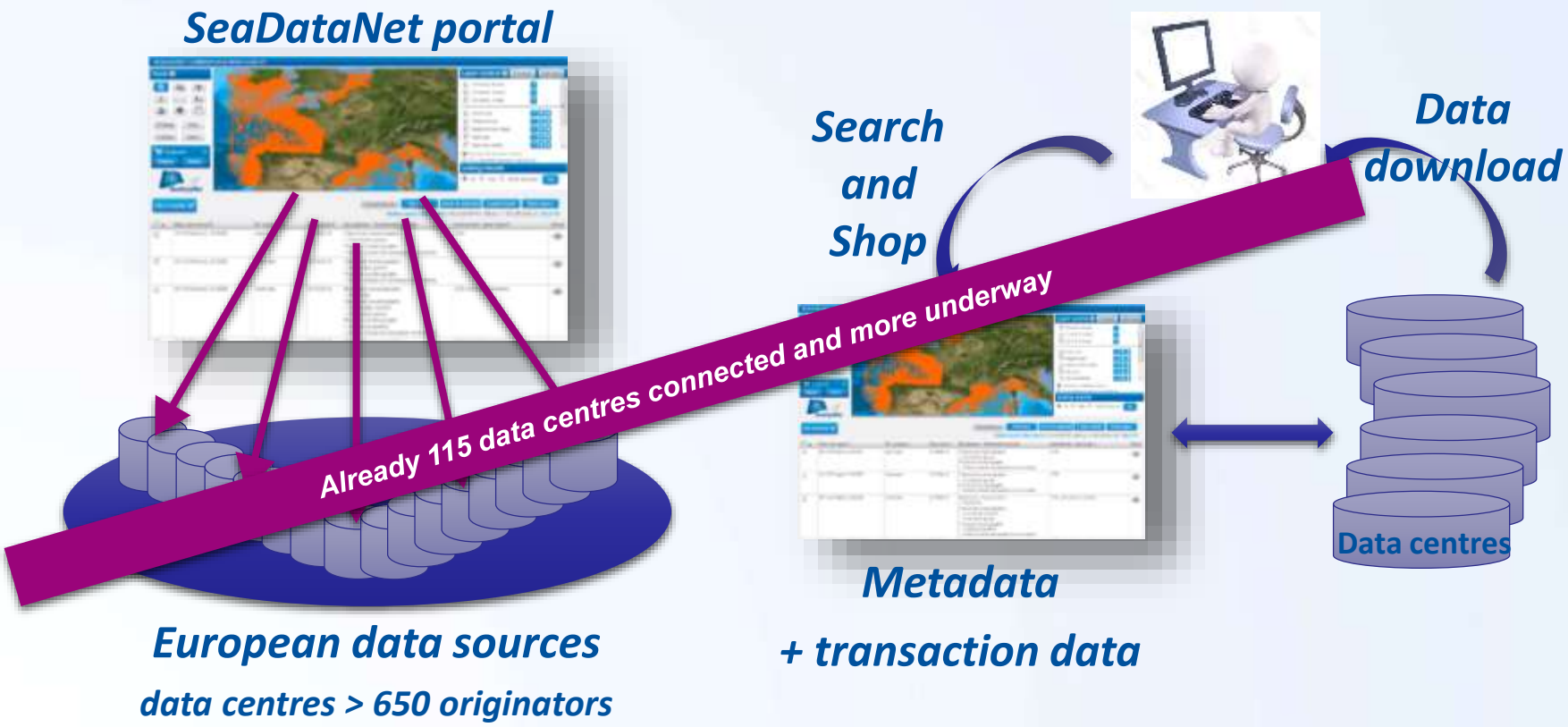
SeaDataCloud

Upgrading the CDI Data Discovery and Access service, adopting the cloud

Dick M.A. Schaap, MARIS, SDC Technical Coordinator

SDC Review, Brussels – Belgium, 6 December 2018
sdn-userdesk@seadatanet.org – www.seadatanet.org

