



# SeaDataCloud

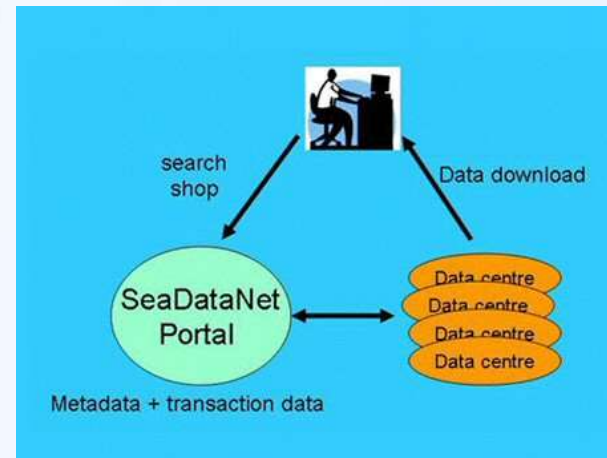
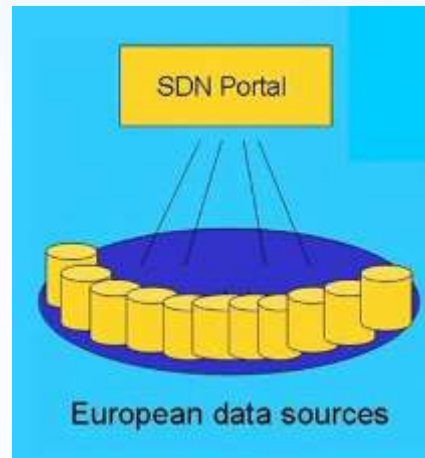
WP9.1: Technical specification of CDI  
service upgrading

Dick M.A. Schaap – Technical Coordinator

SeaDataCloud Plenary Meeting, Athens - Greece, October 2017)  
[sdn-userdesk@seadatanet.org](mailto:sdn-userdesk@seadatanet.org) – [www.seadatanet.org](http://www.seadatanet.org)

# Current CDI Data Discovery and Access service

- one of the core services of the SeaDataNet portal
- providing 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)
- supported by Controlled Vocabularies, and Directories (EDMO, EDMERP, CSR, EDMED)



