SeaDataCloud

Upgrading the Common Data Index (CDI) Data Discovery and Access service

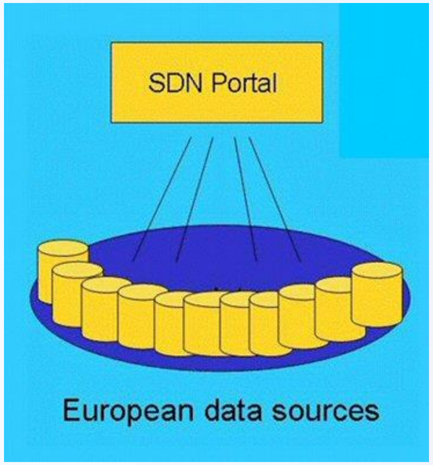
Dick M.A. Schaap - SeaDataCloud technical coordinator (MARIS, Netherlands)

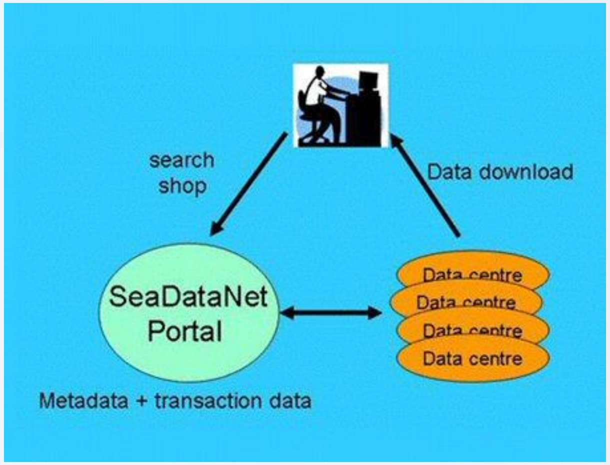
EUDAT conference, Porto, Portugal, 22-25 January 2018

sdn-userdesk@seadatanet.org – www.seadatanet.org

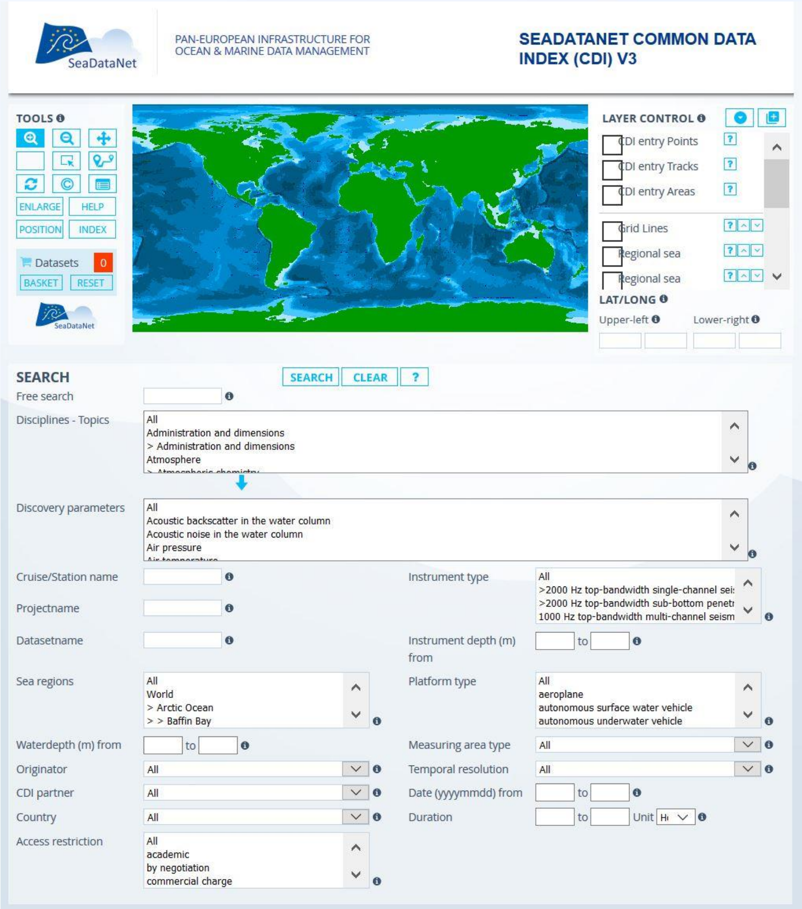
CDI Data Discovery and Access service

- one of the core services of the SeaDataNet portal
- It provides a highly detailed insight and unified access to the large volumes of marine and oceanographic data sets managed by the distributed data centres
- fine-grained index (ISO 19115 – ISO 19139) to individual data measurements (such as a CTD cast or moored instrument record)
- Giving access to associated data sets in unified formats (SDN ODV / NetCDF (CF))
- supported by Controlled Vocabularies, and Directories (EDMO, EDMERP, CSR, EDMED)





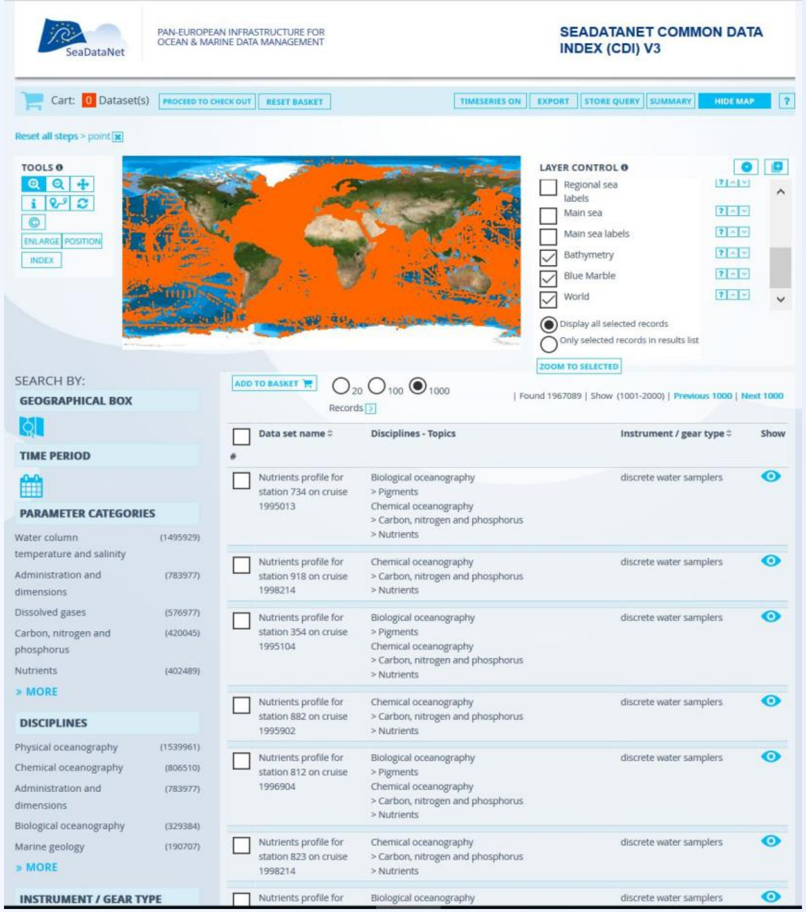
2 types of user interfaces



□ CDI entry Points
□ CDI entry Tracks
□ CDI entry Areas
□ Grid Lines
□ Regional sea
□ Regional sea

SEARCH

Free search	
Disciplines - Topics	All Administration and dimensions > Administration and dimensions Atmosphere
Discovery parameters	All Acoustic backscatter in the water column Acoustic noise in the water column Air pressure
Cruise/Station name	
Projectname	
Datasetname	
Sea regions	All World > Arctic Ocean > > Baffin Bay
Waterdepth (m) from	to
Originator	All
CDI partner	All
Country	All
Access restriction	All academic by negotiation commercial charge
Instrument type	All >2000 Hz top-bandwidth single-channel sei: >2000 Hz top-bandwidth sub-bottom penetr 1000 Hz top-bandwidth multi-channel seism
Instrument depth (m) from	to
Platform type	All aeroplane autonomous surface water vehicle autonomous underwater vehicle
Measuring area type	All
Temporal resolution	All
Date (yyyymmdd) from	to
Duration	to Unit H