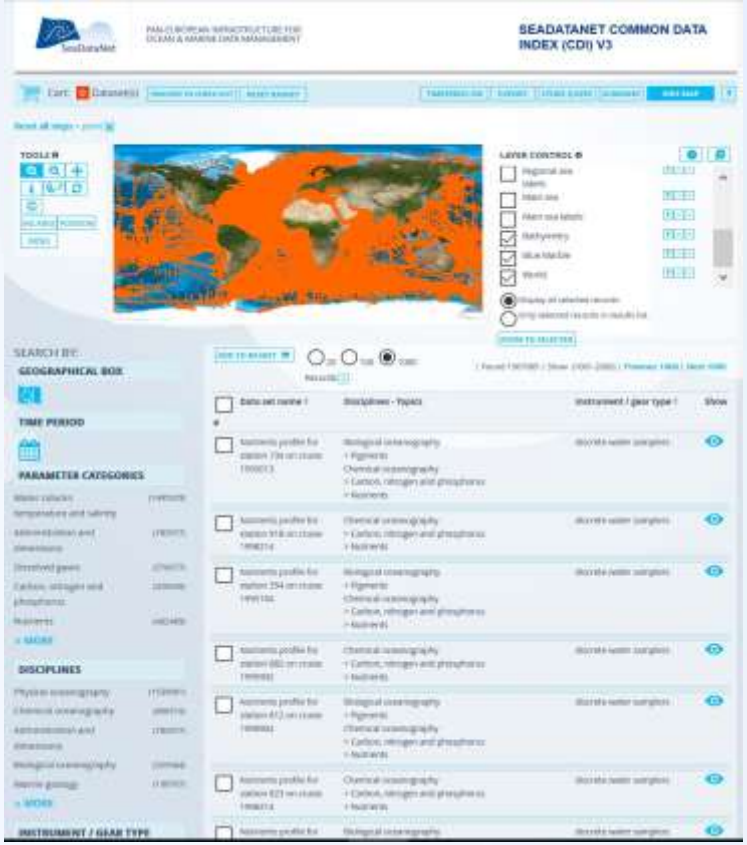
CDI service for discovery and unified data access



