

Introduction to the ACCESS-NRI Intake catalog

Dougal Squire ¹ and Romain Beucher ¹

¹ Australian Earth System Simulator (ACCESS-NRI), Canberra, Australia

DOI: 10.xxxxxx/draft

Software

- [Review](#)
- [Repository](#)
- [Archive](#)

Editor: Romain Beucher

Reviewers:

- [@rbeucher](#)

Submitted: 21 September 2023

Published: unpublished

License

Authors of papers retain copyright and release the work under a Creative Commons Attribution 4.0 International License ([CC BY 4.0](#)).

Summary

The ACCESS-NRI catalog is essentially a table of climate data products that exist on Gadi. Each entry in the table corresponds to a different product, and the columns contain attributes associated with each product—things like the models, frequencies and variables available. Users can search on the attributes to find the products that might be useful to them. For example, a user might want to know which data products contain variables X, Y and Z at monthly frequency. The ACCESS-NRI catalog enables users to find products that satisfy their query and to subsequently load their data without having to know the location and structure of the underlying files.

Statement of need

The ACCESS-NRI catalog provides a catalog of *Intake sources* and associated metadata.

.. topic:: Wait, what are “Intake sources”?

“Intake” is a Python package that provides a general interface for loading data in Python. Intake provides a set of data loaders (called “drivers”) that allow users to load a wide range of data formats into familiar Python data structures (e.g. a pandas dataframe or xarray Dataset) using the exact same code. Some drivers are built into Intake, and some are provided by other “plugin” packages. Documentation of the Intake package can be found here <https://intake.readthedocs.io/en/latest/index.html> and a list of available Intake drivers can be found here <https://intake.readthedocs.io/en/latest/plugin-directory.html>.

“Intake sources” describe data that can be loaded using an Intake driver. For example, an Intake source might describe a simple csv file to be opened using the inbuilt Intake “csv” driver, or it might describe a set of netcdf files to be opened using the plugin Intake-ESM <https://intake-esm.readthedocs.io/en/stable/> “esm_datastore” driver. In fact, the ACCESS-NRI catalog itself is a type of Intake source that uses the plugin Intake-dataframe-catalog <https://intake-dataframe-catalog.readthedocs.io/en/latest/?badge=latest> “df_catalog” driver.

The entries in the ACCESS-NRI catalog are *Intake-ESM datastores* for climate data products that exist on Gadi.

.. topic:: Wait, what are “Intake-ESM datastores”?

Intake-ESM datastores are a type of Intake source that describes a climate data product comprising many files (e.g. netcdf files). Users can search across metadata associated with each file (e.g. a variable name) and open contiguous files into xarray Datasets for subsequent analysis. For example, NCI provides Intake-ESM datastores for the CMIP5 and CMIP6 data collections available on Gadi. In their documentation <https://opus.nci.org.au/pages/viewpage.action?pageId=213713098> they demonstrate how to execute a search across these datastores to find and open a few datasets.

Intake-ESM datastores are often also referred to as Intake-ESM “catalogs”. In this sense, the ACCESS-NRI catalog can be thought of as providing a catalog of catalogs. To try and avoid confusion in these docs, we will use the term “datastore” instead of “catalog” when referring to Intake-ESM.

.. note:: The CMIP5 and CMIP6 Intake-ESM datastores generated by NCI are available as entries in the ACCESS-NRI catalog.

A set of core metadata attributes are associated with each entry in the ACCESS-NRI catalog. At the moment these include:

- The name of the data product
- A short description of the data product
- The model(s) used
- The realm(s) available
- The frequency(/ies) available
- The variable(s) available

A simple search API allows users to filter the entries in the catalog based on these metadata attributes. The idea is that users will:

1. search the ACCESS-NRI catalog for data products, e.g. products containing the models, variables etc that are of interest to them.
2. open the Intake-ESM datastore(s) for the product(s) of interest.
3. search the Intake-ESM datastore(s) for the datasets within each product that are of interest to them. A “dataset” here is a set of files that can be readily opened and combined for analysis.
4. open the datasets of interest as xarray Dataset(s).
5. perform some analysis on the xarray Dataset(s).