□ Regional sea labels
□ Main sea
□ Main sea labels
☑ Bathymetry
☑ Blue Marble
☑ World

● Display all selected records
○ Only selected records in results list

○ 20 ○ 100 ● 1000 Records

| Found 1967089 | Show (1001-2000) | Previous 1000 | Next 1000

SEARCH BY:

GEOGRAPHICAL BOX

TIME PERIOD

PARAMETER CATEGORIES

Water column temperature and salinity	(1495929)
Administration and dimensions	(783977)
Dissolved gases	(576977)
Carbon, nitrogen and phosphorus	(420045)
Nutrients	(402489)

» MORE

DISCIPLINES

Physical oceanography	(1539961)
Chemical oceanography	(806510)
Administration and dimensions	(783977)
Biological oceanography	(329384)
Marine geology	(190707)

» MORE

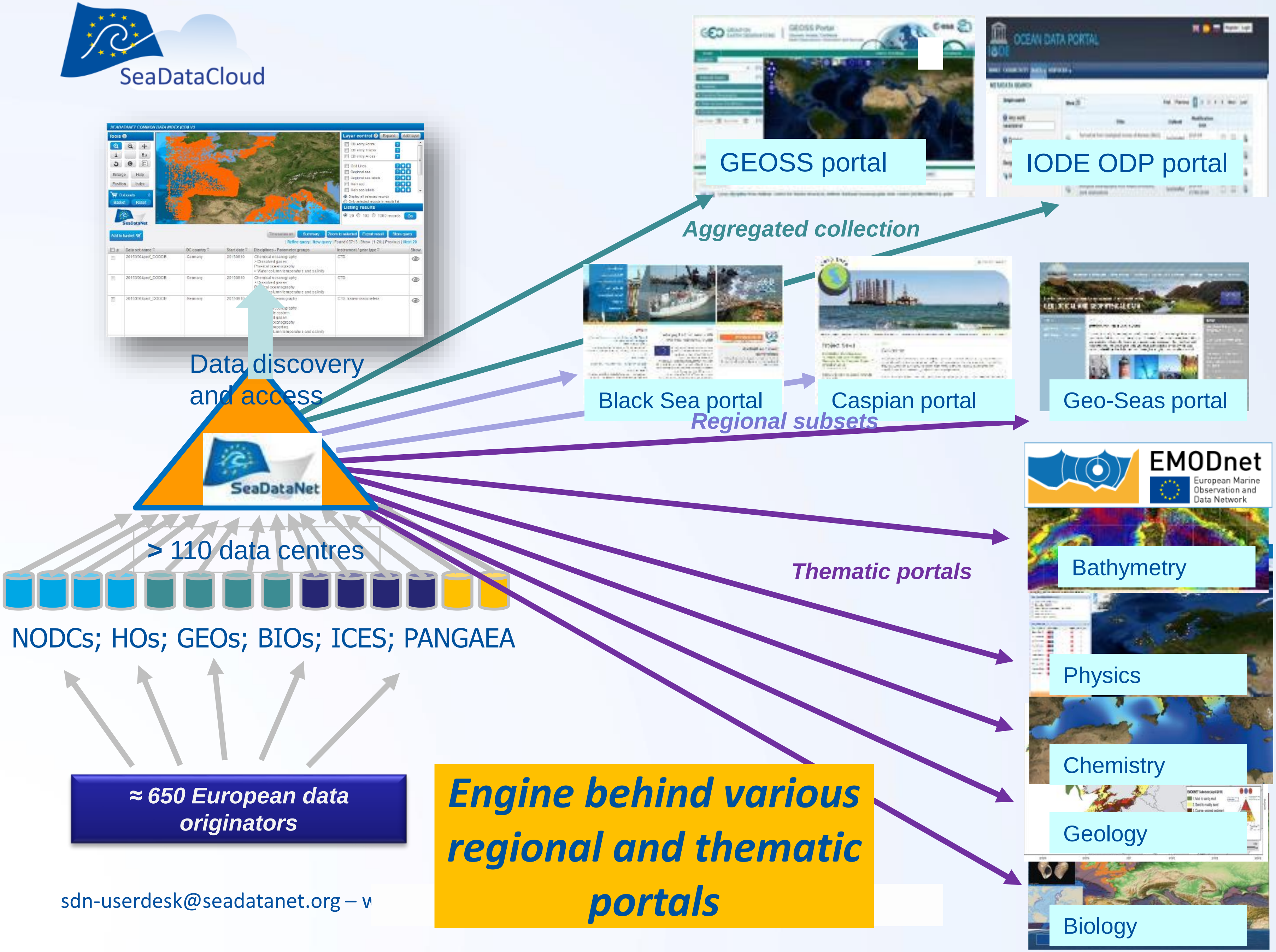
INSTRUMENT / GEAR TYPE

□ #	Data set name	Disciplines - Topics	Instrument / gear type	Show
□	Nutrients profile for station 734 on cruise 1995013	Biological oceanography > Pigments Chemical oceanography > Carbon, nitrogen and phosphorus > Nutrients	discrete water samplers	👁
□	Nutrients profile for station 918 on cruise 1998214	Chemical oceanography > Carbon, nitrogen and phosphorus > Nutrients	discrete water samplers	👁
□	Nutrients profile for station 354 on cruise 1995104	Biological oceanography > Pigments Chemical oceanography > Carbon, nitrogen and phosphorus > Nutrients	discrete water samplers	👁
□	Nutrients profile for station 882 on cruise 1995902	Chemical oceanography > Carbon, nitrogen and phosphorus > Nutrients	discrete water samplers	👁
□	Nutrients profile for station 812 on cruise 1996904	Biological oceanography > Pigments Chemical oceanography > Carbon, nitrogen and phosphorus > Nutrients	discrete water samplers	👁
□	Nutrients profile for station 823 on cruise 1998214	Chemical oceanography > Carbon, nitrogen and phosphorus > Nutrients	discrete water samplers	👁
□	Nutrients profile for	Biological oceanography	discrete water samplers	👁