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

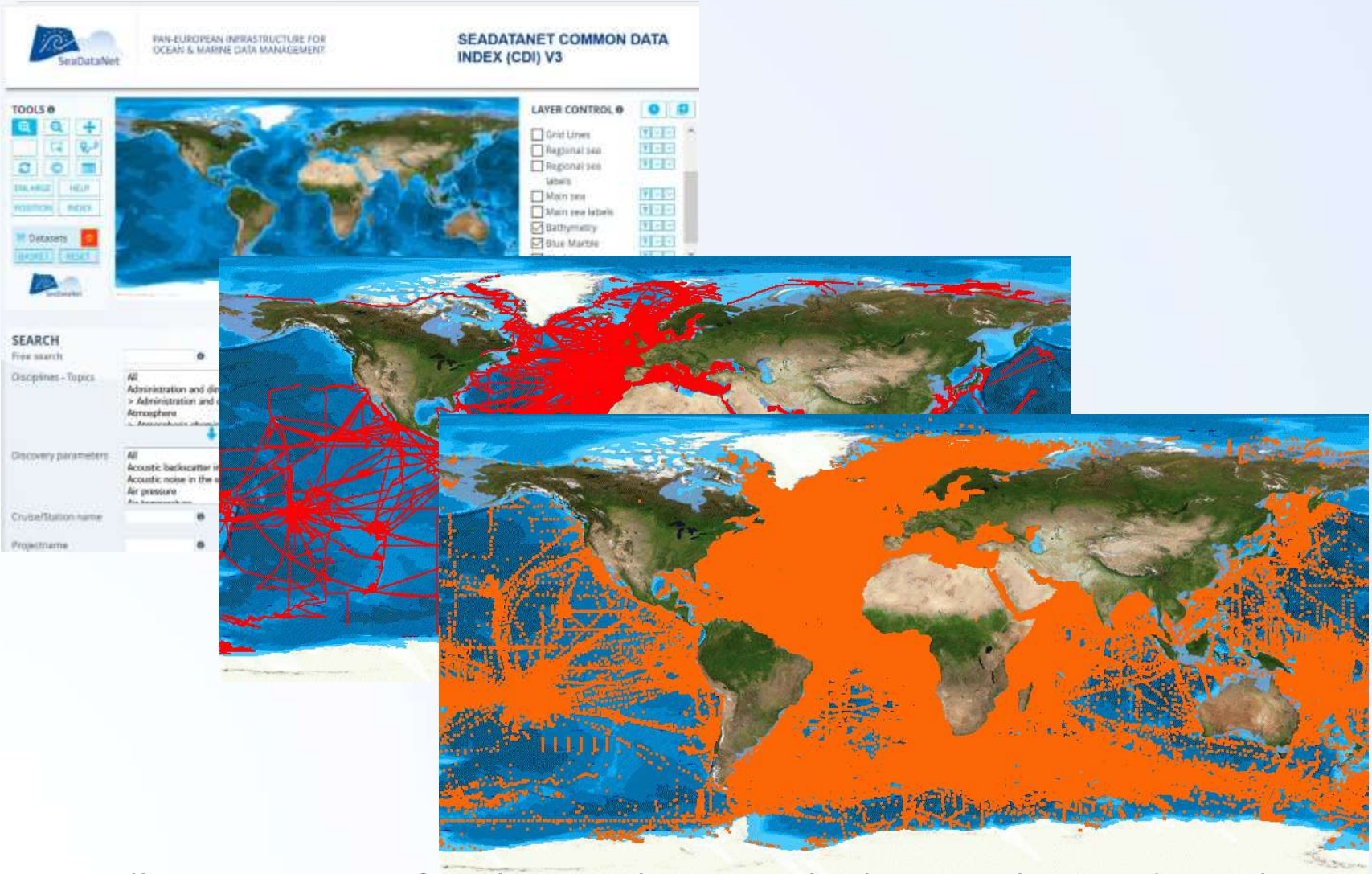
Current CDI user interfaces

Extended Search



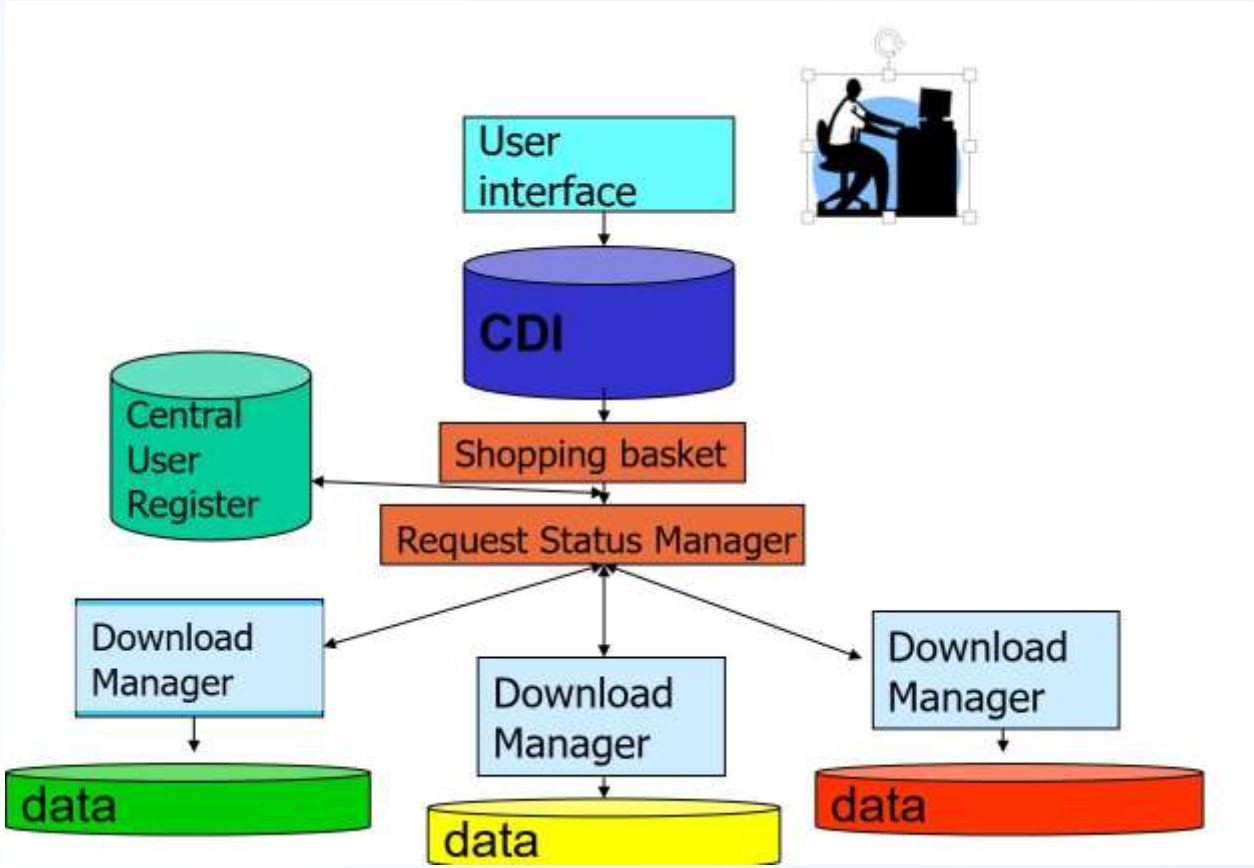
Quick (facet) Search

CDI service with global coverage



> 2.15 Million CDI entries for physics, chemistry, biology, geology and geophysics

