Science & Technology Facilities Council
Rutherford Appleton Laboratory

Bringing NCAS data to the masses:

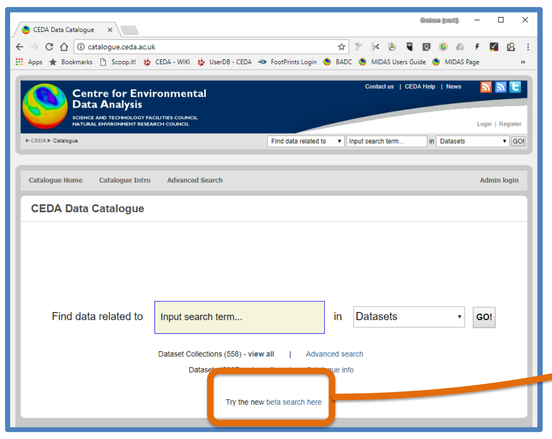
CEDA Catalogue developments and service integrations in 2017

Graham Parton, Ag Stephens, Joe Singleton, William Tucker & Thomas Cairnes
Centre for Environmental Data Analysis (CEDA), STFC Rutherford Appleton Laboratory, Oxfordshire, OX11 0QX
graham.parton@ncas.ac.uk

Introduction

The Centre for Environmental Data Analysis (CEDA) host an archive of over 180 million files across 5000 data sets. Enabling users to find these data and to ensure NCAS meets its EU Inspire requirements is a fundamental operation of CEDA.

Over the last year CEDA has rolled out a number of service updates to aid data discoverability and provide further integration with other CEDA services and provide address specific cataloguing needs for the aircraft, modelling and AMF communities.