# Current CDI Data Discovery and Access service



**SEADATANET COMMON DATA INDEX (CDI) V3**

**TOOLS**

**LAYER CONTROL**

**SEARCH**

**Discovery parameters**

**Characteristics**

**Instrument type**

**Platform type**

**Measuring analytic**

**Geographical box**

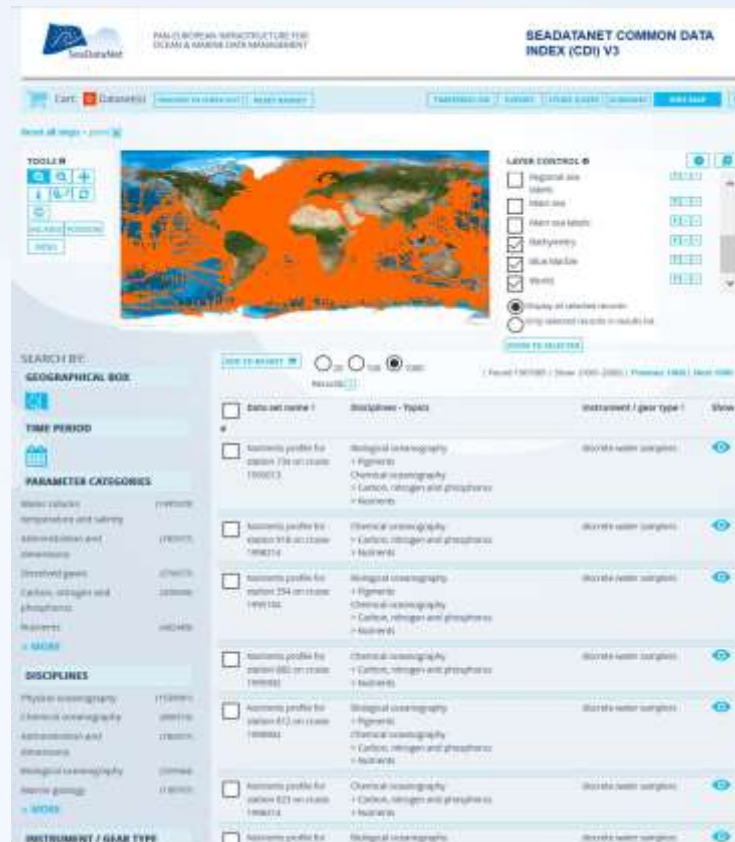
**Time period**

**Parameter categories**

**Dispositions**

**Instrument / gear type**

Extended Search



**SEADATANET COMMON DATA INDEX (CDI) V3**

**TOOLS**

**LAYER CONTROL**

**SEARCH BY:**

**GEOGRAPHICAL BOX**

**TIME PERIOD**

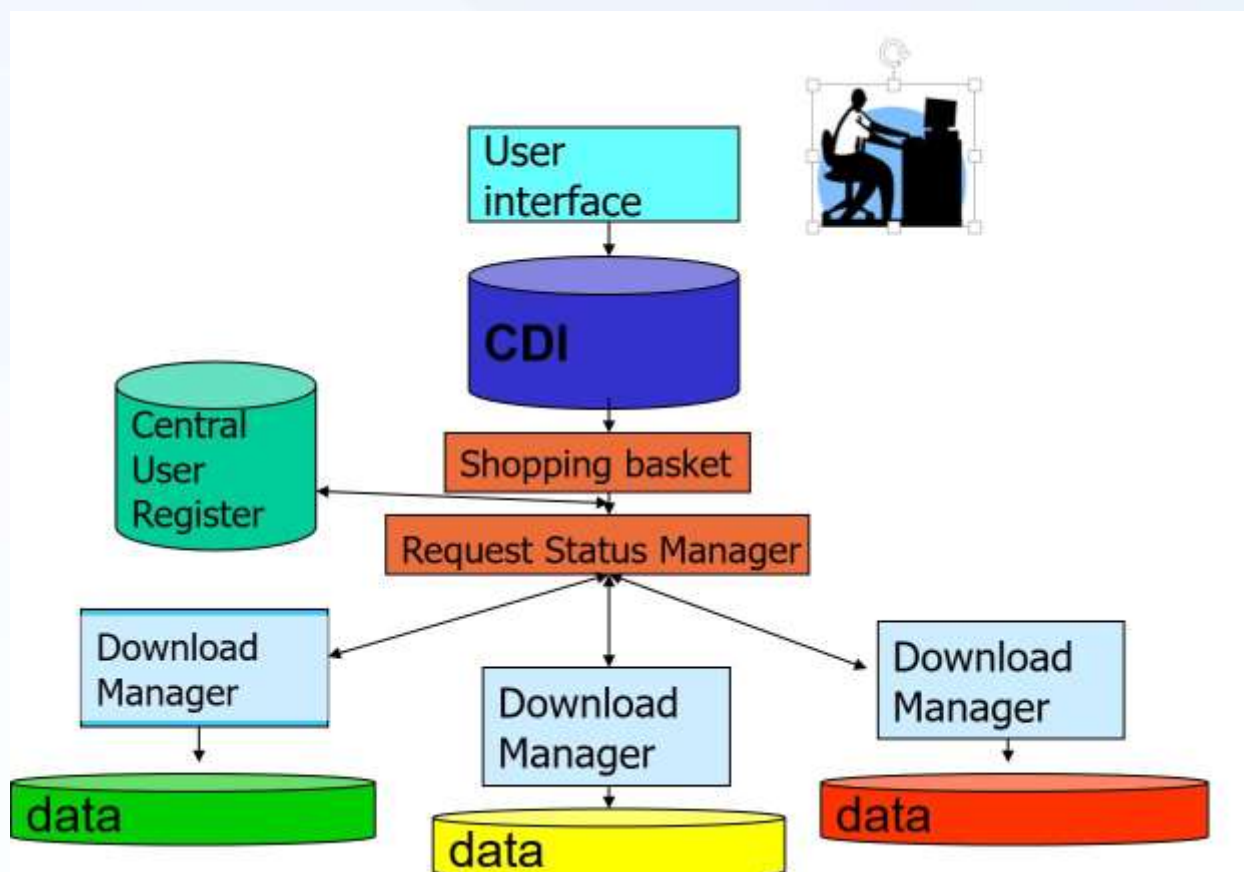
**PARAMETER CATEGORIES**

**DISPOSITIONS**

**INSTRUMENT / GEAR TYPE**

Quick (facet) Search

# Current CDI service architecture



# User search and download workflow



## 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 the CDI Data Discovery and Access service by making use of the cloud and cloud computing

- WP9.1 concerns the following sub-tasks:
  - WP9.1.1: Specification of the SeaDataNet European cloud environment (M6)
  - WP9.1.2: Enhancing the EUDAT B2SAFE replication software (M12)
  - WP9.1.3: Integrating and adapting the EUDAT B2HOST service (M12)
  - WP9.1.4: Configuring the upgraded CDI service (M18)
  - WP9.1.5: Deploying, testing and taking into operation the upgraded CDI service (M18)
- WP9.2 concerns developing and deploying additional services in the cloud for ensuring integrity and conformity of the CDI related cloud data resources (M18)
- upgraded CDI service fully deployed – M24 (*via* WP5 and WP3)