Current CDI service architecture



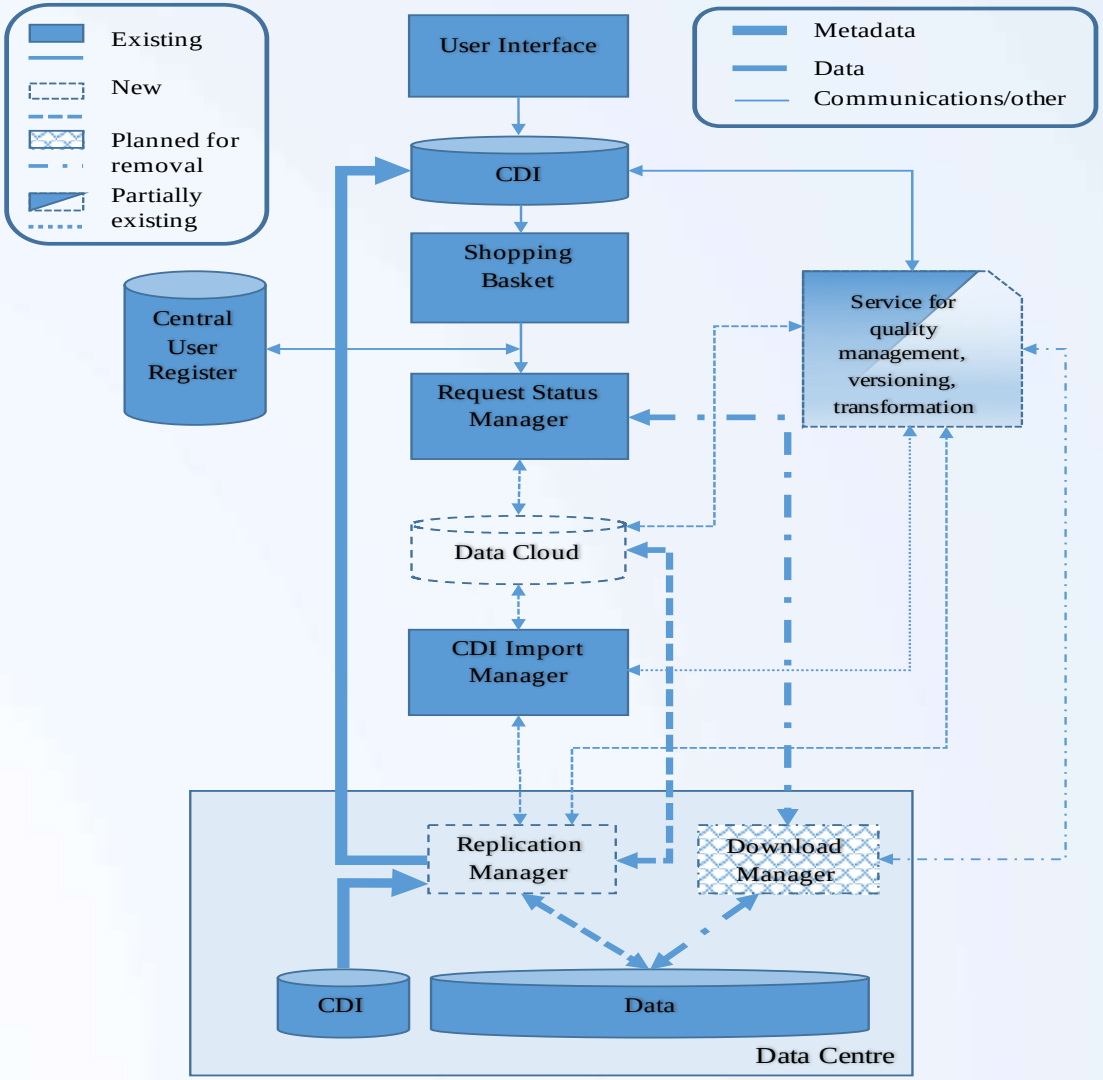
Issues with current CDI service

- **performance for users**: CDI data access service interacts with the distributed data collections and databases 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

Principles for upgrading the CDI service using the cloud

- To configure and maintain a **CLOUD** environment with High Performance Computing (HPC) facilities to host **copies of unrestricted data resources**
- Exchange by dynamic **replication** from the individual data centres, following their updating of the CDI catalogue service
- In the cloud buffer:
  - checking overall quality of metadata and data, as extra check on top of local QA-QC by data centres
  - checking integrity of data files and metadata relations.
  - results of checks to be reported back to data centres for amendments of their submissions and/or local configurations for mapping data and metadata.
- Include transformation services for converting data sets to SeaDataNet ODV and NetCDF formats and relevant INSPIRE data models.
- Introduce versioning of metadata and data as part of provenance

New CDI service architecture



New CDI service components

- **Local software tools** at data centres to prepare ingestions
- **Replication Manager** (RM) at data centres for exchanging to Import Manager and EUDAT cloud
- **Import Manager dashboard** to steer import and validation process
- **EUDAT cloud** with adapted EUDAT services
- Upgraded **CDI User Interface**, ordering and downloading facility

