Abstract ID : **98**

STAC at CEDA - a scalable, standards-based search system

Content

The Centre for Environmental Data Analysis (CEDA) stores over 20 Petabytes of atmospheric and Earth observation data. Sources for the CEDA Archive include aircraft campaigns, satellites, automatic weather stations and climate models, amongst many others. The data mainly consists of well-described formats such as netCDF files but we also hold historical data where the format cannot be easily discerned from the file name and extension.

CEDA are investigating the SpatioTemporal Asset Catalogue (STAC) specification to allow for user interfaces and search services to be enhanced and facilitate interoperability with user tools and our partners. We are working to create a full-stack software implementation including an indexing framework, API server, web and programmatic clients, and vocabulary management. All components are open-source so that they can be adopted and co-developed with other organisations working in the same space.

We have built the "stac-generator", a tool that can be used to create a STAC catalog, which utilises a plugin architecture to allow for more configurability. A range of input, output, and extraction methods can be selected to enable data extraction across the diverse archive data and its use by other organisations. Elasticsearch was chosen to host the indexed metadata because it is performant, highly scalable and supports semi-structured data - in this case the faceted search values related to different data collections. As STAC's existing API was backed by an SQL database this called for the development of a new ES backed STAC API, which has now been merged back into the community developed API as an alternate database backend. We have also developed several extensions to the STAC framework to meet requirements that weren't met by the core and community functionality. These include an end-point for interrogating the facet values, as queryables, and a free-text search capability across all properties held in the index.

The developments of our search system has also included pilots for the Earth Observation Data Hub (EODH) and a future version of the Earth System Grid Federation (ESGF) search service, in which we have created an experimental index containing a subset of CMIP6, CORDEX, Sentinel 2 ARD, Sentinel 1, and UKCP data to investigate performance and functionality.

With the increasing demand on cloud-accessible analysis-ready data we are seeing in several of our upcoming projects. We have started to explore Kerchunk a lightweight non-conversion approach for referencing existing data, which works with open-source python packages like fsspec and xarray. And are looking to integrate this with our STAC work.

It is the aim of project to increase the interoperability of our search services, as well as foster collaboration with other organisation who share our goals. Additionally, it is hoped that this work will allow for greater and easier access to the data held at CEDA.

Topic

Data innovations: Data Analytics, Sensitive Data/FAIR Data

Primary author: EVANS, Rhys

Presenter: EVANS, Rhys

Contribution Type: Short Talk

Submitted by **EVANS, Rhys** on **Monday, 6 May 2024**