# Adopting cloud and HPC facilities - requirements

- Copies of all unrestricted and SeaDataNet-licensed data should be replicated from data centres to the SeaDataNet data cache in an efficient way.
- To ensure the quality of the delivered data sets for users, a number of technical quality checks should be applied before making the data available.
- Results of checks should be reported back to SeaDataNet data centres for possible amendments



# Adopting cloud and HPC facilities - requirements

- Enabling versioning of datasets
- Facilitating transformation services to convert incoming SeaDataNet ODV files to SeaDataNet NetCDF files, support INSPIRE and to enrich data files with their CDI metadata for delivery to users.
- Authentication and authorization of users' access need to be in place and compatible with the currently used Marine-ID, while for access to EUDAT services (internally) use should be made of B2ACCESS.

# Use Cases for the CDI service

- two main use cases:
  - advanced CDI import
  - advanced data download
- **CDI import process:** Currently each data centre only uploads the metadata records (CDI entries) to the central CDI catalogue. In the new situation the ingestion phase will include additional steps such as upload / replication of data, quality control of data attributes and characteristics, control of duplicates, etc.
- **Data download process:** Currently access requests are forwarded to data centres for serving users by Download Manager + RSM. In the new situation RSM will interact with the EUDAT data cache for delivering data sets to users.