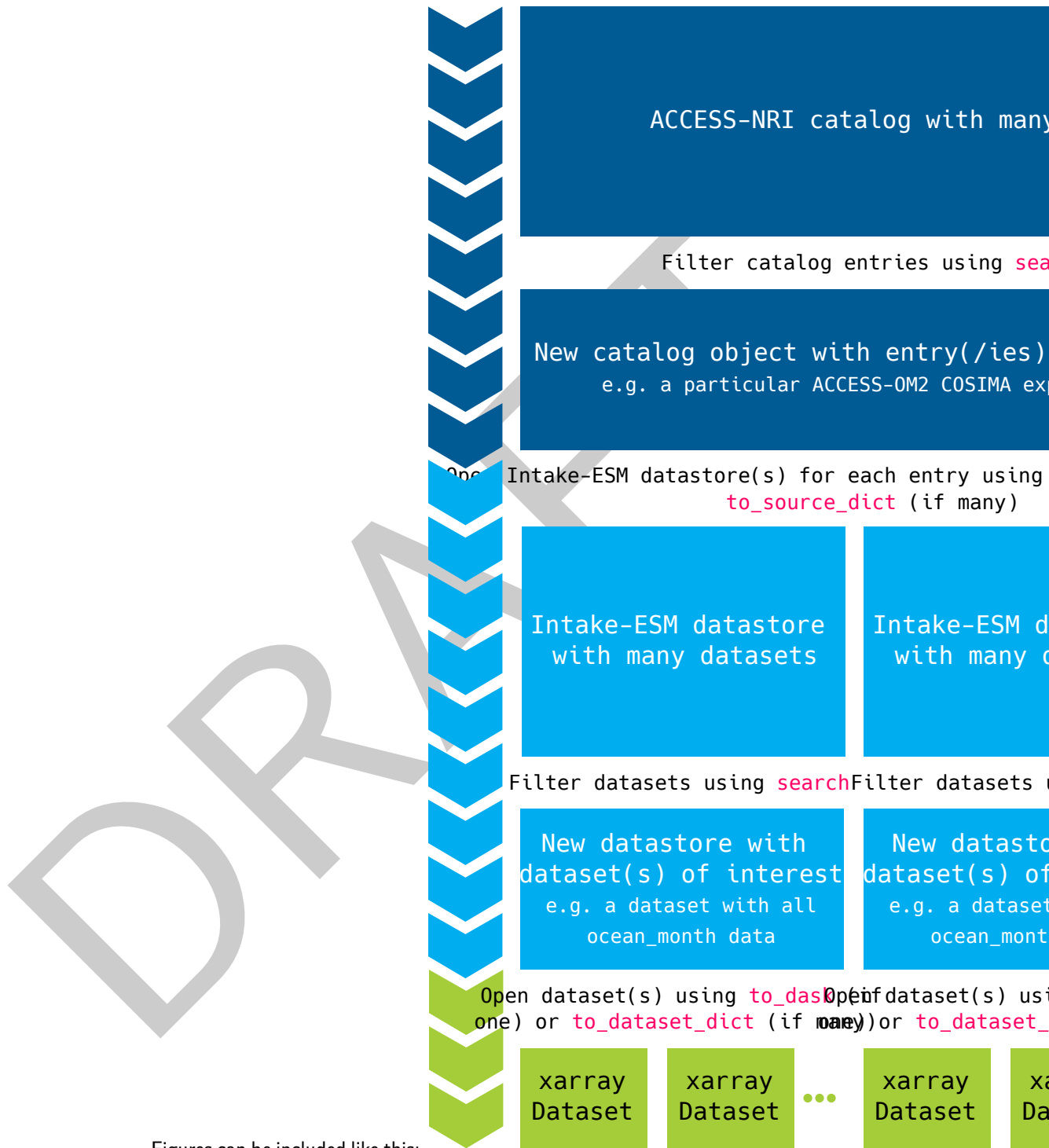
Citations

Citations to entries in paper.bib should be in [rMarkdown](#) format.

If you want to cite a software repository URL (e.g. something on GitHub without a preferred citation) then you can do it with the example BibTeX entry below for ([fidgit?](#)).

For a quick reference, the following citation commands can be used: - Banihirwe et al. ([2023](#)) -> “Author et al. (2001)” - ([Banihirwe et al., 2023](#)) -> “(Author et al., 2001)” - [Banihirwe et al. ([2023](#)); Banihirwe et al. ([2022](#)); rocklin2015dask] -> “(Author1 et al., 2001; Author2 et al., 2002)”

73 Figures



74 Figures can be included like this:
75 and referenced from text using ??.



76 Figure sizes can be customized by adding an optional second parameter:

77 Acknowledgements

78 This work has been supported by ACCESS-NRI. This research was undertaken with the
79 assistance of resources from the National Computational Infrastructure (NCI Australia), an
80 NCRIS enabled capability supported by the Australian Government.

81 References

- 82 Banihirwe, A., Long, M., bonnland, Grover, M., Kent, J., Bourgault, P., Squire, D., Busecke,
83 J., Spring, A., Schulz, H., Paul, K., RondeauG, & Kölling, T. (2023). *Intake/intake-esm:*
84 *v2023.07.07* . <https://doi.org/10.5281/zenodo.8124292>
- 85 Banihirwe, A., Long, M., bonnland, Kent, J., Grover, M., Bourgault, P., Spring, A., Busecke,
86 J., Paul, K., RondeauG, & Kölling, T. (2022). *Intake/intake-esm: v2022.9.18* . <https://doi.org/10.5281/zenodo.7089684>