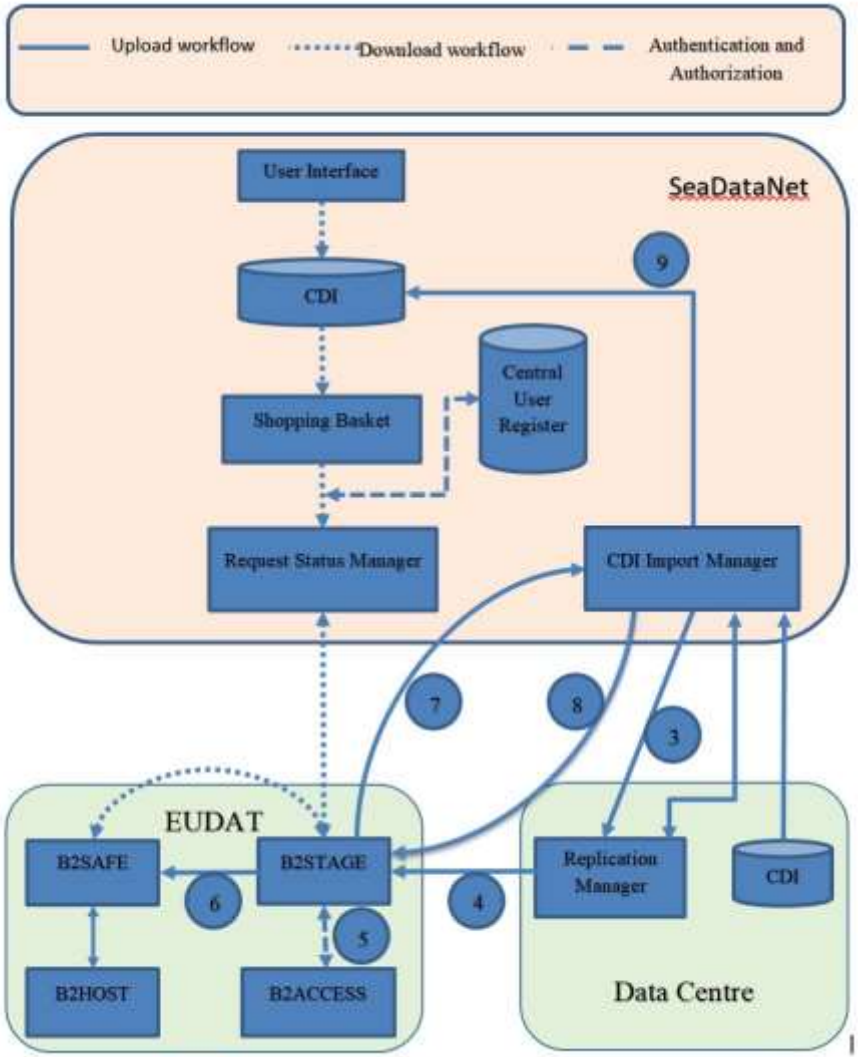
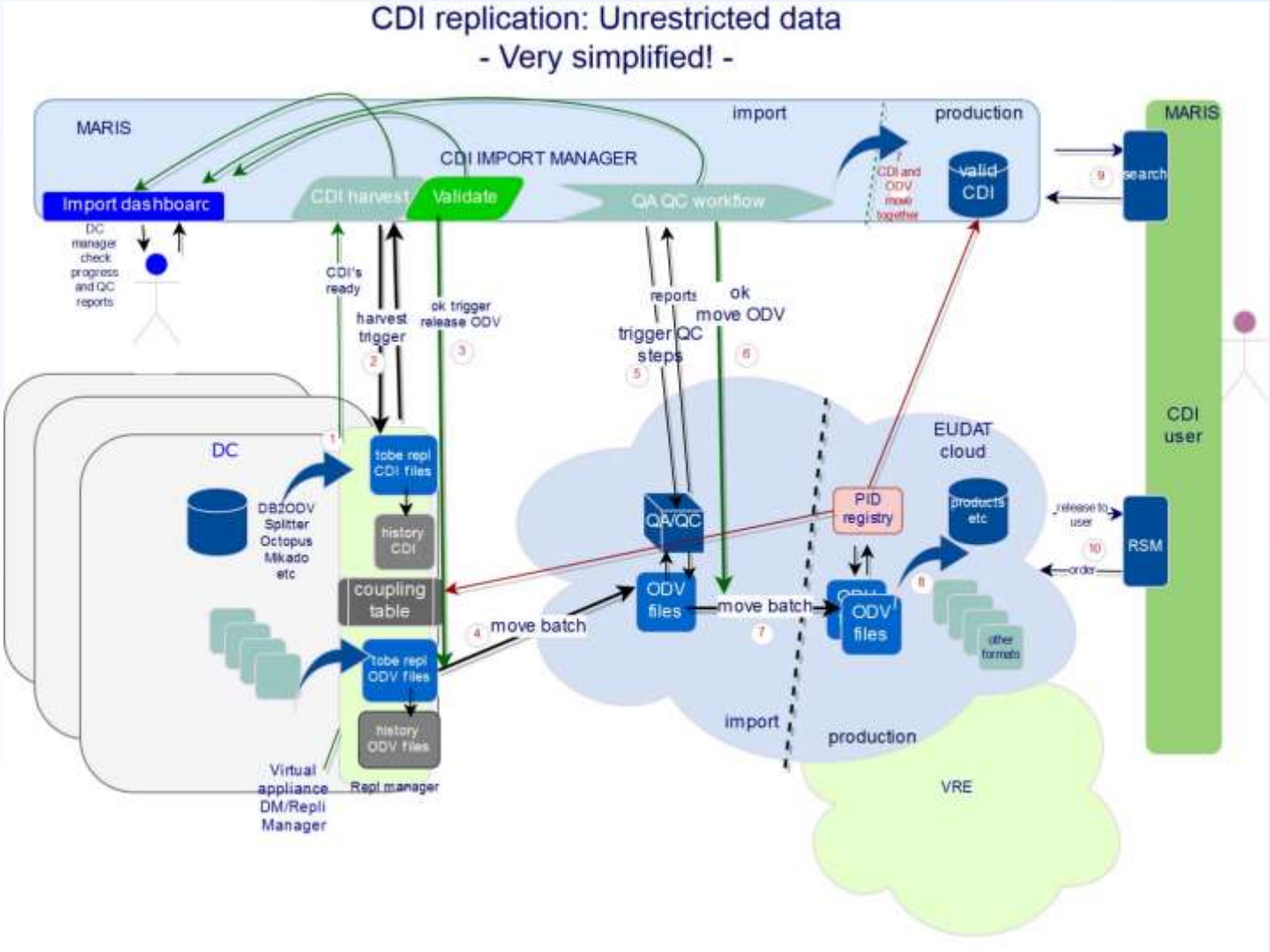