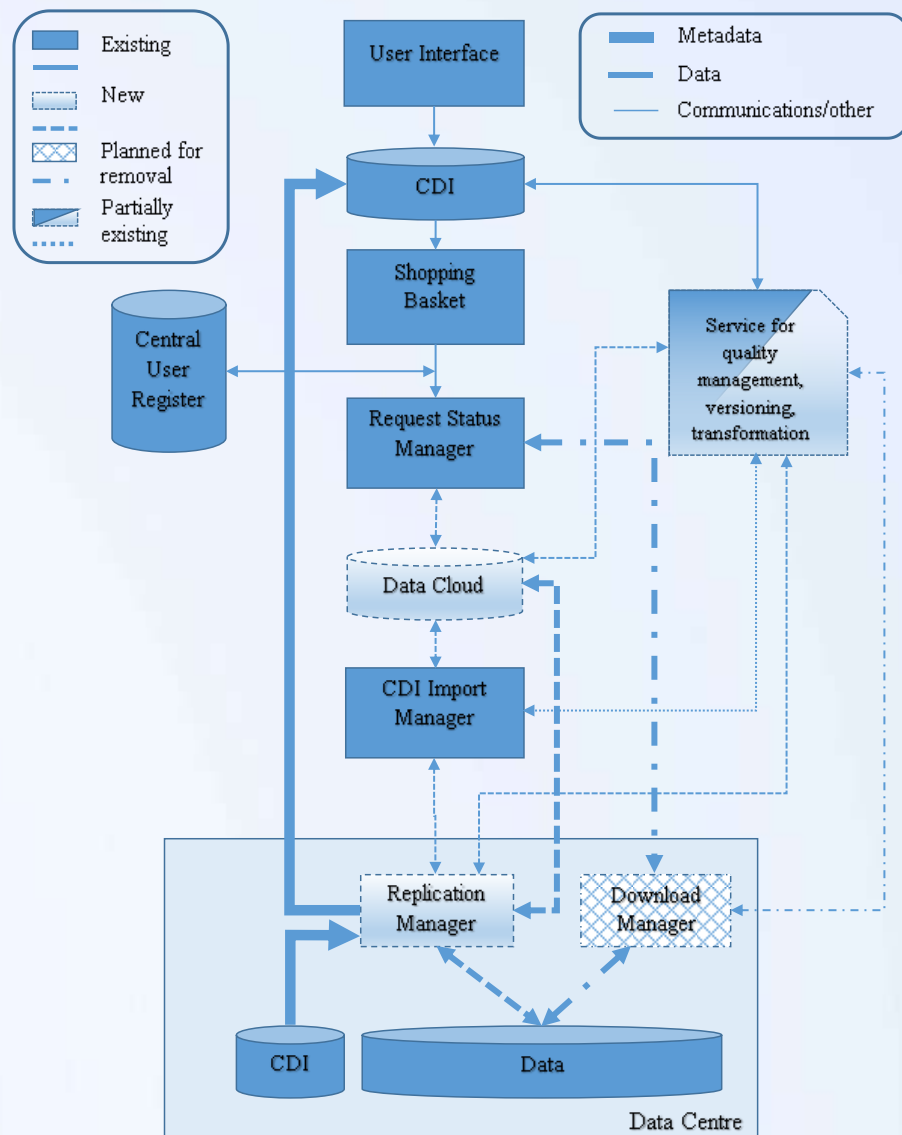
# SeaDataCloud Infrastructure

- Data replication from Data Centres to EUDAT Cloud allows:

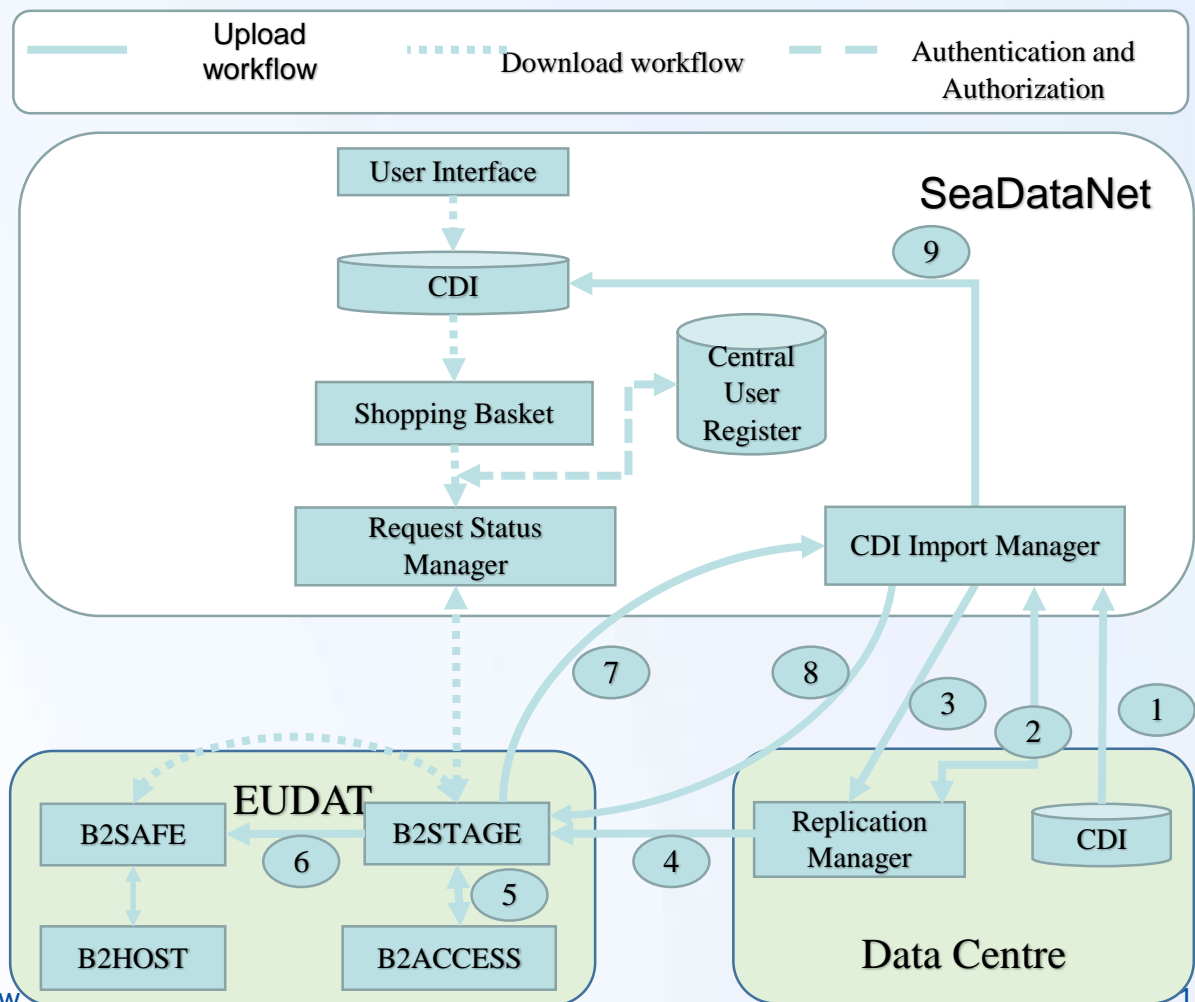
- Unified access (Download)
- Extra QCs
- Transformations
- Versioning
- ...

- But also

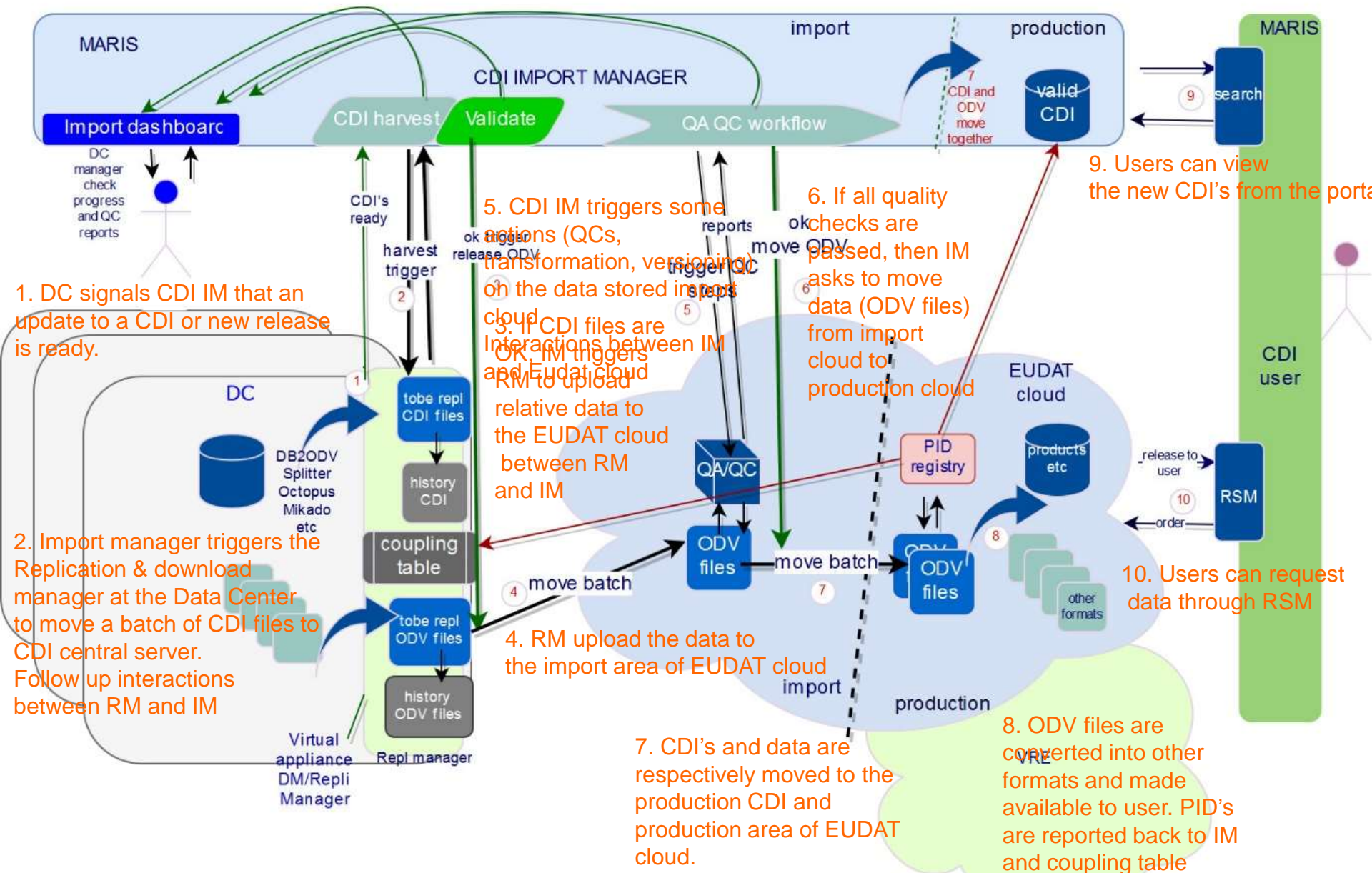
- Digital preservation
- Long term curation



