2019-10-08

CARDAT Infrastructure Report

By Ivan Hanigan & Christy Geromboux, The University of Sydney School of Public Health.

Overview

The Centre for Air pollution, energy and health Research (CAR) is a Centre of Research Excellence funded by the National Health and Medical Research Council. The centre brings together more than 30 researchers at the forefront of their fields, based in seven of Australia's leading universities. CAR is funded by NHMRC from 2017 to 2021.

CAR is the only group of its kind nationally to bring together researchers focusing on health impacts of air pollution, and new versus traditional forms of energy. The centre supports teams of researchers in the fields of epidemiology, exposure assessment, toxicology, chemistry, biostatistics and clinical respiratory medicine to pursue collaborative projects and to develop their capacity. Our centre's vision for a healthier community is the driving force behind our research.

CAR's Data and Analysis Technology (CARDAT)

The CAR Data and Analysis Technology (CARDAT) project aims to deliver a collection of IT infrastructure that enables easy data sharing and reuse, and reproducible data analysis to CAR staff and affiliates. This infrastructure consists of four components:

- the CAR Data Inventory (a catalogue of available datasets and their associated metadata,
- a cloud based file store (CloudStor) which enables shared access to datasets, and the ability to sync data automatically,
- an online platform for data analysis (CoESRA) which provides a secure environment for reproducible, collaborative data analysis, and
- an internal wiki site containing documentation and examples for users.

The shared datasets are described in our data inventory. The data can be accessed via the file store on CloudStor or accessed and analysed via the online virtual environment, CoESRA. The CARDAT Data Platform is shown in a schematic Figure 1.

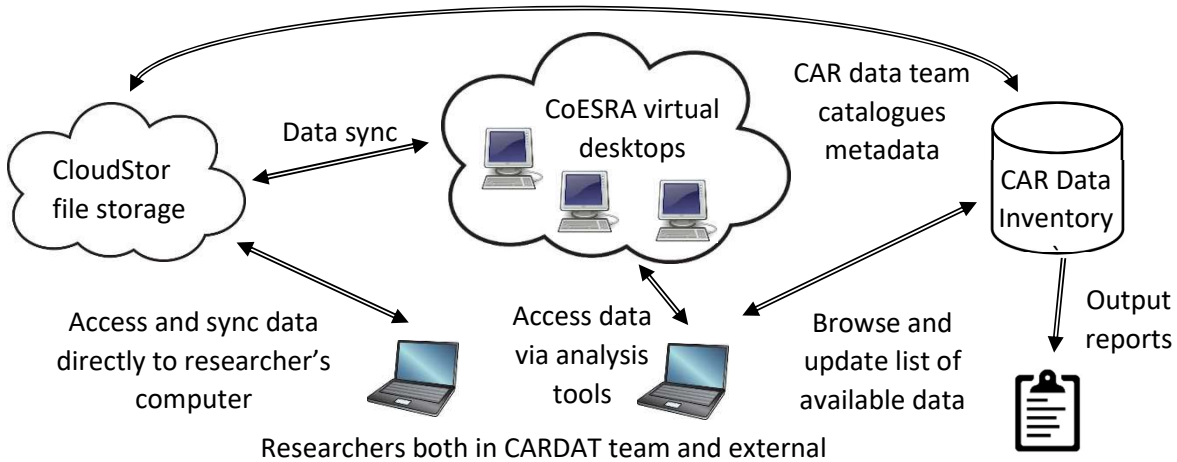

Figure 1: CARDAT data platform overview. Data can be downloaded directly from CloudStor or accessed via the virtual desktops on CoESRA. The CARDAT Data Inventory lists all the data available and produces reports.

User management

Access rules (who can request and who can grant access) are governed by the CARDAT user access guidelines document. The CAR Data Management team can be contacted via a shared email inbox (car.data@sydney.edu.au) and is responsible for enabling access requests as per access guidelines.

CARDAT IT infrastructure

CAR Data Inventory

Metadata records of all datasets are managed through a custom built opensource web interface called the CAR Data Inventory. Alternative proprietary and opens source options were investigated, however, most of them were too expensive or not functionally suitable.

The CAR Data Inventory was released as free open source software using a creative commons licence. This tool creates a catalogue of metadata that uses the Ecological Metadata Language (EML) which is an international metadata standard (https://knb.ecoinformatics.org/external//emlparser/docs/index.html).

This software is pre-installed on the CoESRA platform, and can also be downloaded from the following website: https://github.com/ivanhanigan/data_inventory.

The full catalogue of data is automatically published as a pdf and website, and is available through our website https://cardat.github.io.

The Data Inventory is used:

1. to track and document the details of the datasets, including metadata, access restrictions and access requests; and
2. to generate reports on datasets and to produce dynamic catalogues of the available data.

Dataset storage (CloudStor)

CARDAT stores its datasets on the cloud based file storage CloudStor. This interface enables synchronisation of datasets between the cloud and local machines. It is also available through CoESRA.

CARDAT stores three types of data:

- Open data - data that has been published on both CloudStor and in the public domain. These data don't require any special access requirements.
- General data - data that is published to CloudStor and is available to all CAR users on request.
- Restricted data - data that is published to CloudStor but require permission from the Data Owner before access can be granted.

Data analysis platform (CoESRA)

Collaborative Environment for Scholarly Research and Analysis (CoESRA - https://coesra.tern.org.au) is an online platform that has been developed collaboratively between CAR and the University of Queensland. It is hosted on the Australian National eResearch Collaboration Tools and Resources project (https://nectar.org.au/).

This platform provides a collaborative environment that supports transparent and reproducible workflows and data analysis. It also synchronises with CloudStor, removing the necessity for researchers to download large datasets locally.

Internal Wiki

The CARDAT internal wiki site (https://osf.io/82sjz/) is hosted by Open Science Framework (OSF). It provides a central repository for:

- user guideline documents,
- examples of data analysis,
- background on the project, and
- data cleaning tips and examples.

In order to login to this site, users must be able to log in to OSF (either by using their ORCID, or by creating an account), and be given access by the CAR Data Management team (car.data@sydney.edu.au).

Contact

Ivan Hanigan PhD
Data Scientist (Epidemiology)
The University of Sydney
University Centre for Rural Health
School of Public Health
and
Centre for Air pollution, energy and health Research (CAR)
Woolcock Institute of Medical Research
Ph: 0428 265 976
Email: ivan.hanigan@sydney.edu.au