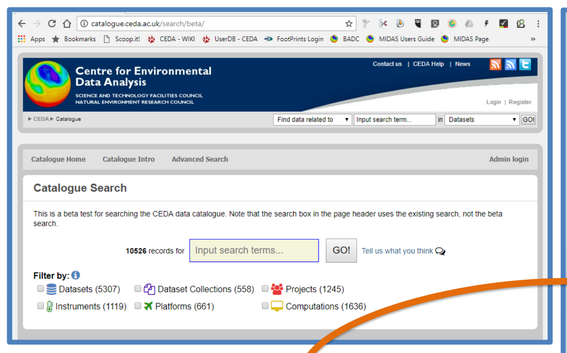


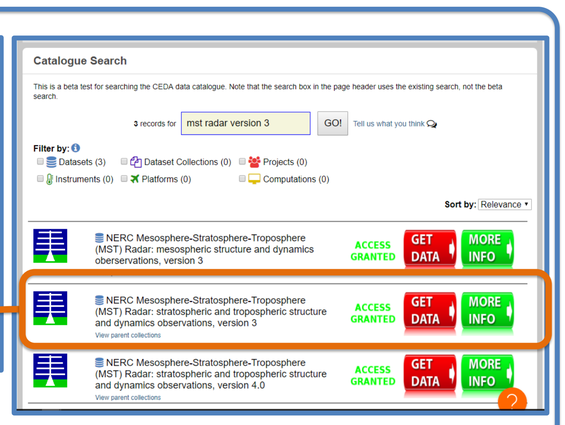
Improved Catalogue Search

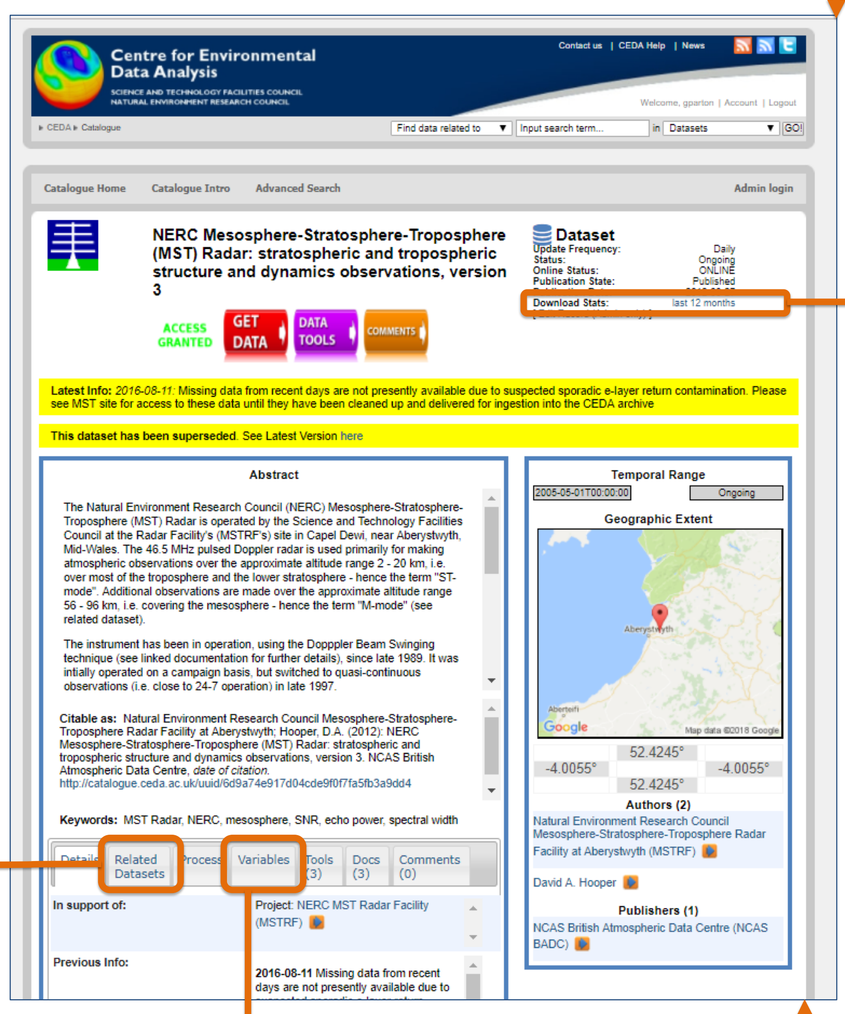
A new (beta) catalogue search tool has been produced based on the Django Haystack module, utilising off-the-shelf search engine technology (ElasticSearch). This is already giving our users much improved response times, search accuracy and filtering whilst making it easier and more intuitive to search for data and information via the various record types in the catalogue.

This approach will allow CEDA to index additional aspects of dataset records for more powerful search and filtering options in 2018. These include:

- Geo-spatial searching
- Temporal searching and filtering
- Variable information (see below)

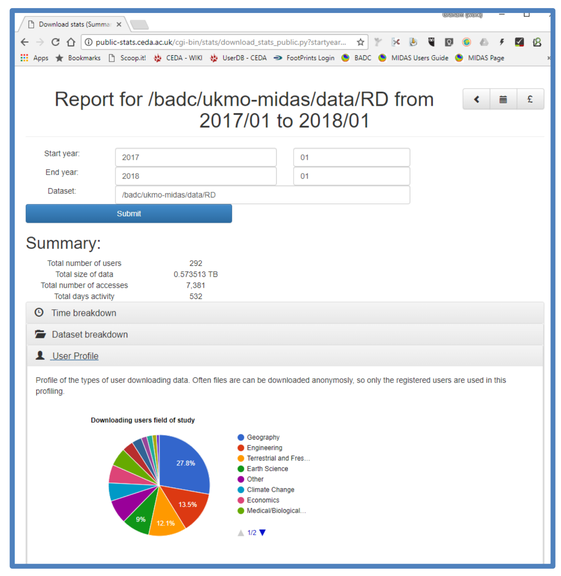






Usage Statistics

Data providers wishing to report on usage of their datasets can now find this information quickly and easily through dedicated links on each dataset record in the catalogue. The easy and intuitive interface provides access to public download statistics from 2012 to present.
Note – it doesn't cover JASMIN direct access and fully public data are only represented by IP addresses.