# New CDI service components - workflow



# Upload & Download

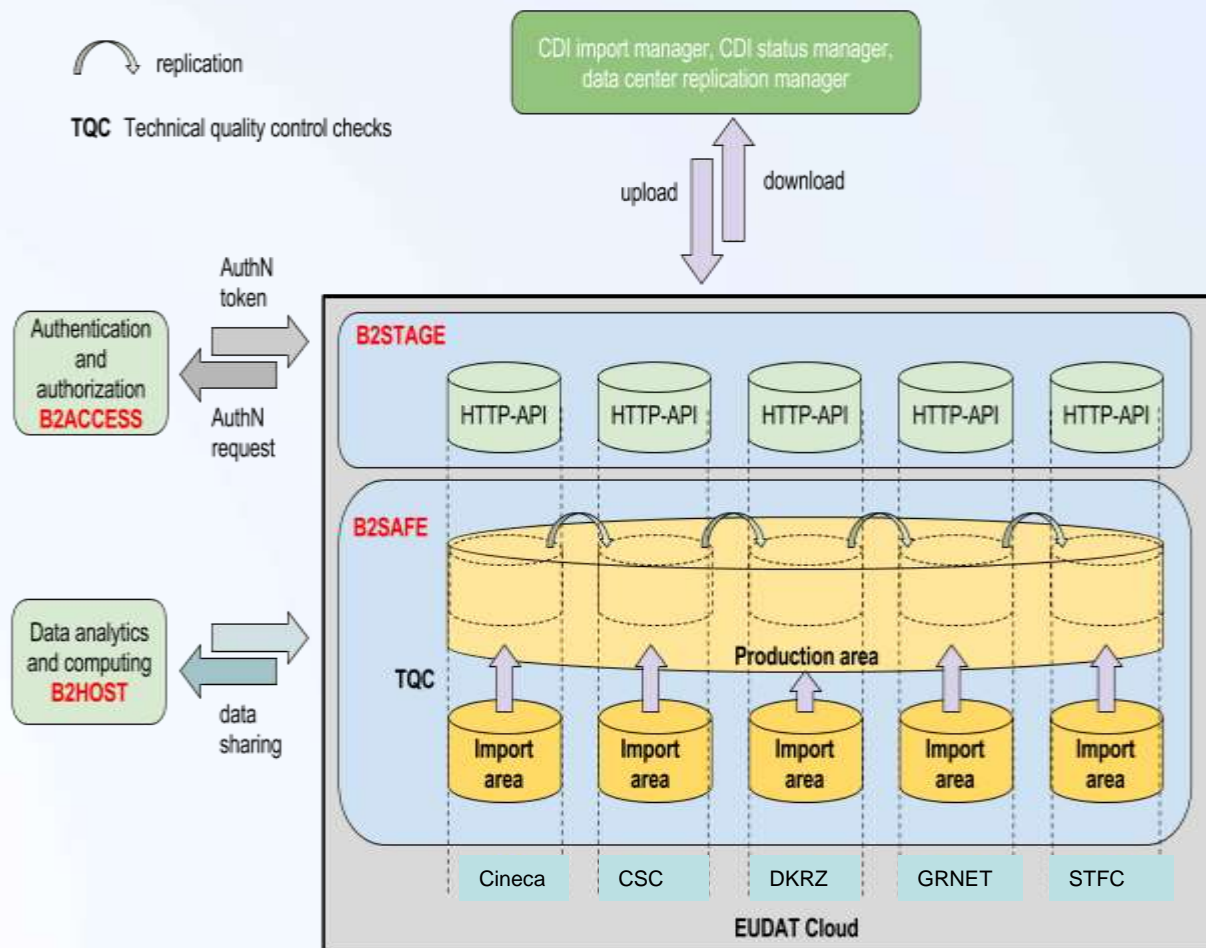






# EUDAT Cloud

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

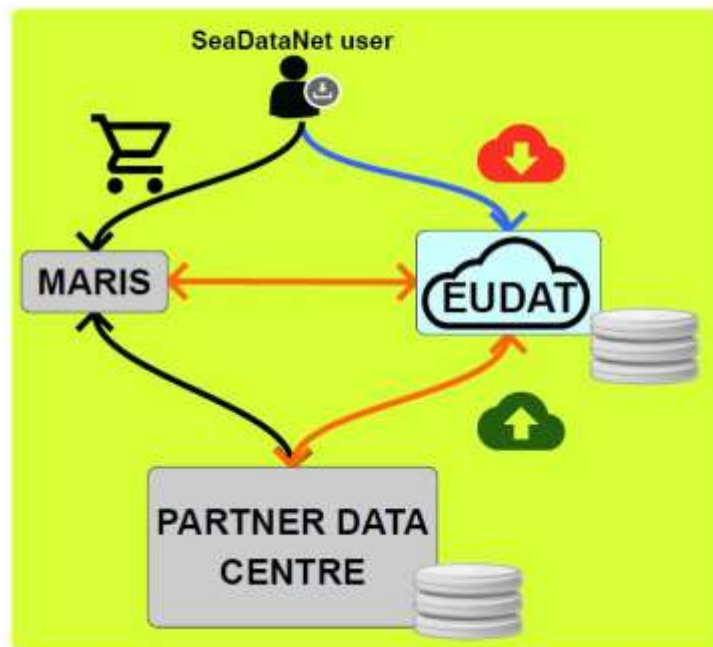
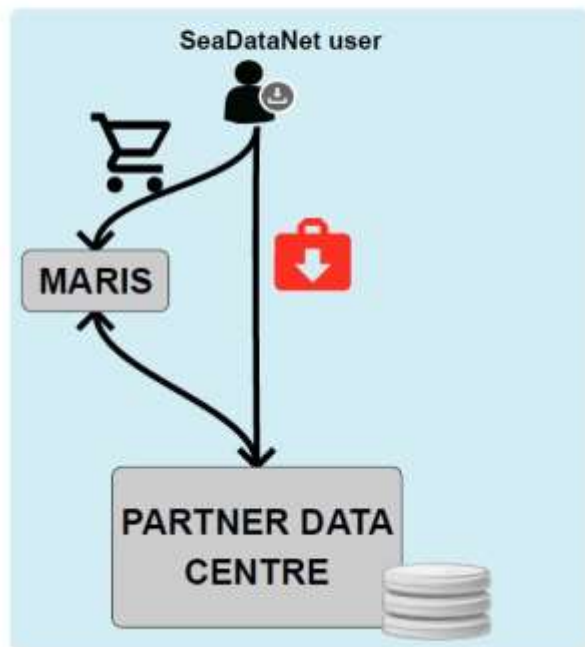




