Indis

International conference on Marine Data and Information Systems



27-29 May 2024 🗮





neosc



Next-Generation Data Lake and Subsetting Solution

<u>Peter Thijsse</u>, Robin Kooyman, Dick Schaap, Tjerk Krijger (all MARIS)









Challenge

- Data organised in millions of files (e.g. >2.8M for SeaDataNet) with metadata
- Human user interfaces exist to query metadata, select files, order and download.
- But how to optimise systems like SeaDataNet CDI, ARGO, .. for Machine2Machine access to subsets, for access by VRE's, Jupyter Notebooks, by other applications?
- How to go from files to serving exactly data as needed, on the fly, in one file, as a true "Data lake" component?

NEW SEARCH REFINE	SEARCH S	EARCH RESULTS		SUMMARY	TIMESERIES
Filter search		1	island reserve	i i Monett - 1003 j me	e i Prev i Sever i Lave
You searched for:	Reset all	101107			
Measuring area type:		DATE OF STREET	Include Course states		u
point		CRAINSET RAME	Protoc Datas Toros		
Discipline (POR):		COUNTRY ORDENATOR	Platineri - edelaridei		
Physical oceanography		STATIONE	discusto andre sometico.		
Descention in the second	and the second second	SUPER	D.		P.W
EXPORT RESULT SAVE	SARKA 🖓	SOLUT	Include Annual Contra		12
INPUT PIELDS		COUNTRY ORIGINATOR	Remote Lateration		100
Free search O		CONTRACT CARD	1000000		
		SIMO DATE	19761008		ar ar ar
Date search (D		SIGNAL PROPERTY GOAT THE	and the water samplers.		1.1
From momental To m	even ald	DATASET NAME	AUTTING No.		100
		COLUMN COLUMN	Automobile .		100
Geographic search O		COONTINE ORIGINATION	manual resources		
C Search within house line h	20.0	SOAT DATE	19770407		
- sector warm outling t	-	SHE ADMENT / GEAR TYPE	tostrene water samplers		
SEARCH R	ESET	see	U		100
SEA REGIONS		EADAGET NAME	HACOC, HARDA, KITO		100
Contraction of the		COUNTRY ORIGINATOR	Ukrame		
Allantic Ocean	(993946)	START DATE	19710804		Burgaro,
North Allantic Ocean Montheast Atlantic Ocean (All	(830769)	INSTRUMENT / GEAR TYPE	discrete water samplers		
	(537980)	SLUCY			
Mediterranean Sea	(360347)	DATASET NAME	ANGOC_BOTH_R010		1000
Battic Sea	(336438)	COUNTRY ORIGINATOR	Ukrame		A.8/a.;
+ More		START DATE	19710804		(240.2)
DISCIPLINE (POR)		INSTRUMENT / GEAR TYPE	discrete water samplers		- and a
		SELECY	0		24 90
Administration and dimensio	(1391443)	DATASET NAME	RNDOC_BATHE_ETTE		and the
Bolorical oceanography	(276157)	COUNTRY ORIGINATOR	Uleaine		
Marine pictury	(70326)	START DATE	19710723		
Enverorment	(33886)	INSTRUMENT / GEAR TYPE	discrete water samplers		Madra, in dorpe
Human activities	(33886)	SELECT	0		0
Terrestrui	(17308)	EATASET NAME	PNDDC_DUTH_8310		
Atmosphere	(4733)	COUNTRY ORIGINATOR	Likeline		Everyone
Feberies and aqueculture	(52)	START DATE	19710804		1
		INSTRUMENT / GEAR TYPE	discrete water samplers.		
www.antitikesborg	an a	SELECT			Parameters.
Water column temperature a	(2289128)	DAZASET NAME	RNDDC Bottle #310		886470
Administration and dimensio	(1391443)	COUNTRY ORIGINATOR	Ukraine		887700
Dissolved geses	(677615)	START DATE	19710804		CDOM.
Nutrients	(343824)	INSTRUMENT / GEAR TYPE	absorate water samplers.		Deployment
					3024
					2021
					2023
					ew1
					Position & da
					Good Store
					River and

Emodnet Physics





Example user query

• Request:

Give me all the temperature data in the North Sea, from 2010-2012, in degrees celsius, at a depth from 0-50 m.

• Response: **One** NetCDF file containing exactly this data. On the fly, directly usable in a Jupyter notebook and for HPC, with all original metadata accessible.