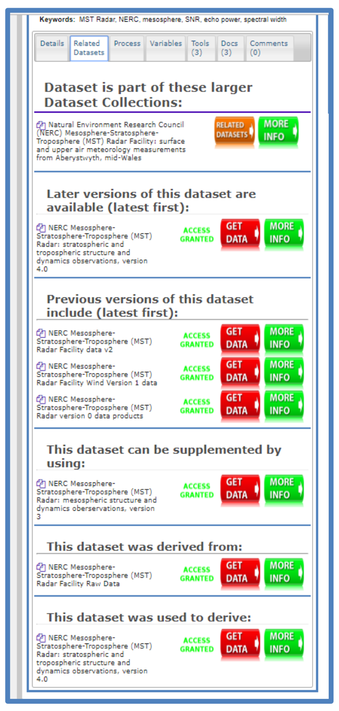
Connecting Related Data

Though datasets may be used in isolation there are often important connections between related datasets that are useful to know about. The CEDA catalogue's underlying data model already provides connections via related concepts such as data for the same Project or produced by the same Instrument, or within a wider Dataset Collection. However, in 2017 additional, specific dataset relationships were implemented covering:

- Data versioning/supersedence
- Derivation relationships
- Supplementary data
- Subsetted datasets

Where lists of datasets appear on catalogue records (i.e. Instrument, Platform, Computation, Project or Collection records) these are also broken down into groupings depending on where the dataset is in its life cycle:

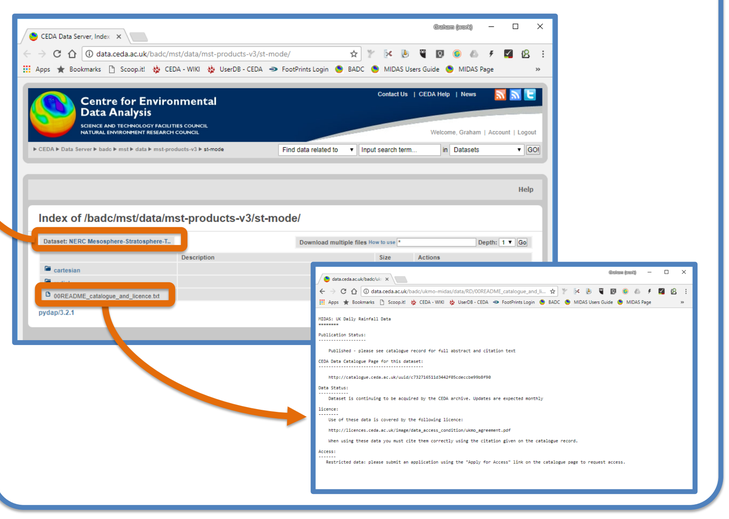
- Current datasets are listed first to ensure that users quickly find relevant datasets
- Upcoming datasets are then listed indicating when a dataset is in preparation for formal publication in the archive
- Historic datasets are ones which may have been marked as being held for historic/reference purposes or removed from the archive entirely.



Archive linking

An HTTP API for the catalogue now allows real-time polling for relevant dataset or collection level records based on a file or directory path. This approach enables other services to dynamically interrogate the catalogue service to improve automated interactions between CEDA systems.

Additionally, for those directly accessing datasets via JASMIN or FTP, signposting to useful catalogue and background information is available in "00README_catalogue_and_licencing.txt" files in for each dataset. This includes links to the relevant catalogue record alongside access, licencing information and data status.



Aircraft Data

CEDA holds over 2000 datasets related to FAAM, EUFAR and ARSF aircraft flights, most of which are covered by the Elastic Search indexes underpinning the EUFAR flight finder tool.

Thanks to these detailed indexing of files for each flight, and careful curation over the years capturing key information such as related Projects, the vast majority of these flights have been accurately catalogued for the first time, enhancing discoverability and connections with related datasets in the CEDA archive.

The FAAM, EUFAR and ARSF Flight Finder indexes were able to provide geographic and temporal information scraped from the archive, whilst project information was supplied from readme files curated alongside the flight information.