# Replication Manager

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

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

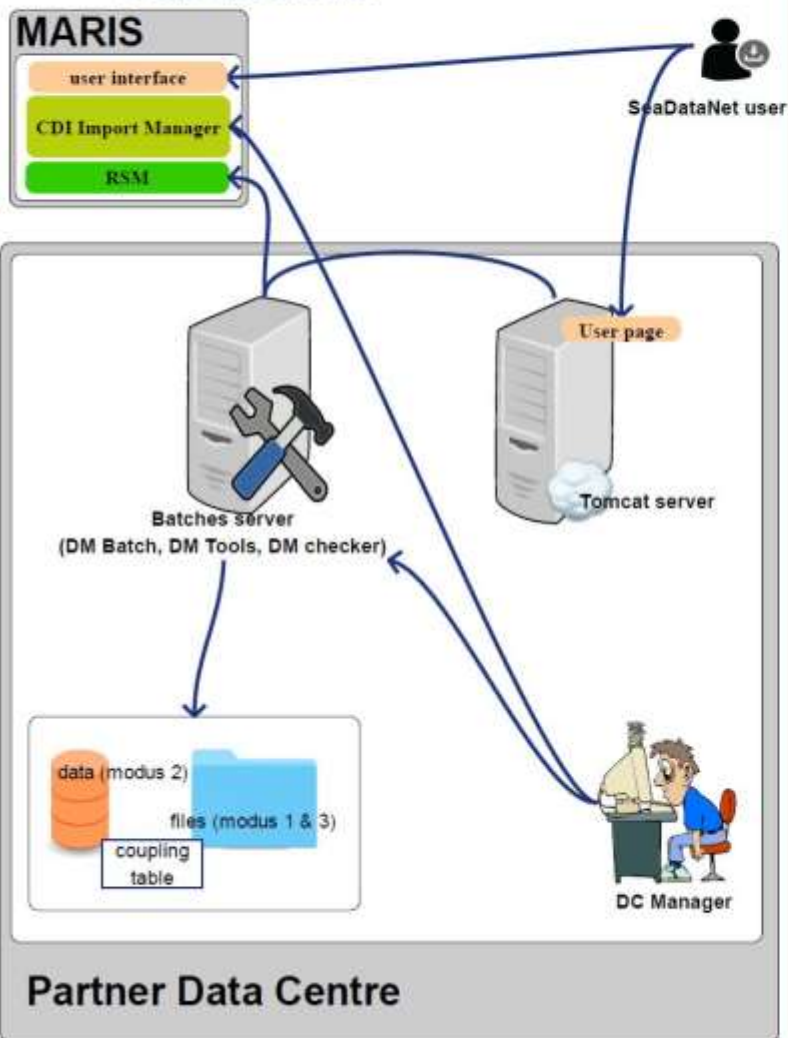


# Replication Manager

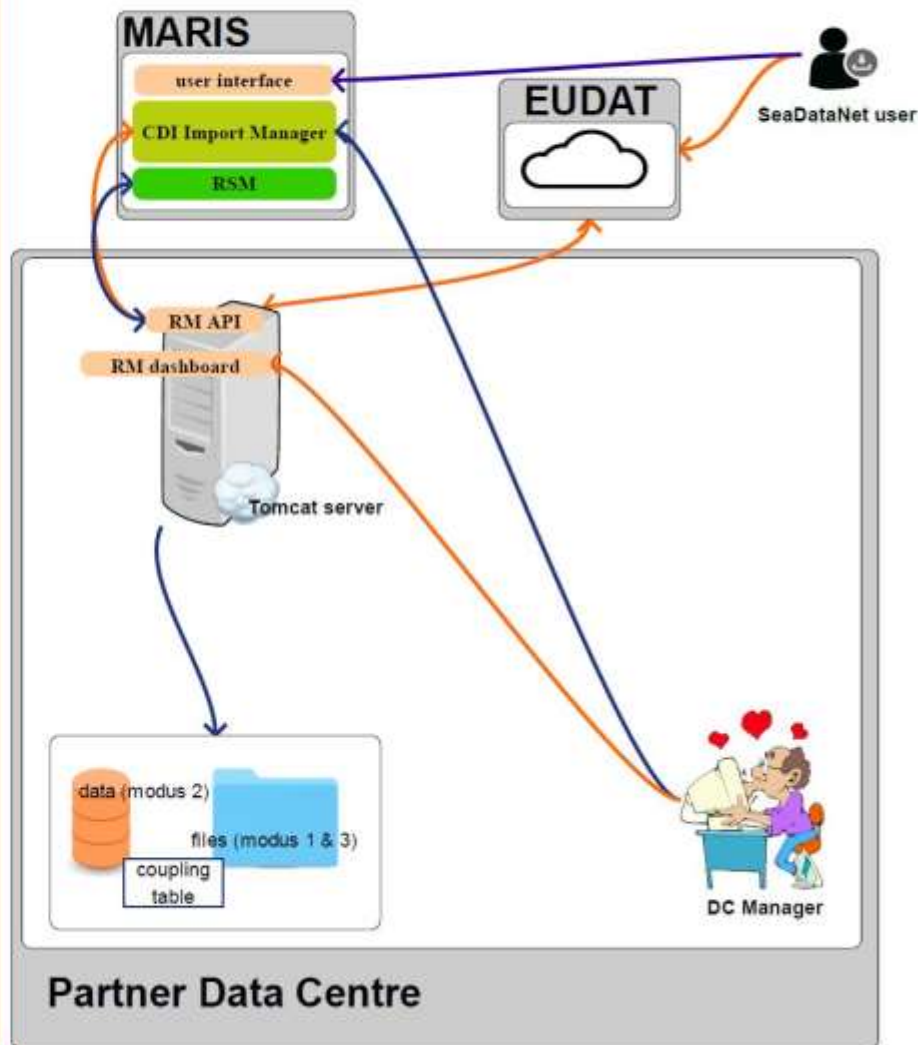
- send CDIs to the Import Manager (Maris)
- listen to RSM (Maris) for user's restricted data requests
  - prepare data
  - upload data to the cloud (temporary storage)
- listen to Import Manager (Maris) for unrestricted data (public and SDN license) import requests
  - prepare data
  - upload data to the cloud (permanent storage)



