Unifying specialized databases for a central search portal – the HCDC approach

Linda Baldewein, Helmholtz-Zentrum Geesthacht (Germany), linda.baldewein@hzg.de
Ulrike Kleeberg, Helmholtz-Zentrum Geesthacht (Germany), ulrike.kleeberg@hzg.de
Dietmar Sauer, Helmholtz-Zentrum Geesthacht (Germany), dietmar.sauer@hzg.de
Robin Luckey, Helmholtz-Zentrum Geesthacht (Germany), robin.luckey@hzg.de
Philipp S. Sommer, Helmholtz-Zentrum Geesthacht (Germany), philipp.sommer@hzg.de
Housam Dibeh, Helmholtz-Zentrum Geesthacht (Germany), housam.dibeh@hzg.de

Coastal research is characterized by a large variety of overlapping research fields. Each of them has its own requirements for data formats, storage and retrieval solutions. For example, observational oceanographers use sensors for a continuous collection of near real-time data, while climate modelers generate terabytes of high resolution model data, oftentimes on unstructured grids. Providing access through a single portal for a wider scientific community as well as stakeholders to these different data sets is a challenging task for data managers and providers.

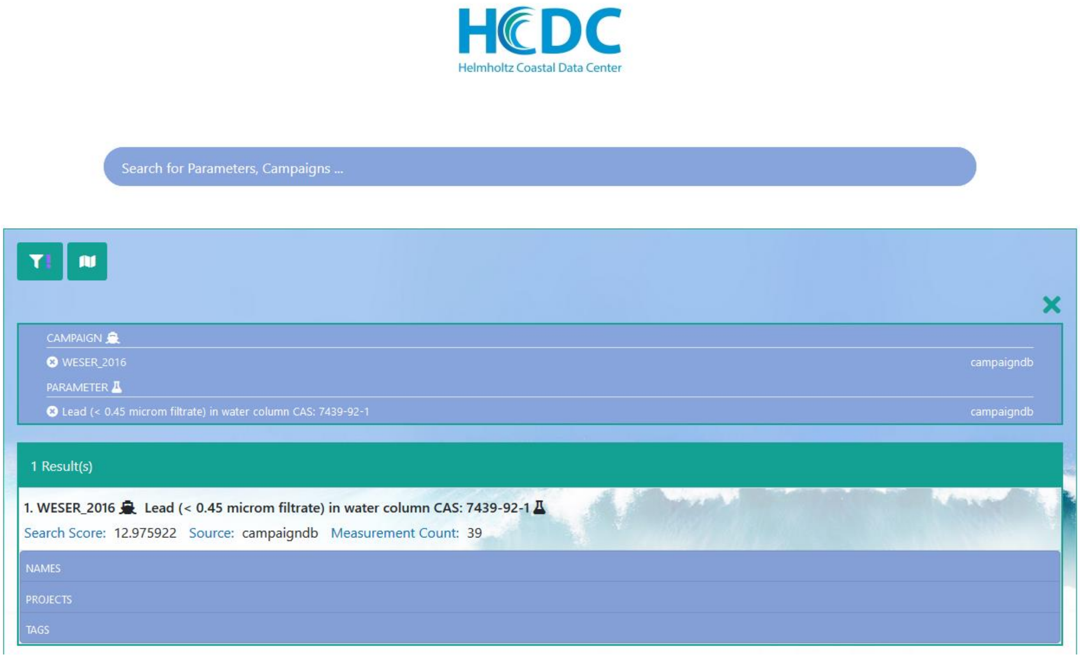
At the Helmholtz Coastal Data Center (HCDC) we have three main branches of data with their own storage solutions. Biogeochemical campaign data, received as either ASCII or Excel file from the author, are stored in a relational database. Metadata for biogeochemical data are available for each individual measurement. The second type, near real-time observational data stored in ASCII files, are sent via web transfer protocols directly from the sensors, on for example ferries or underwater knots, into a database set up for time series data. Related metadata are aggregated for each platform that is sending data. The third data type is model output. They are stored in NetCDF files directly at the site of the High Performance Computing system for Earth system research at our partner institute, the German Climate Computing Center (DKRZ), and the metadata are available through the data publishing platform CERA (cera-www.dkrz.de).