Figure ...: High level architecture

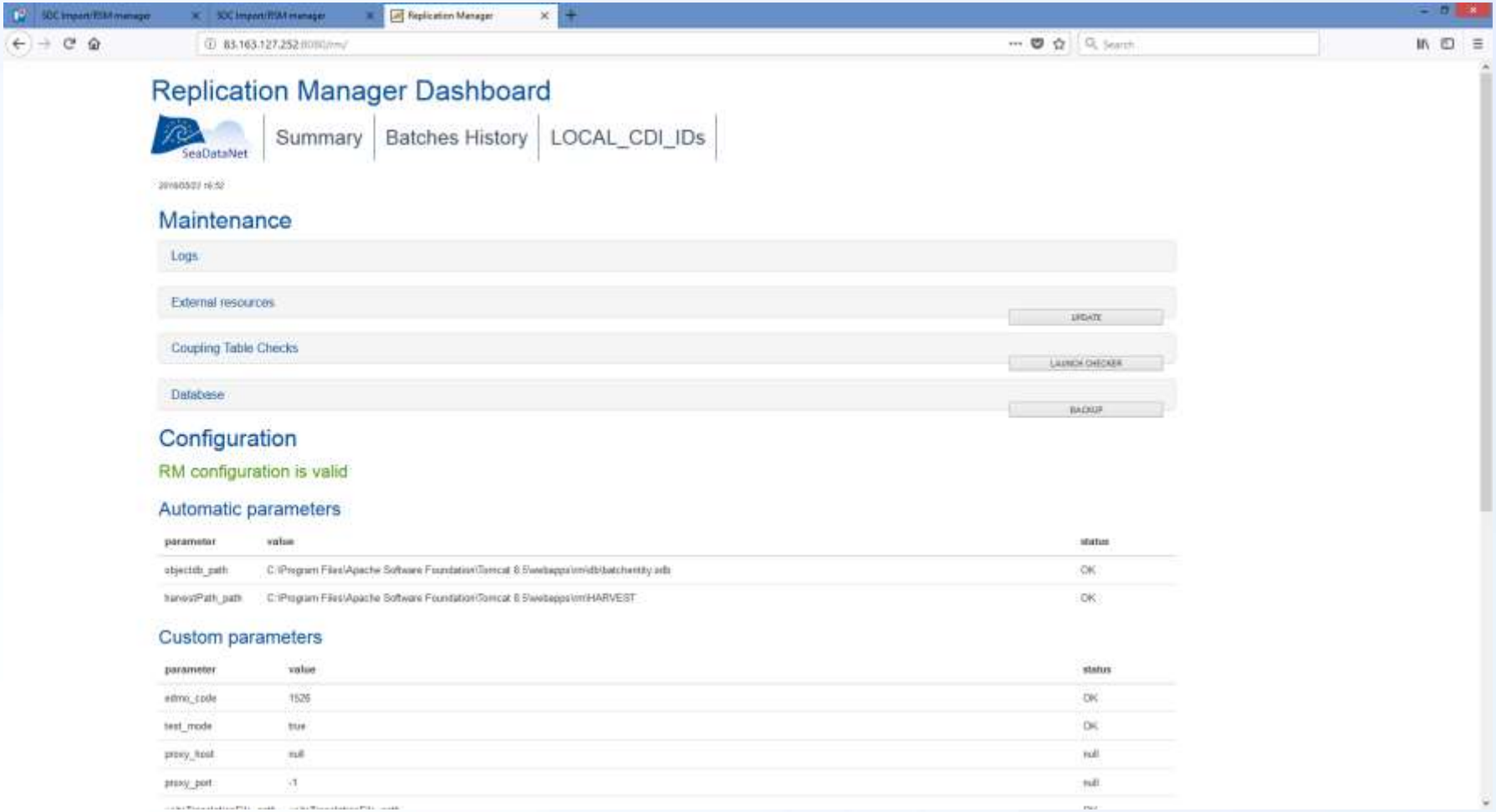
Unrestricted data flow



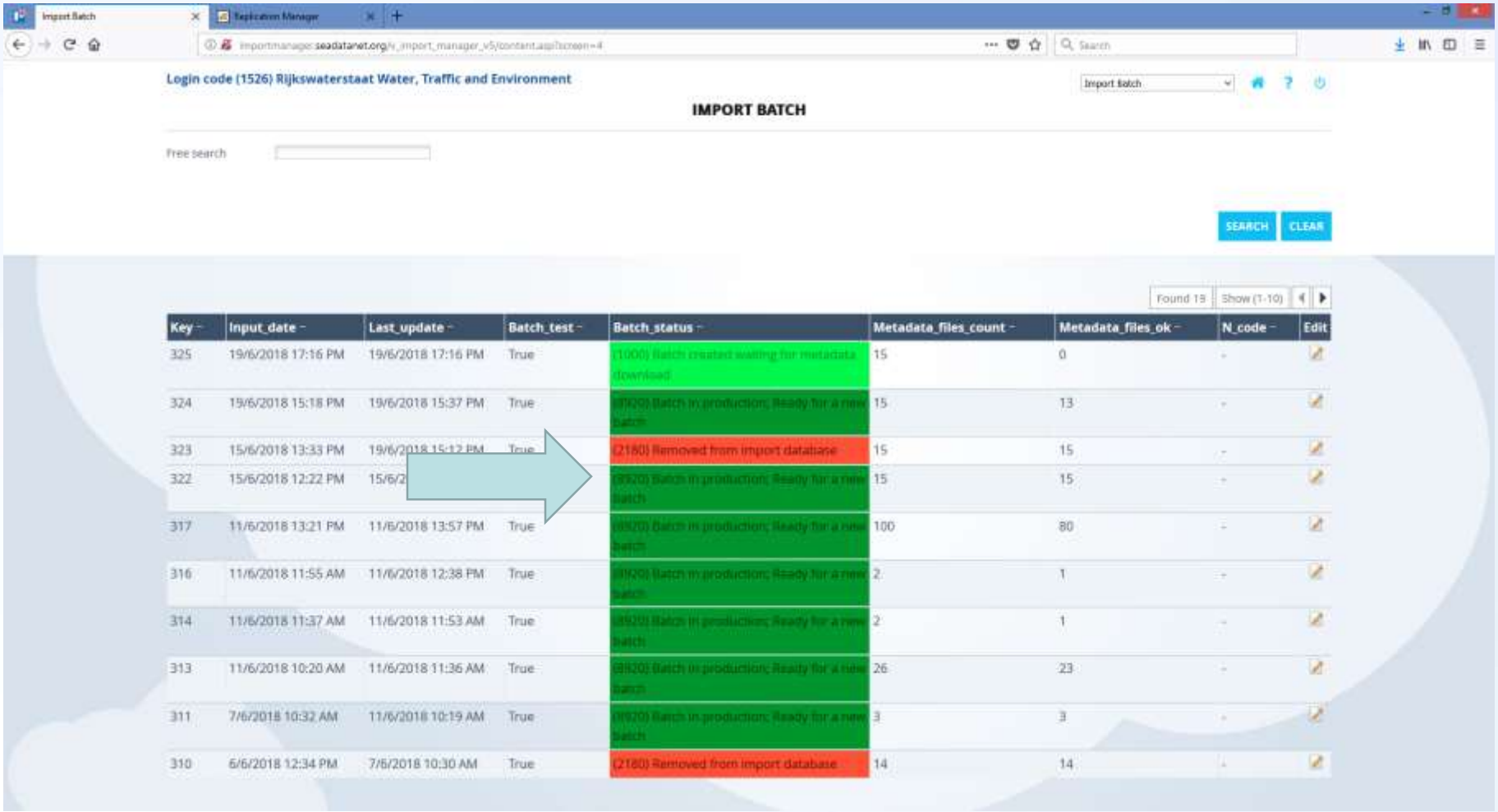
Unrestricted data ingestion process

1. Authentication
2. Publishing metadata
3. Checking metadata
4. Uploading a databatch
5. Unzip and run data checks
6. Move validated metadata and data in production