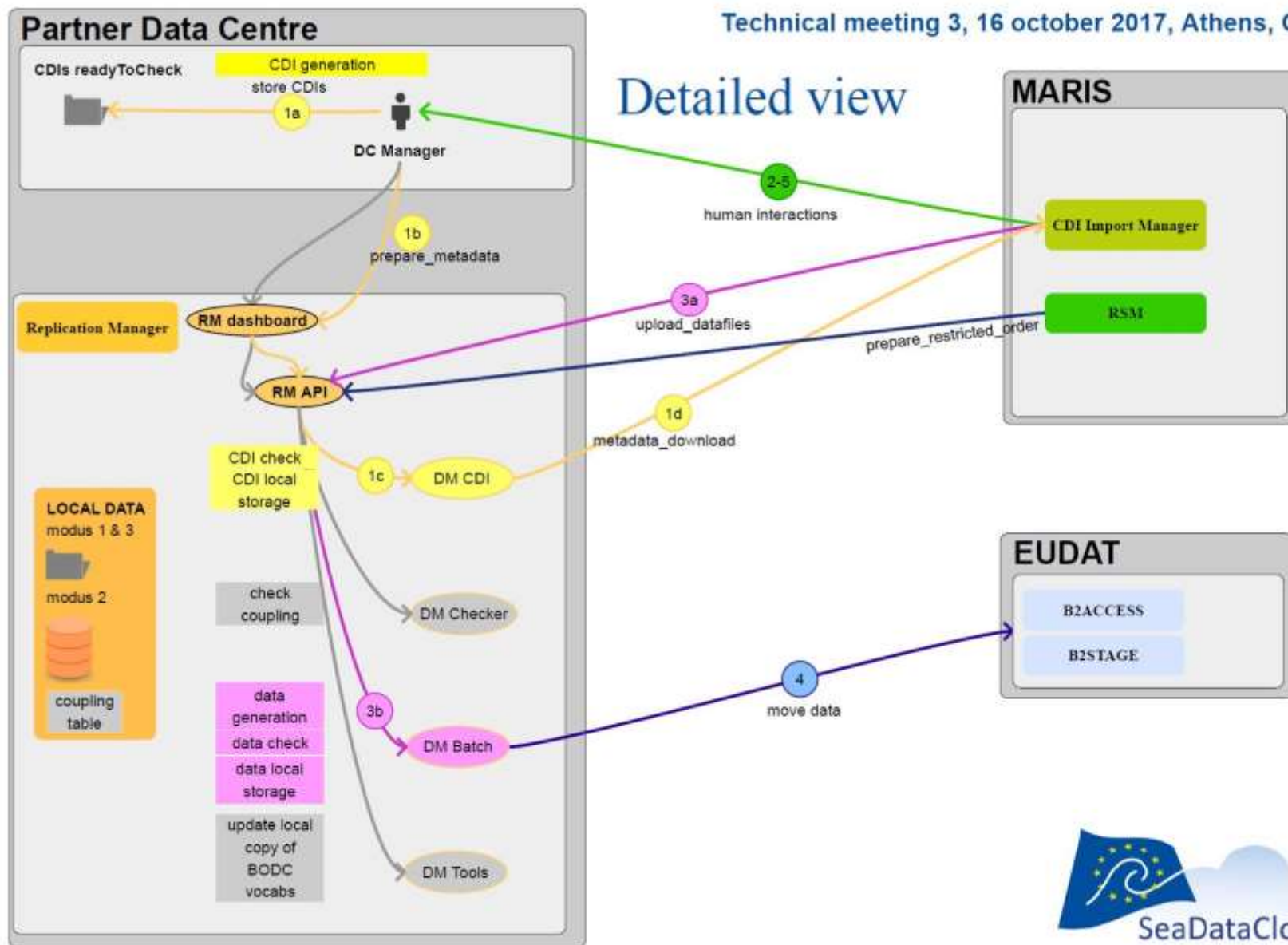
## Download Manager



## Technical meeting 3, 16 october 2017, Athens, Greece Replication Manager

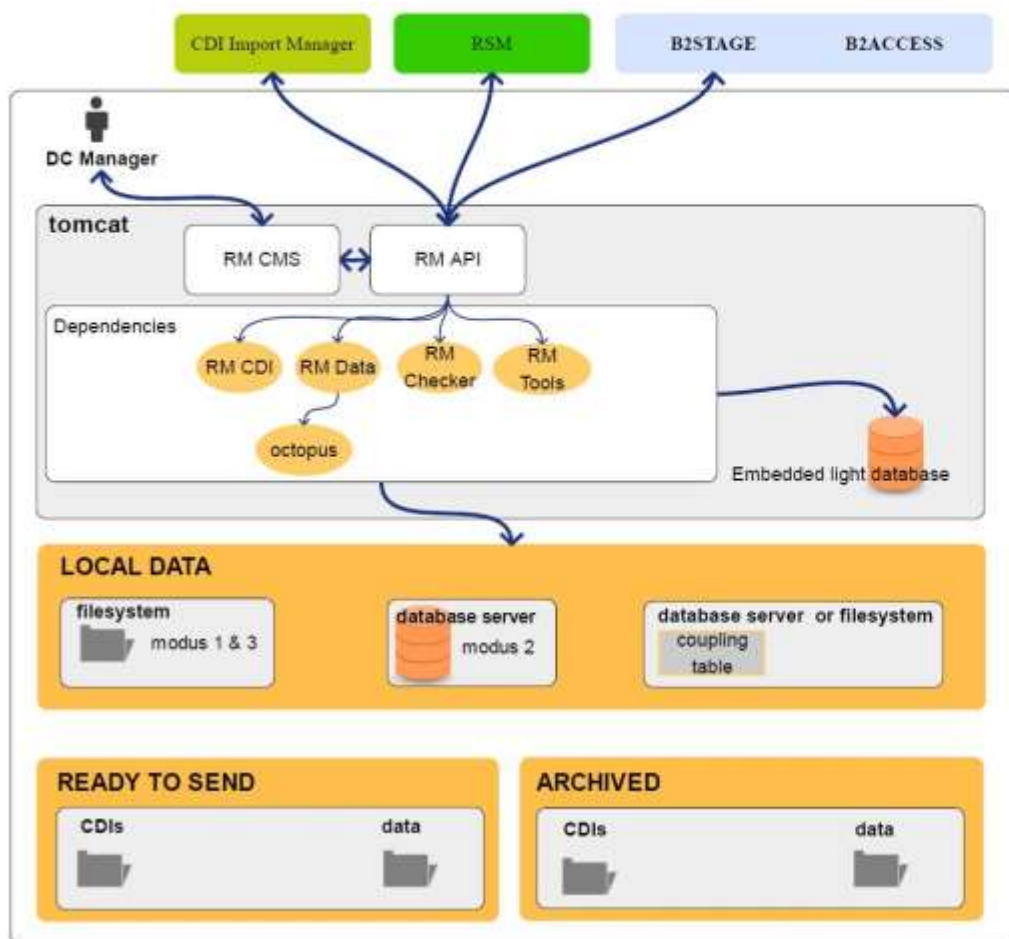


## Detailed view





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

# Upgrading the CDI Data Discovery and Access service

- Deliverable 9.1 published end June 2017: It includes the requirements to build the environment and the specification for upgrading the CDI service infrastructure using EUDAT services.
- Deliverables 9.2 and 9.3 well underway:
  - **D9.2:** Enhanced EUDAT B2SAFE replication software (**M12**)
  - **D9.3:** Adapted EUDAT B2HOST service operational (**M12**)
- Technical documents underway:
  - MARIS: specifying dialogues
  - MARIS: drafting new user interface wireframe
  - IFREMER: specifying Replication Manager

# Development and implementation

- Implementation plan for installing, configuring and taking into operation local components for the upgraded metadata directory will be provided in WP5 at M16 (D5.1) while a pilot group of data centres will be operational with upgraded directory services at M18 (D5.2). All data centres will be operational at M24 (D5.3).
- Development plan:
  - M6-16 Development and Implementation
  - M17-M18 Test of the upgraded CDI service
  - M19 – M24 Uptake