=> In these cases **Beacon** provides the solution

Beacon in a nutshell

- High Performance Data Lake
- Written in Rust + C from scratch
- Easy to deploy (Docker)
- On the fly subsetting out of millions of datasets
- Harmonized single file output (Eg. 1 netcdf file)
- Powerful query capabilities
 - Filter on:
 - Ranges
 - Polygons
 - Metadata

	2	
	2	
	2.5	
(sp	2	
(Seonci	1.5	
Duration		
	1	
	0.5	
	0	10_000
		Index Tra

PERFORMANCE



Beacon performance for extracting dataset of approx. 10M datapoints (max), from increasing number of datasets, containing billions of datapoints.

How does Beacon work?

- Build indexes
- Reading only the necessary data
- Harmonization of files types
- Filtering, calls to the API
- Conversions of units, aggregation into parameters
- Writing output to various formats (a.o. NetCDF, Beacon binary format)



Beacon architecture

Current State

- Beacon pre-registration available at:
 - https://beacon.maris.nl/
 - Github available for community: sharing notebooks, python scripts, drivers
- Several Beacon deployments running for tests as part of BlueCloud2026
 - Argo
 - SeaDataNet CDI
 - ERA5/ORAS5 (sample set)
 - EMODnet Physics (operational)
 - ..
- Example notebooks created, and more will follow
- EOSC-Future demonstrator for subsetting CDI and ARGO, co-location with CMEMS

Beacon

The high-performance climate & marine O data lake solution used to store and subset millions of NetCDF datasets and terabytes of data with powerful query possibilities and lightning 4 fast retrieval.





Demo

Beacon is the data lake engine that powers the marine data viewer 😤 and provides it with all the millions of observational data records on the fly 🞇. Together with powerful broker features













Usage example in Notebook: Creating an ARGO product in Pluto

Perform request and write into netCDF file

- If the file is already there, the Beacon API is not called.
- The file name is created according to the region name, the period of interest and the depth range.

```
\odot

 begin

        filename = joinpath(datadir,
        "Argo_$(parameter)_$(regionname)_$(Dates.format(datestart, "yyyymmdd"))-
        $(Dates.format(dateend, "yyyymmdd"))_$(Int(mindepth))-$(Int(maxdepth))m.nc");
        @info("Data will be written in file:\n$(filename)")
  .
        if isfile(filename)
  .
            @info("File already downloaded")
  .
        else
  .
            @time open(filename, "w") do io
  .
                r = HTTP.request("POST", "$(beaconURL)/api/query",
  .
                ["Content-Type" => "application/json"], query,
  .
                response_stream=io);
                @info(r.status)
            end
  .
  .
  .
        end;
  .
        @info("NetCDF file size: $(round(filesize(filename)/1000^2, digits=1))M")
  .
  end;
 ① Data will be written in file:
    ../data/Argo_sea_water_temperature_Canary_Islands_20000101-20101231_0-400m.nc

 200

 ① NetCDF file size: 1.8M
 >____
       2.249019 seconds (2.16 k allocations: 2.017 MiB)
Read netCDF content
```







Interactive map (Leaflet)

Leaflet

The coordinates obtained from the input file (netCDF) are directly fed into Leafle Changing the region or the period of interest will automatically update the interest



<pre>filename);</pre>
0
et. active map.
enStreetMap rto servation locations
EMODnet Pathymetry

Notebook output: Creating an ARGO product in Pluto

🔭 Create the plot

The figure title is set according to the depth and the time period

field2D = @view field[:,:,depthindex,timeindex];



Thanks Charles!





EOSC-Future demo (Beacon as back-end and supporting server side drawing)



https://eosc-future.maris.nl/

Current release (V0.9)

- Supports ODV pipeline for QC for the Workbenches
- Improved additional metadata handling and query options \bullet
- New Beacon feature: Dynamic Chunking •
 - supports high-performance queries on large datafiles (remote sensing, products, etc).
 - Can extract e.g. timeseries on certain location
 - Like ZARR but more powerful β and faster 4
- Improved Stability of the Beacon Platform \bullet

What's next?

- Deployment of Beacon for various data infrastructures under Blue-Cloud2026
- Integration and use into BC2026 and Fair-Ease VRE's
- Towards first stable release:
 - Available for academic use
 - Open source available: API, Python scripts, Notebooks.
- Slowly rolling out the software into the community, and building a community

Questions and further information?

- Email to:
 - <u>dick@maris.nl</u>
 - peter@maris.nl
 - <u>robin@maris.nl</u>
- Join and connect to us via <u>https://beacon.maris.nl</u>





Indis

International conference on Marine Data and Information Systems



27-29 May 2024 🗮