Before starting the journey of unifying all three data systems in a single search portal, we questioned our stakeholders, mainly colleagues from our research institute, other research centers and public authorities, and defined the required functionalities of the search portal. The users desire a portal with a search engine that “is like Google”, meaning a single search field that allows them to find all data, like a one-stop shop. At the same time they want to be able to filter by time, geographic locations and a variety of other metadata fields.

Such a search process is only possible if the metadata systems are unified, because it simplifies the process compared to searching for data across three different metadata solutions with different technical and logical setups. Thus, a separate metadata database is set up, where metadata are aggregated per data source. A list of common metadata fields is created to which all three systems are mapped as well as possible. For example all three metadata systems contain parameters, measurement / file counts and the data source, e.g. a platform, model or campaign. This metadata database is continuously updated to reflect changes in the three source databases / storage locations.

The number of aggregated datasets for the combined sample, observational and model data is in the range of several millions. The use of a highly scalable full-text search engine on top of the relational database facilitates a real-time search experience. Thus, the metadata are transferred into an Elasticsearch cluster, an open-source NoSQL search engine.

To generate a “Google”-like search feeling while using filters at the same time, the intelligent criteria search has been developed. While the user is typing a search term on the responsive website, a fuzzy search is started over all keywords in the Elasticsearch cluster. The most likely matches as well as the search criterions to which they belong are displayed, independent of the database from which they originate. The user can now select the keyword from the resulting list and use it as a filter. The search results are displayed immediately and can be further refined by adding more filters (see Figure 1).


HCDC
Helmholtz Coastal Data Center
Search for Parameters, Campaigns ...
NAMES
PROJECTS
TAGS

CAMPAIGN
WESER_2016 campaigndb
PARAMETER
Lead (< 0.45 microm filtrate) in water column CAS: 7439-92-1 campaigndb

1 Result(s)

1. WESER_2016 Lead (< 0.45 microm filtrate) in water column CAS: 7439-92-1
Search Score: 12.975922 Source: campaigndb Measurement Count: 39

Figure 1: Unified HCDC data portal for different databases. A single search field at the top is followed by selected search criteria and the result list.

The resulting datasets are then listed to best match the searched terms. With a single click, the dataset can be added to the cart and the download can be requested. Different output file formats are offered to the user so that each specialist can continue working in their desired format. The user will receive an e-mail once the dataset is ready for downloading.

The data portal is build using state-of-the-art web frameworks and techniques. The front end is a progressive webapp based on Angular 8. It is set up as a series of services and components. The services store the application's state and provide the means of inter-component information transfer and communication with the backend. The services are being injected into predominantly stateless view components by the angular dependency injection system. The backend provides an API for searching for Metadata in the Elasticsearch cluster. It is a stand-alone, lightweight server based on ExpressJs. Its functionalities include the real-time, prefix-based, fuzzy completion suggestion for the search in the front-end, the data aggregation to provide an overview over the existing data measurements that fit the filter and a functionality for downloading these measurements in various data formats. In order to allow minimal response time albeit the multitude of requests necessary, JavaScript promises are heavily made use of for the means of asynchronous communication.

The unified HCDC data portal provides a single access point to three different data bases and storage locations for the individual user. As such, it serves as a real-life example for users and data managers how to access various subject specific databases through a single and intuitive entry point. The search field allows for filtering and downloading data across different coastal research disciplines. The data portal, using state-of-the-art web technologies, simplifies data access for all stakeholders and prevents time consuming searches across different platforms and heterogeneous user interfaces.