Extended Search

Quick (facet) Search

SeaDataCloud



Aggregated collection

GEOSS portal

IODE ODP portal

Regional subsets

Black Sea portal

Caspian portal

Geo-Seas portal

Thematic portals

EMODnet European Marine Observation and Data Network

Bathymetry

Physics

Chemistry

Geology

Biology

Data discovery and access

SeaDataNet

> 110 data centres

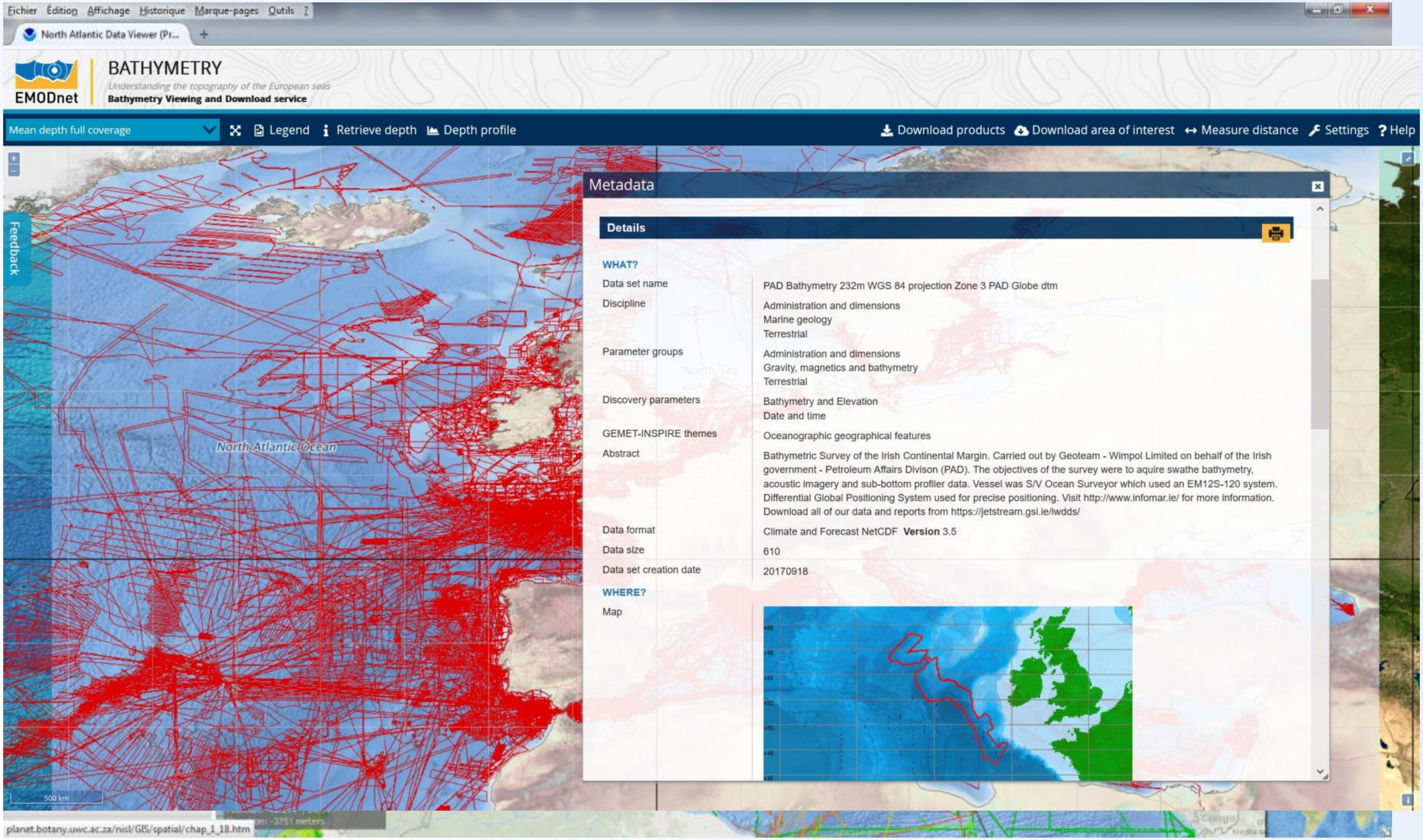
NODCs; HOs; GEOs; BIOs; ICES; PANGAEA

≈ 650 European data originators

Engine behind various regional and thematic portals

sdn-userdesk@seadatanet.org – w

Web services



Metadata

Details

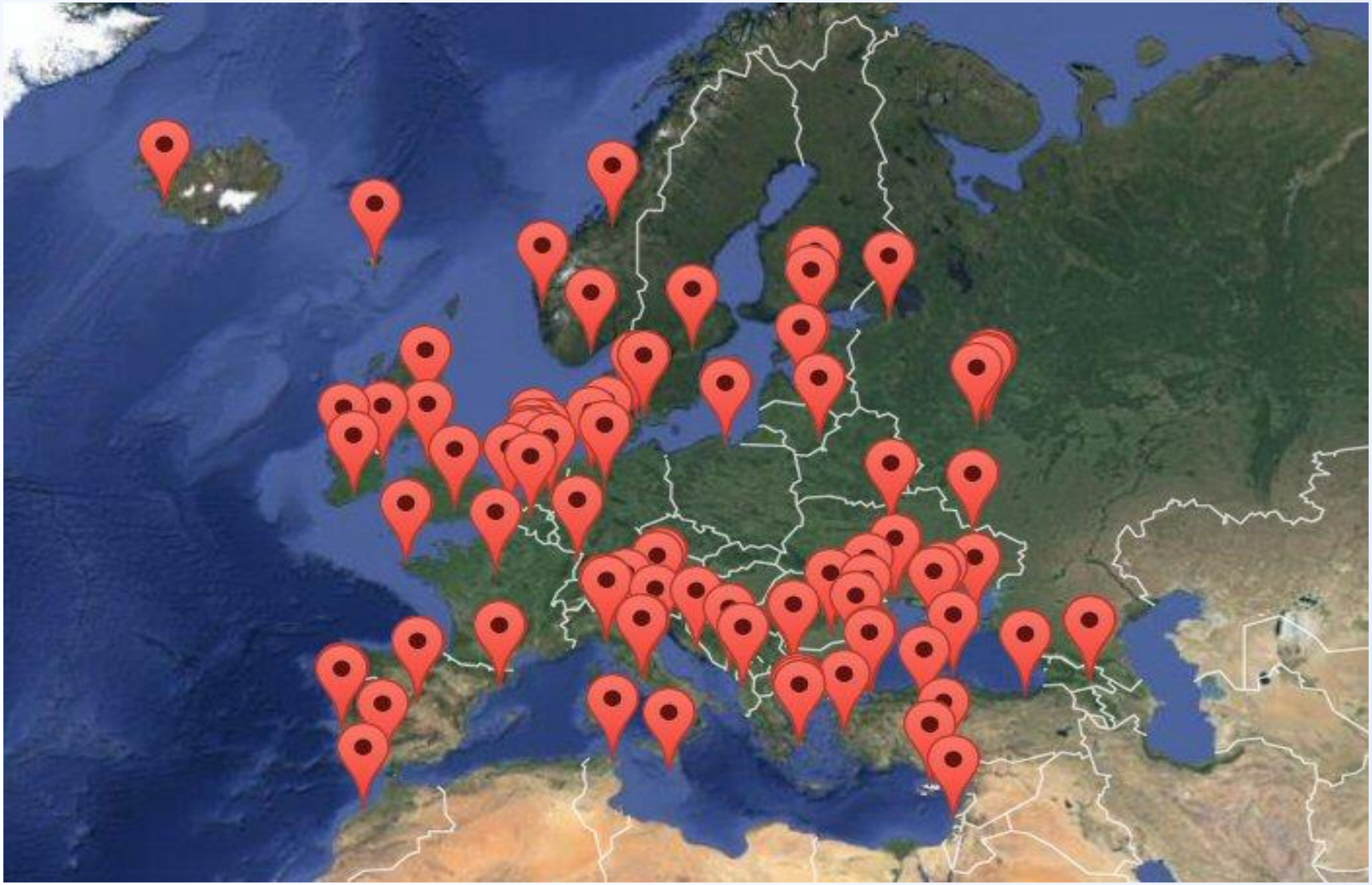
WHAT?

Data set name	PAD Bathymetry 232m WGS 84 projection Zone 3 PAD Globe dtm
Discipline	Administration and dimensions Marine geology Terrestrial
Parameter groups	Administration and dimensions Gravity, magnetics and bathymetry Terrestrial
Discovery parameters	Bathymetry and Elevation Date and time
GEMET-INSPIRE themes	Oceanographic geographical features
Abstract	Bathymetric Survey of the Irish Continental Margin. Carried out by Geoteam - Wimpol Limited on behalf of the Irish government - Petroleum Affairs Divison (PAD). The objectives of the survey were to aquire swathe bathymetry, acoustic imagery and sub-bottom profiler data. Vessel was S/V Ocean Surveyor which used an EM12S-120 system. Differential Global Positioning System used for precise positioning. Visit http://www.infomar.ie/ for more information. Download all of our data and reports from https://jetstream.gsi.ie/iwdds/
Data format	Climate and Forecast NetCDF **Version** 3.5
Data size	610
Data set creation date	20170918

WHERE?