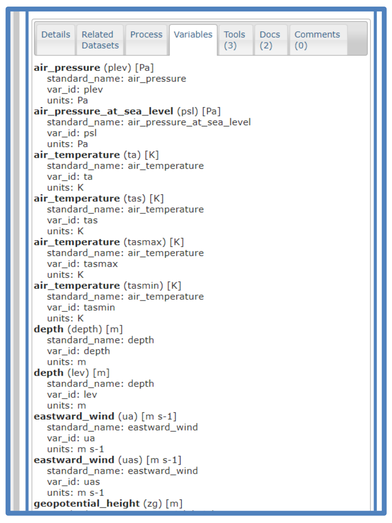
Where the index was unable to help, location names in the readme files were used to poll the Google Maps API to return likely geographic locations. Coarse temporal information was also determinable from directory naming conventions.

Archive indexing with ElasticSearch and Variables

Over 2017 CEDA have been refining a full, file-by-file index of CEDA's entire archive holdings using ElasticSearch. This initially included just file information such as size, location and format, but increasingly larger parts of the archive are now being scanned and variable information is being yielded directly from the file metadata.

At present we're looking at how best to aggregate this information up to the Dataset record level and then into our new Haystack catalogue search service. The information gathered will allow searching by:

- CF standard names
- MIP table variable IDs
- Longer parameter names

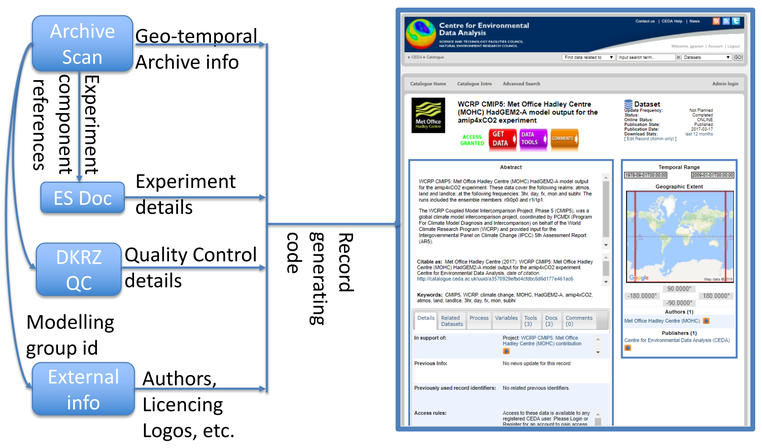


Additionally, other variable identifiers are captured alongside the units used, whilst the way information if linked preserves relationships between these various pieces of information for each variable.

Climate Modelling data

Cataloguing CMIP5/CMIP6/CCMI data

Climate modelling programmes produce a large proportion of the datasets in the CEDA archives with over 1800 to date, requiring automated record generation tools to allow cataloguing to scale with these ever increasing numbers of modelling datasets. Cataloguing the CMIP5 dataset in CEDA archives has paved the way to further automation, though these are highly reliant on structured sources of required metadata.



AMF holdings in the CEDA archive

As part of the NCAS Observations Data Project CEDA have been working with members of the NCAS Atmospheric Measurement Facility (AMF) to improve the coverage of AMF holdings in the CEDA archive and catalogue.

Following comparisons with AMF instrument information and checking archive holdings CEDA produced a series of bespoke catalogue views allowing users to dynamically obtain lists of NCAS observation datasets from both AMF itself and other NCAS observation facilities. These can be viewed at the following locations:

- AMF instrument datasets by Campaign : http://catalogue.ceda.ac.uk/listings/amf/campaigns
- AMF instrument datasets by Instrument: http://catalogue.ceda.ac.uk/listings/amf/instruments
- Long term observations by Facility (excepting FAAM): http://catalogue.ceda.ac.uk/listings/amf/longterm

For further details on the NCAS Observations Data Project, including details on metadata conventions, visit:

https://sites.google.com/a/ncas.ac.uk/ncas-data-project/

National Centre for Atmospheric Science
NATURAL ENVIRONMENT RESEARCH COUNCIL

National Centre for Earth Observation
NATURAL ENVIRONMENT RESEARCH COUNCIL

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