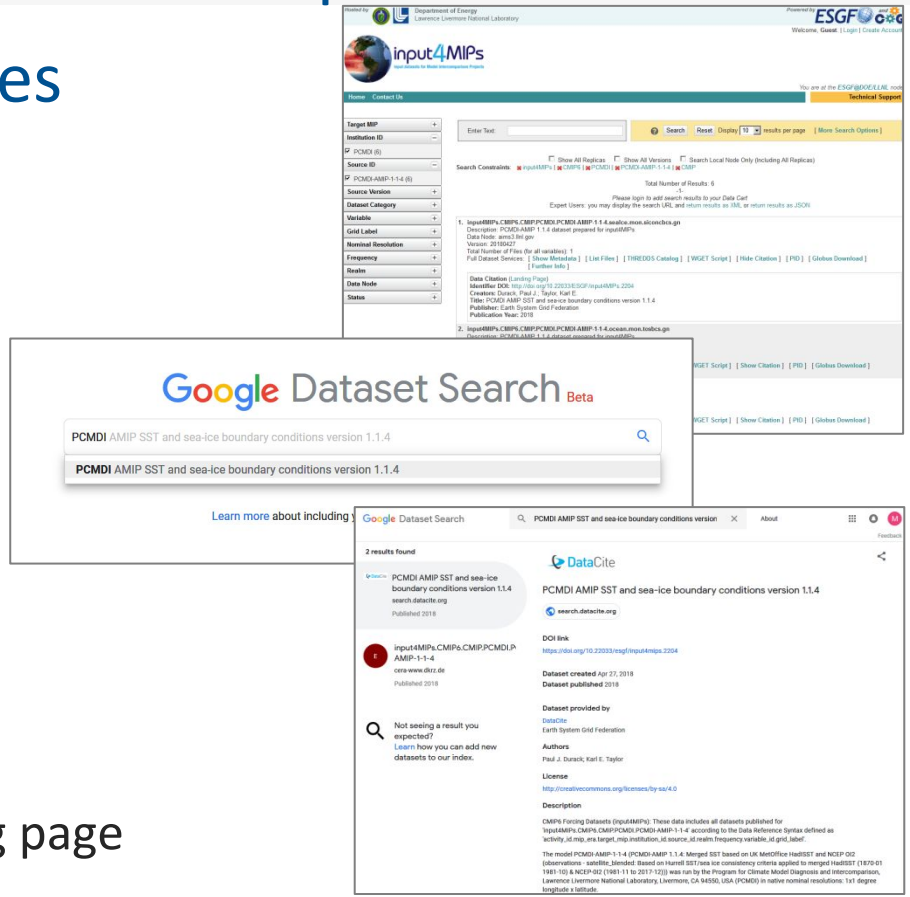
- ESGF CoG portal

  https://esgf-node.llnl.gov/search/input4mips/

- Google Dataset Search

  https://toolbox.google.com/datasetsearch/

- NetCDF file □ ES-DOC □ DOI landing page

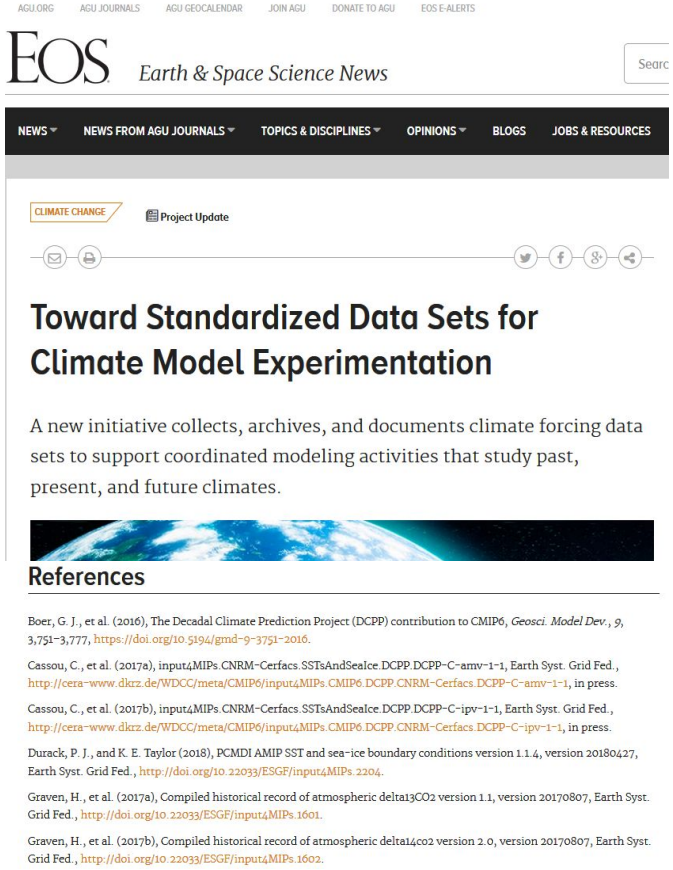


CMIP6 Data: Three Steps for Data Citation

I. Find CMIP6 Data References

II. Cite Data

Cite data in the text and include data references in reference lists of articles, e.g. https://doi.org/10.1029/2018EO101751 (according to the recommendation of the "Enabling FAIR Data Project")



AGU.ORG AGU JOURNALS AGU GEOCALENDAR JOIN AGU DONATE TO AGU EOS E-ALERTS

EOS *Earth & Space Science News*

NEWS | NEWS FROM AGU JOURNALS | TOPICS & DISCIPLINES | OPINIONS | BLOGS | JOBS & RESOURCES

CLIMATE CHANGE | Project Update

Toward Standardized Data Sets for Climate Model Experimentation

A new initiative collects, archives, and documents climate forcing data sets to support coordinated modeling activities that study past, present, and future climates.

References

Boer, G. J., et al. (2016), The Decadal Climate Prediction Project (DCPP) contribution to CMIP6, *Geosci. Model Dev.*, *9*, 3,751–3,777, https://doi.org/10.5194/gmd-9-3751-2016.

Cassou, C., et al. (2017a), input4MIPs.CNRM-Cerfacs.SSTsAndSealce.DCPP.DCPP-C-amv-1-1, Earth Syst. Grid Fed., http://cera-www.dkrz.de/WDCC/meta/CMIP6/input4MIPs.CMIP6.DCPP.CNRM-Cerfacs.DCPP-C-amv-1-1, in press.

Cassou, C., et al. (2017b), input4MIPs.CNRM-Cerfacs.SSTsAndSealce.DCPP.DCPP-C-ipv-1-1, Earth Syst. Grid Fed., http://cera-www.dkrz.de/WDCC/meta/CMIP6/input4MIPs.CMIP6.DCPP.CNRM-Cerfacs.DCPP-C-ipv-1-1, in press.

Durack, P. J., and K. E. Taylor (2018), PCMDI AMIP SST and sea-ice boundary conditions version 1.1.4, version 20180427, Earth Syst. Grid Fed., http://doi.org/10.22033/ESGF/input4MIPs.2204.

Graven, H., et al. (2017a), Compiled historical record of atmospheric delta13CO2 version 1.1, version 20170807, Earth Syst. Grid Fed., http://doi.org/10.22033/ESGF/input4MIPs.1601.