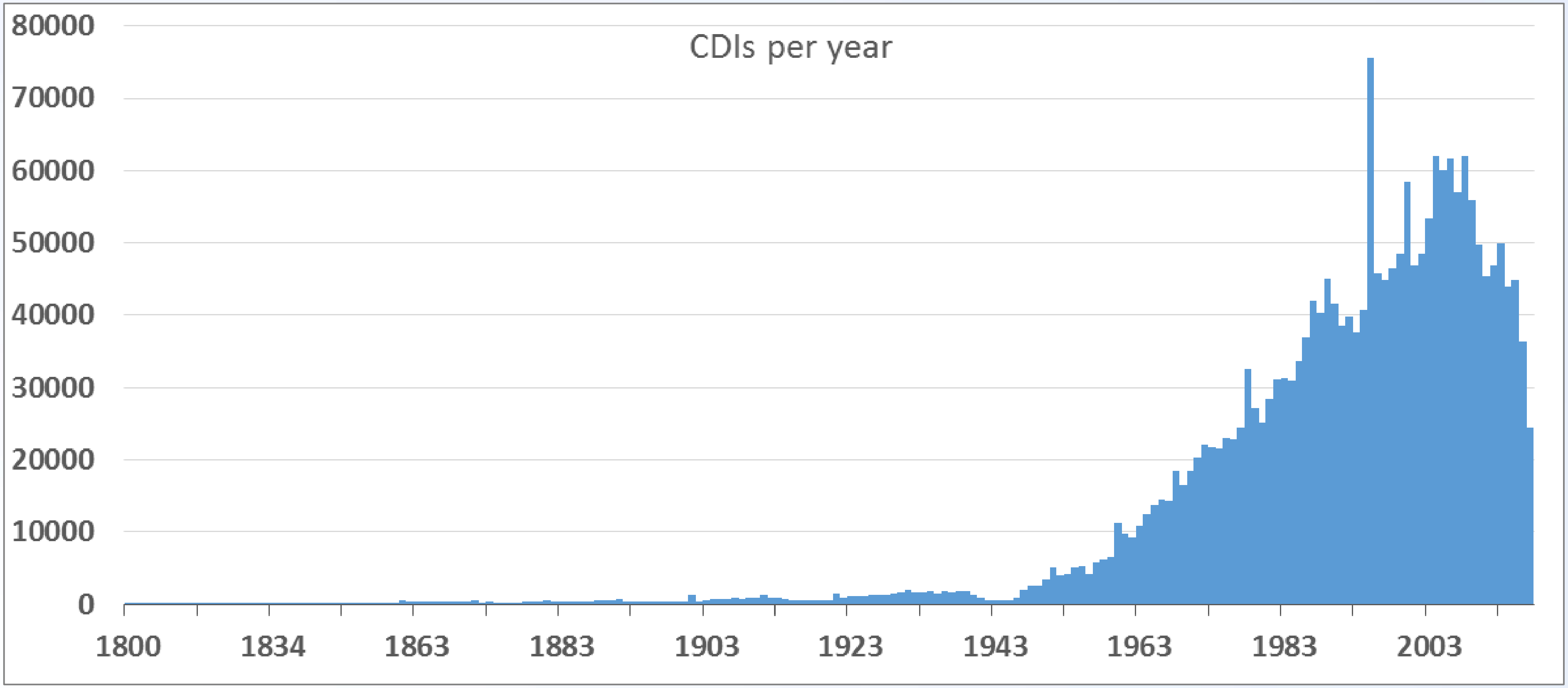
Map

110 connected data centres

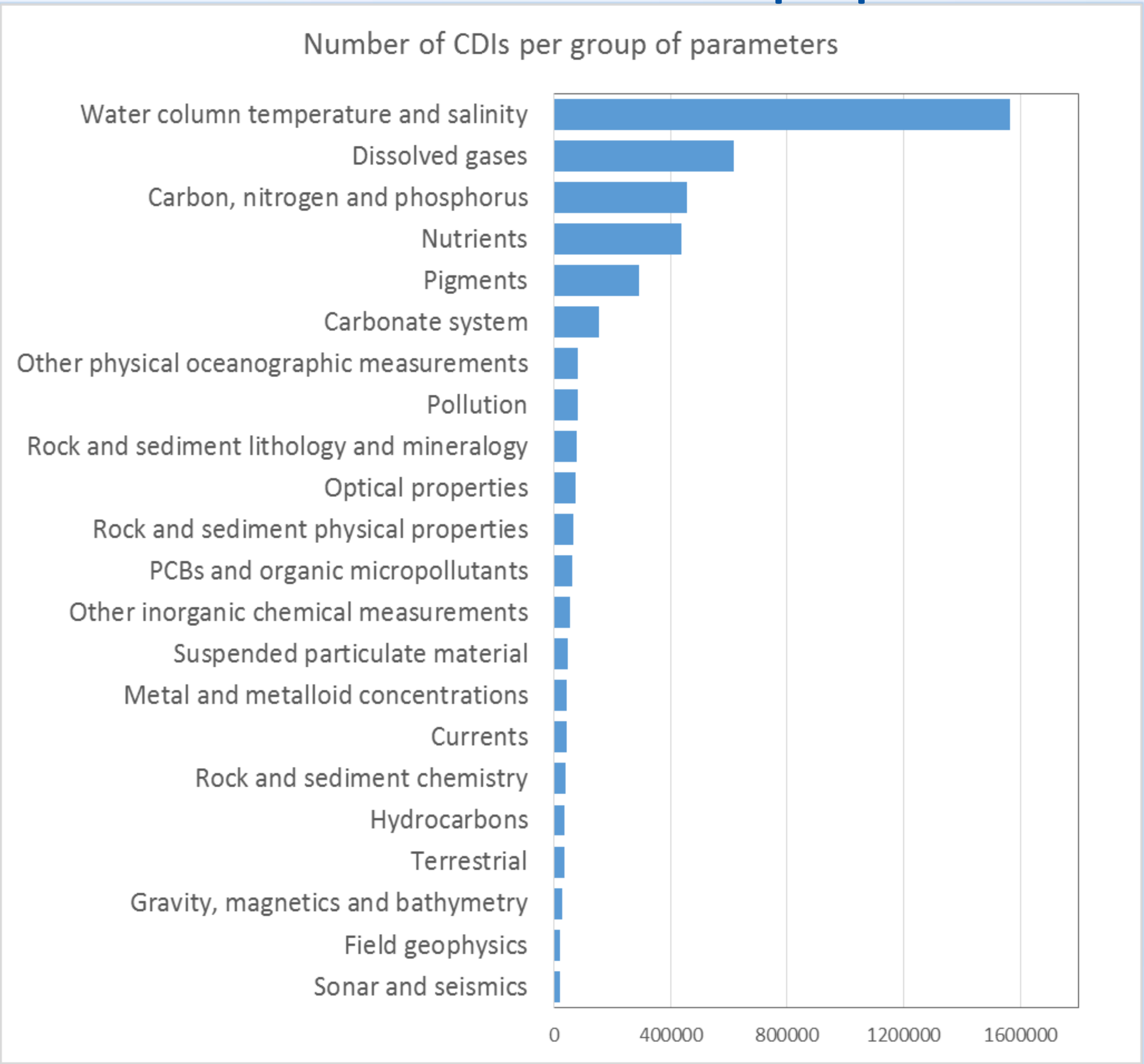


SeaDataCloud

CDI data population

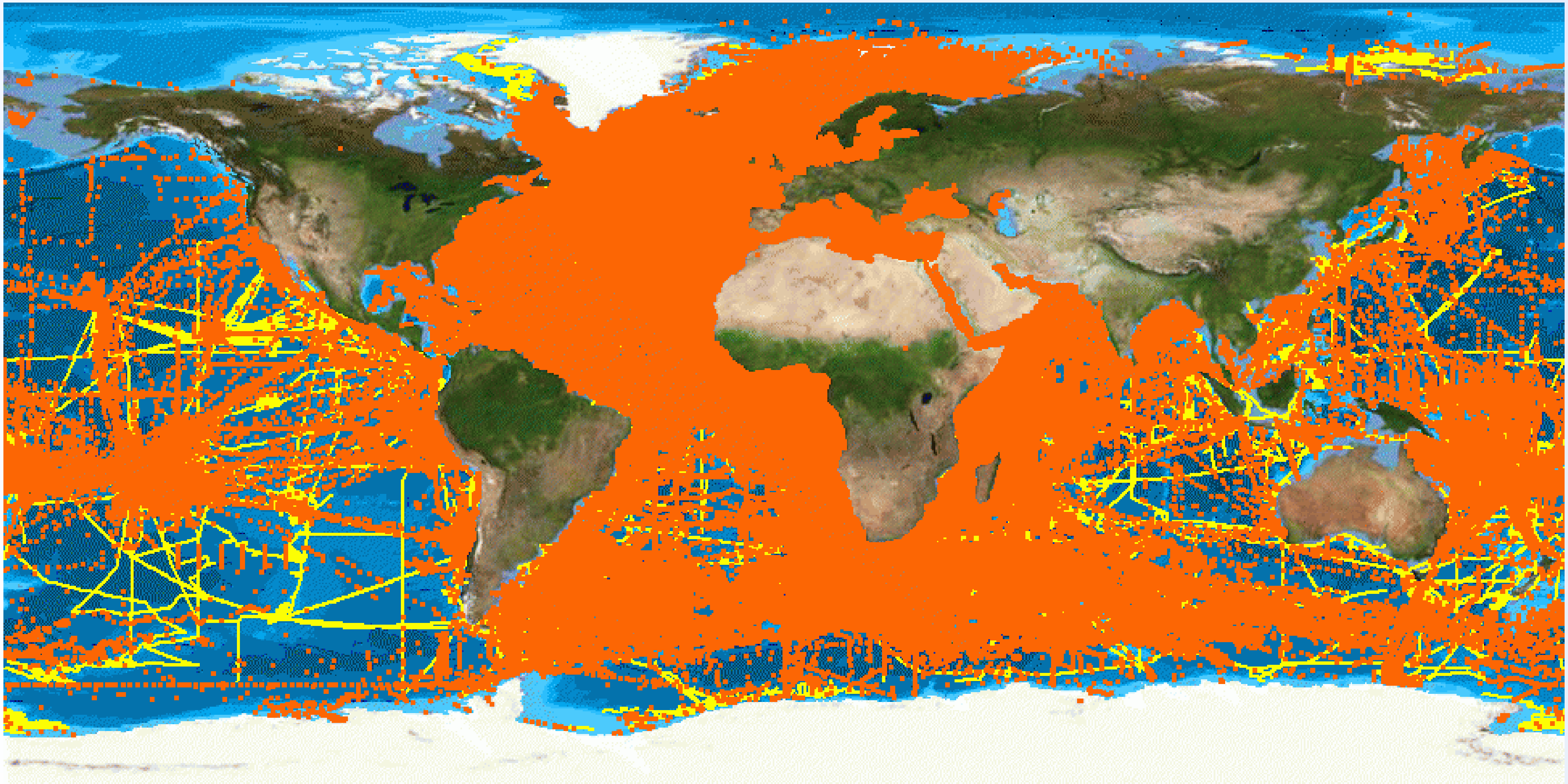


CDI data population



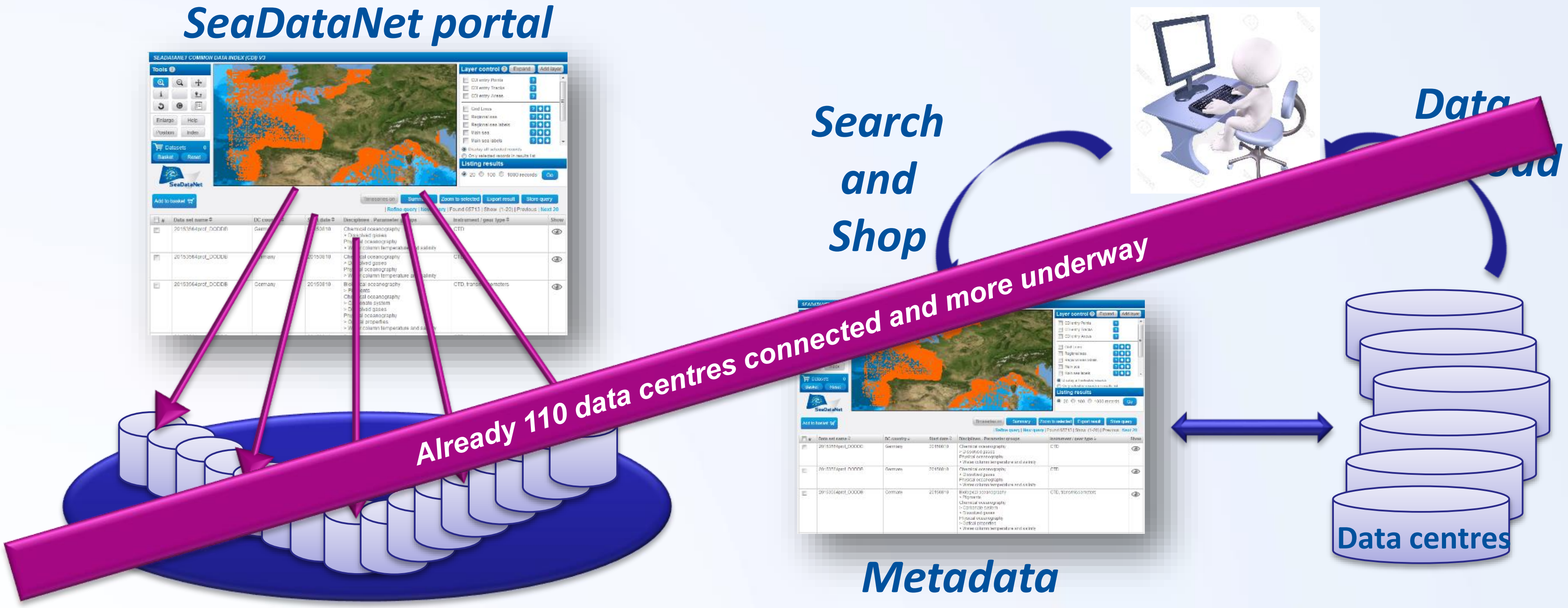
SeaDataCloud

CDI data population



- **2 million** CDI entries from **34** countries, **110** data centres and > **650** originators
- physics, chemistry, geology, geophysics, bathymetry and biology;
- from **1805 to 2017**; **86%** unrestricted or under SDN License