Graven, H., et al. (2017b), Compiled historical record of atmospheric delta14co2 version 2.0, version 20170807, Earth Syst. Grid Fed., http://doi.org/10.22033/ESGF/input4MIPs.1602.

CMIP6 Data: Three Steps for Data Citation

I. Find CMIP6 Data References

II. Cite Data

III. Credit and Reuse

- Impact of CMIP6 data reaches the data providers
  via services of the publishers (e.g. WoS)
  or via Scholix and data publisher services
- Article readers can reuse the data
  by resolving the DataCite DOI
  (part of the data reference),
  e.g. **https://doi.org/10.22033/ESGF/input4MIPs.2204**

Data Provenance

Data provenance describes how the dataset (e.g. figure) was created, including information on used source data and applied analysis. The international standard is W3C PROV.

Provenance Documentation (e.g. a figure) - Best Practice:

- Add relations (references) for persistent (e.g. DOI) resources: applied software DOI, source data DOI
- Store resources without DOIs: W3C PROV record, source data...
- Brief human-readable provenance description in abstract

Proposed Framework (for discussion)

- Create a repository for each figure of family of related figures;
- After submission of SOD, create a Zenodo entry and reserve a DOI for the software (embargoed) and a catalogue entry (embargoed) for any data files created in support of the assessment; add relations to source data, software and provenance information to data file entries.
- The catalogue will provide links to underlying data sources (including links to data holdings, not just the relevant publications -- but for CMIP6, see Martina's talk);

Proposed Framework (for discussion)

- Add information for figure reproducibility to figure caption:
  - Place figure data DOIs in the report figure captions for direct access to the plotted data
  - Add citations for used data and software to captions (for multiple data citations these could be put into a separate table referenced in the figure's caption - example: FOD Fig. 5.13)
- Add data and software references in a dedicated section of the chapter's reference list

THE END