discovery and unified data access



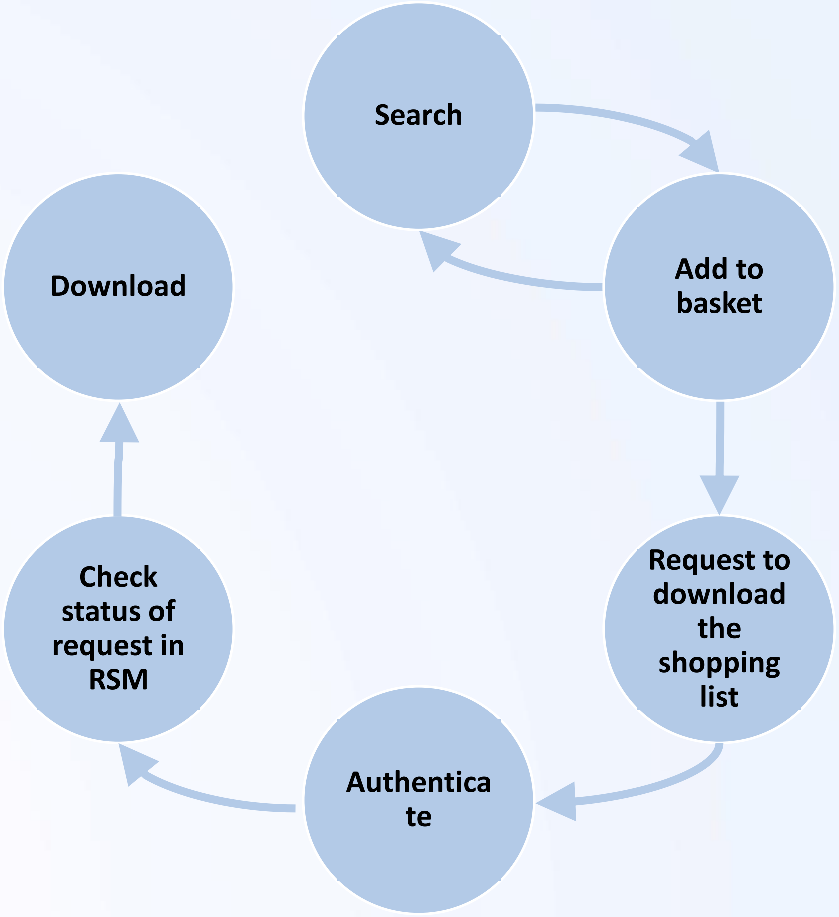
Metadata

+ transaction data

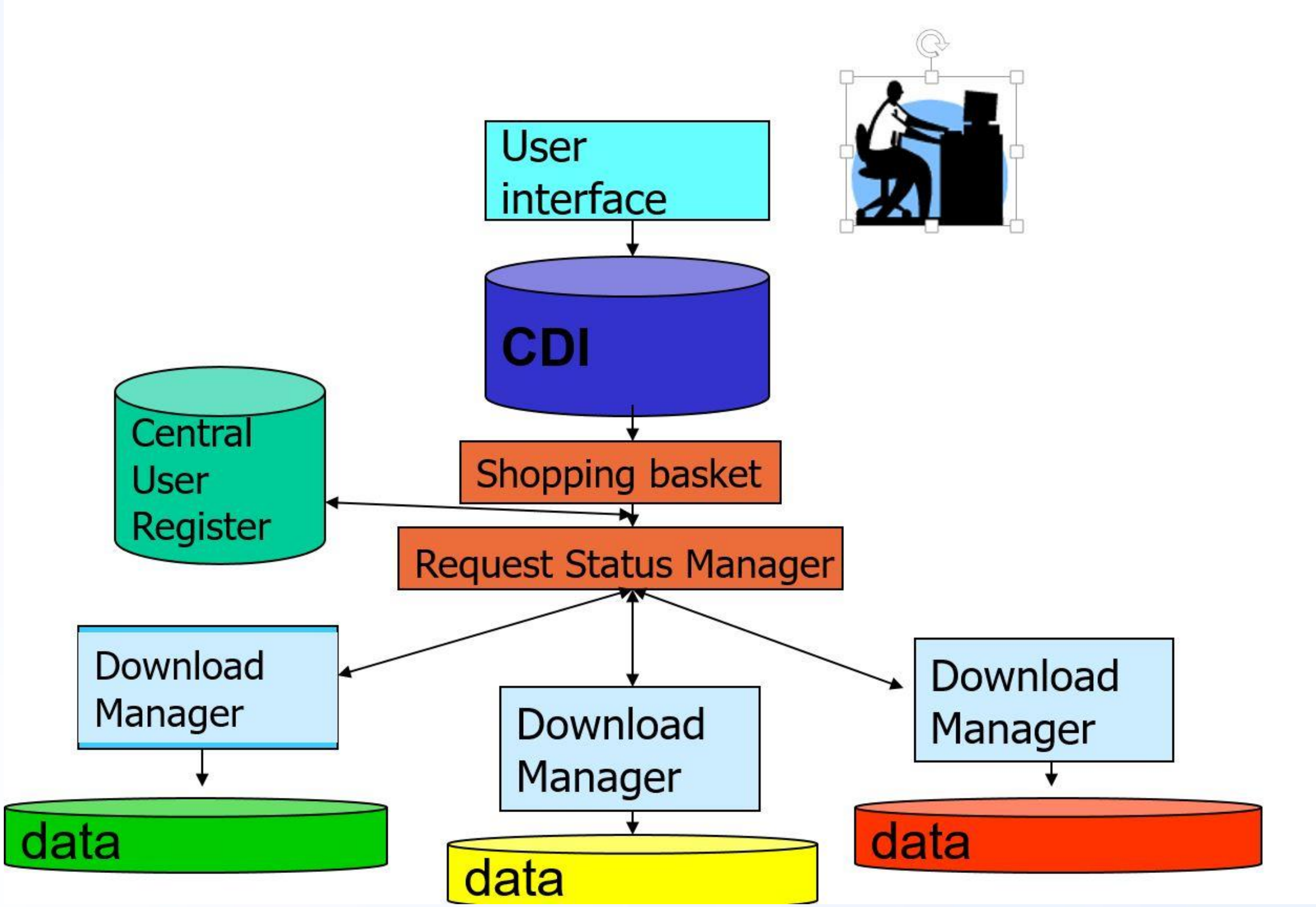
European data sources

data centres ← ≈ 650 originators

User search and download workflow


Search
Add to basket
Request to download the shopping list
Authenticate
Check status of request in RSM
Download

Current CDI service architecture



Issues with current CDI service

- **performance for users**: CDI data access service interacts with the distributed data collections and data bases at the connected data centres.
  - user can submit a shopping basket with requests for data from multiple data centres.
  - user must await the automatic data preparation by each of these data centres
  - user must download resulting data sets through the RSM as packages directly from each data centre, which implicates multiple download transactions
- **performance for users**: data centres are not always online, operational and have different machine capacities which might give extra delays
- **quality issues**: concerning formats of data files (ODV + NetCDF) and their consistency with CDI metadata.
- **installation and configuration** of the Download Manager software can be challenging due to different configurations, firewalls etc., which in practice results in having different versions installed

Upgrading CDI service using the cloud

- Configure and maintain a cloud environment as a ‘cache’ to host copies of all data resources (from the distributed data centres)
- Exchange by dynamic replication from the individual data centres, following their updating of the CDI catalogue service
- In the cloud buffer new functions:
  - checking possible duplicates
  - Checking overall quality of formats
  - Checking integrity of data files and metadata relations.
  - Results of checks reported back to data centres for amendments of their submissions

- Develop a Virtual Research Environment (VRE) to facilitate collaborative and individual research by users
- Provide customised services (MySeaDataCloud) to let users have search profile, receive alerts on new available data, ingest and manage their own datasets

Upgrading CDI service using the cloud

SeaDataCloud

New modern interfaces

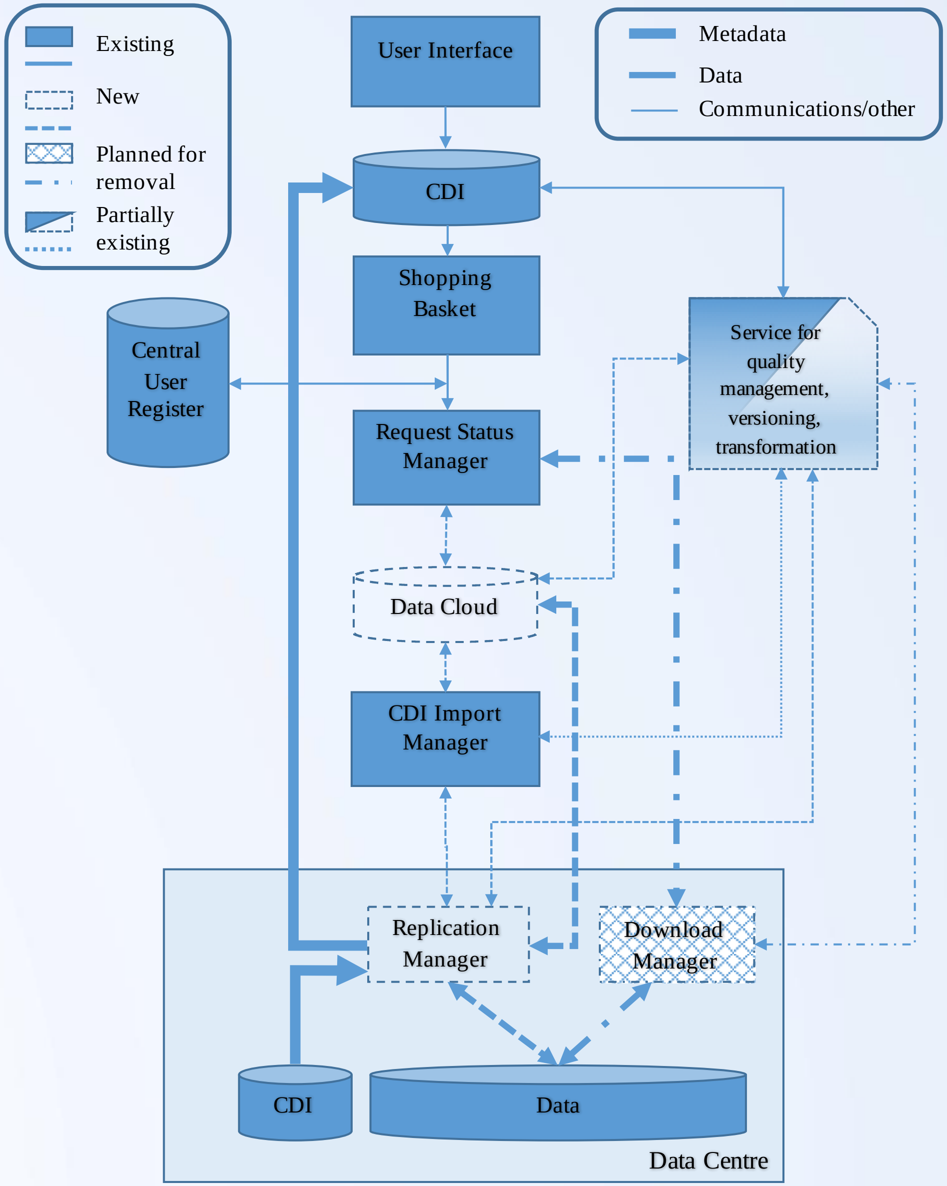
Transformation services

Buffer quality control

Data caching

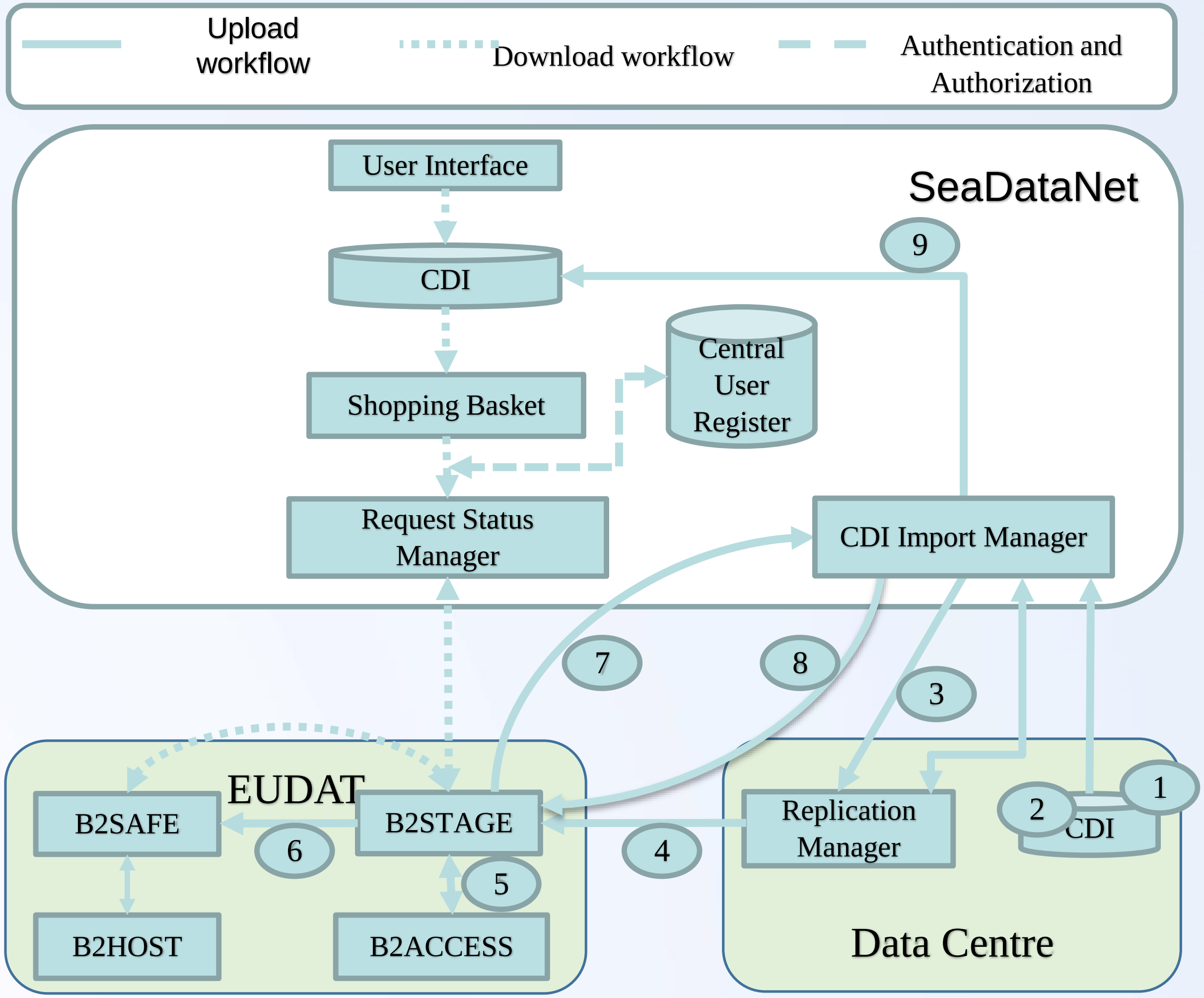
Import manager

Replication Managers at Data nodes



sdn-userdesk@seadatanet.org – www.seadatanet.org

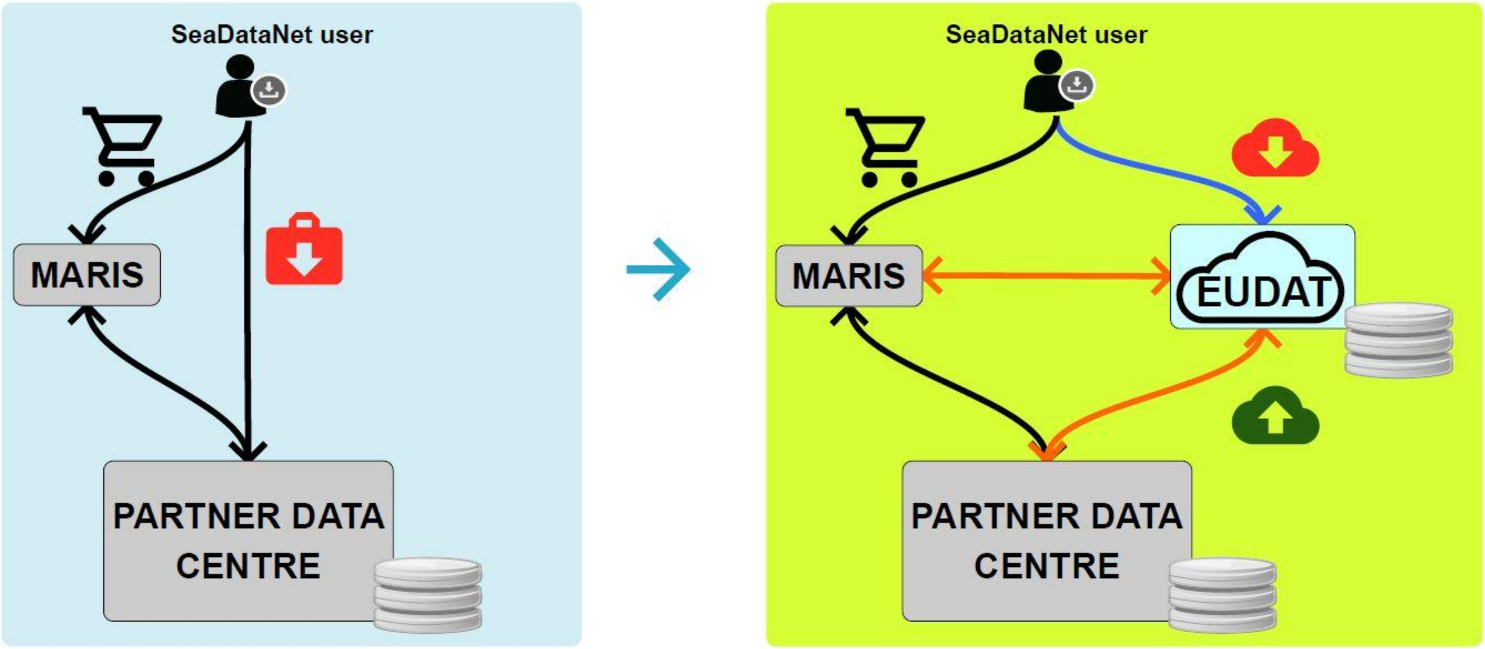
New service components - workflow



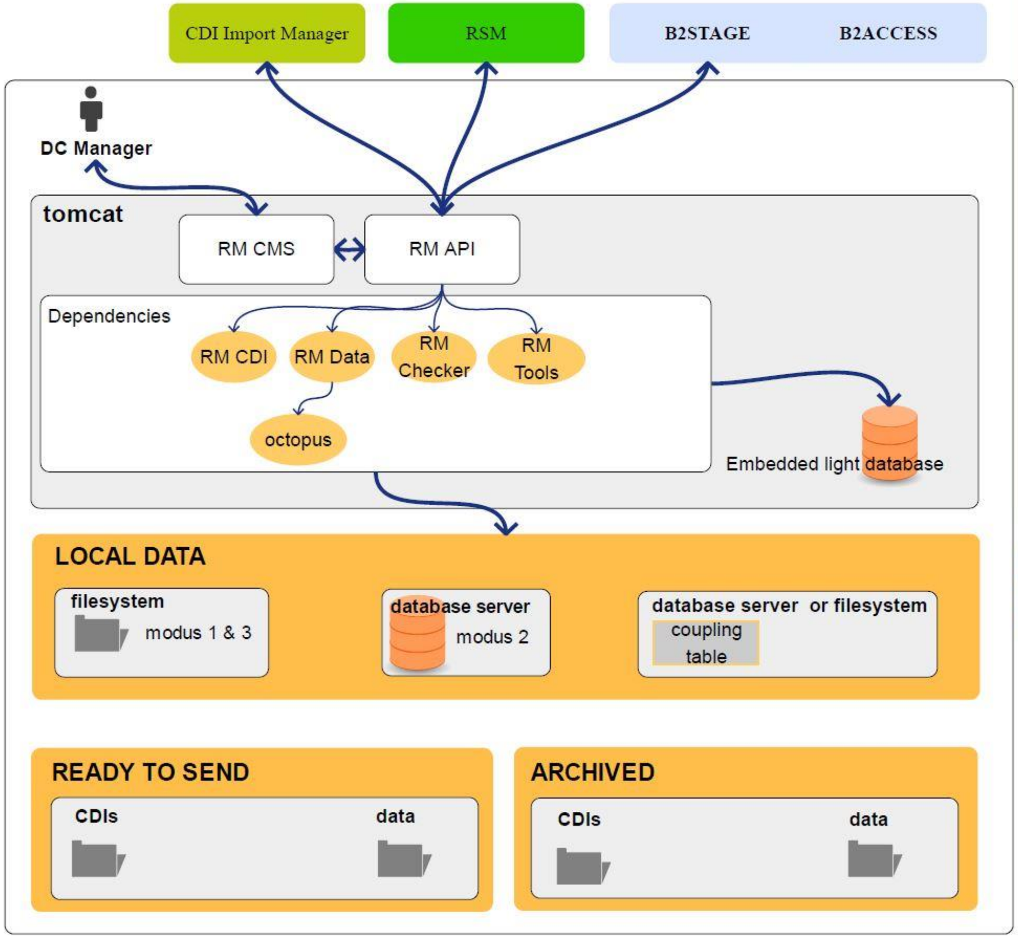
Replication Manager

3 parts instead of 2: **EUDAT** is a new element in the workflow

MARIS + DATA CENTRE → MARIS + DATA CENTRE + **EUDAT**



Architecture



old DM -> new RM:

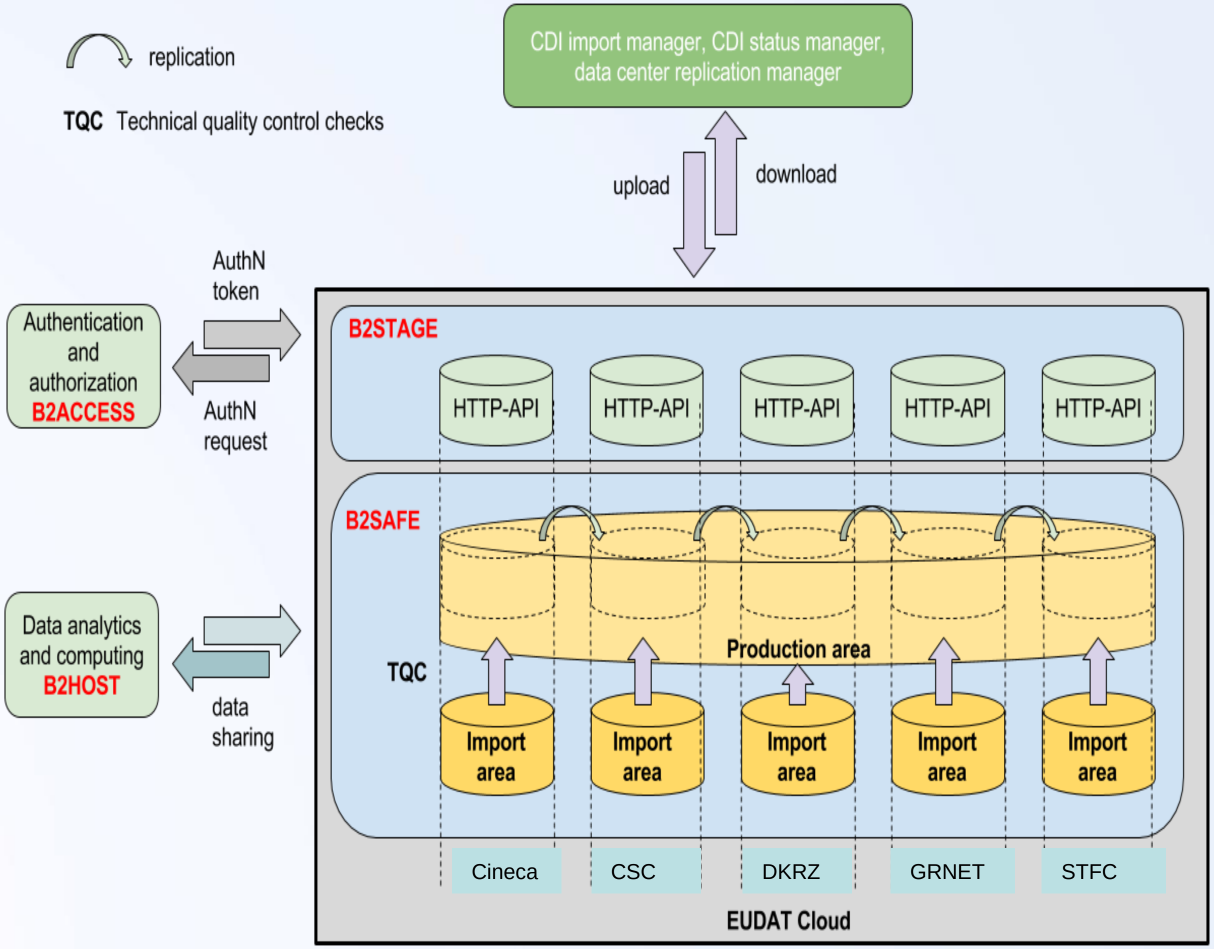
- DM_Servlet (tomcat app)
  -> RM API (tomcat REST service)
  -> RM CMS
- DM Batch -> RM Data
- DM Checker -> RM Checker
- DM ToolsBatch -> RM Tools
  (cleaner is deleted, update vocabs stays)
- -> RM CDI

These are not batches any more, but java libraries.

Modus 2 database is optionnal.
Data centres with modus 2 database can use the same server,
even same database, for modus2 and coupling,
but this is not mandatory.

EUDAT Cloud

- Import area distributed across the five EUDAT partners
- Production area replicated across the five EUDAT partners



Benefits for CDI service and its users

- Cloud buffer in combination with the CDI service will
  - speed up the performance,
  - expand discovery and ease of use of the data access and downloading
  - provide users with one integrated download package instead of multiple packages from multiple data centres.
- Overall quality and coherence (data – metadata) will improve
- Data replication will be triggered per data centre by CDI updates. The replication module might have less complexity than the present Download Manager module
- A system of versioning will be introduced which is required in the context of the MSFD for facilitating repeated analysis of environmental assessments after many years, and for scientific papers.