Replication Manager dashboard



Import Manager dashboard

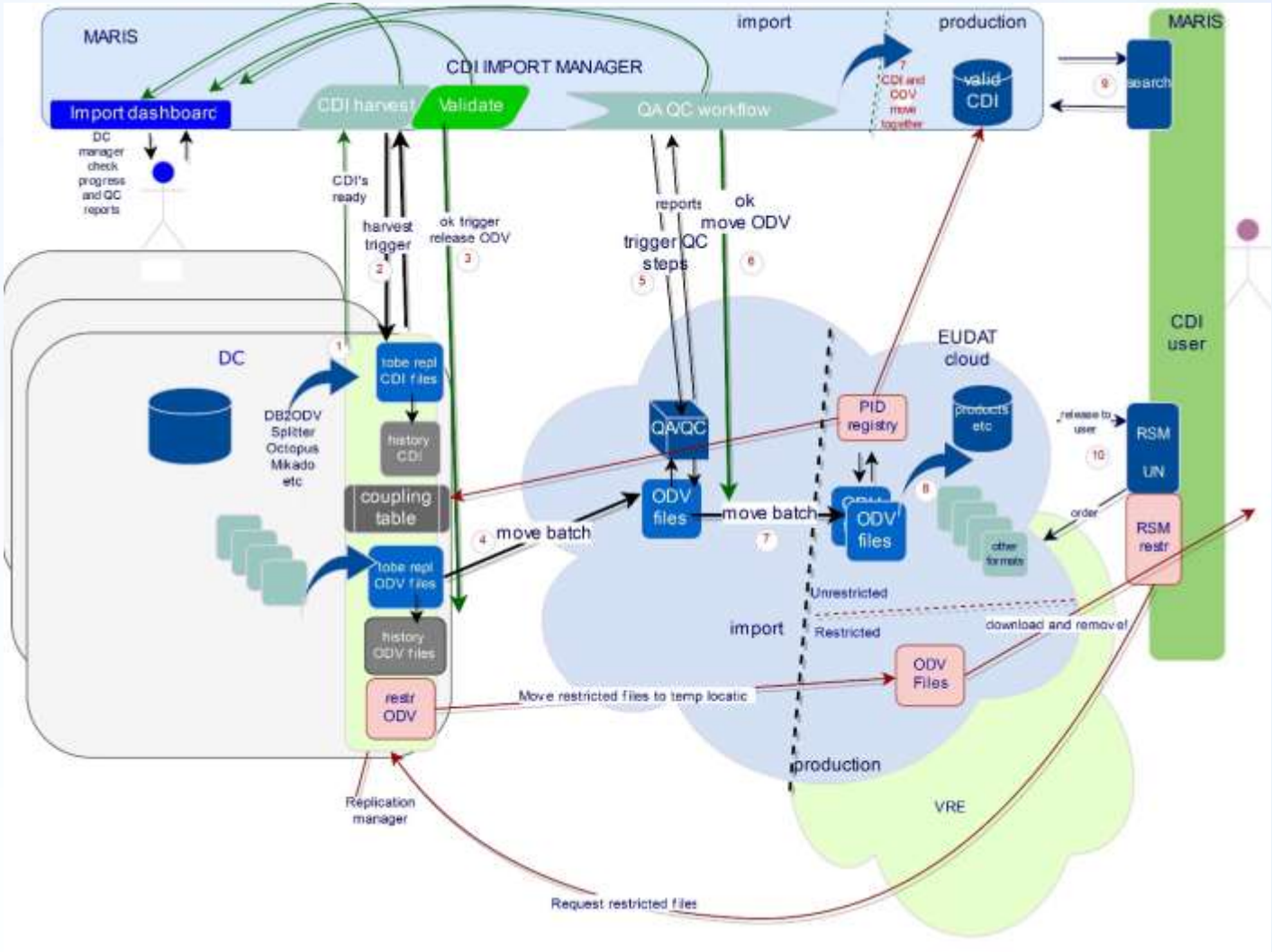


Key	Input_date	Last_update	Batch_test	Batch_status	Metadata_files_count	Metadata_files_ok	N_code	Edit
325	19/6/2018 17:16 PM	19/6/2018 17:16 PM	True	(1000) Batch created waiting for metadata download	15	0	-	
324	19/6/2018 15:18 PM	19/6/2018 15:37 PM	True	(8920) Batch in production; Ready for a new batch	15	13	-	
323	15/6/2018 13:33 PM	19/6/2018 15:12 PM	True	(2180) Removed from import database	15	15	-	
322	15/6/2018 12:22 PM	15/6/2		(8920) Batch in production; Ready for a new batch	15	15	-	
317	11/6/2018 13:21 PM	11/6/2018 13:57 PM	True	(8920) Batch in production; Ready for a new batch	100	80	-	
316	11/6/2018 11:55 AM	11/6/2018 12:38 PM	True	(8920) Batch in production; Ready for a new batch	2	1	-	
314	11/6/2018 11:37 AM	11/6/2018 11:53 AM	True	(8920) Batch in production; Ready for a new batch	2	1	-	
313	11/6/2018 10:20 AM	11/6/2018 11:36 AM	True	(8920) Batch in production; Ready for a new batch	26	23	-	
311	7/6/2018 10:32 AM	11/6/2018 10:19 AM	True	(8920) Batch in production; Ready for a new batch	3	3	-	
310	6/6/2018 12:34 PM	7/6/2018 10:30 AM	True	(2180) Removed from import database	14	14	-	

Restricted data flow

- CDI metadata same flow
- Data files extracted at same time as metadata and stored as local copies (to maintain versions)
- Shopping requests in RSM are evaluated by data centre
- If ok, data files will be temporarily released from local to cloud directory of user, upon request RSM

Restricted data flow



Benefits for users

- The performance will be increased, discovery and data requests improved, and downloading made more easy as each shopping request will provide one integrated download package instead of multiple packages from multiple data centres.

- Overall quality and coherence (data – metadata) will improve

- Tracking and tracing of data transactions will continue to be administered by an upgraded and much faster RSM service to oversee shopping requests and deliveries. The user RSM will be integrated as MySeaDataCloud service in the CDI user interface.

- Versioning of metadata and data will facilitate repeated analysis of e.g. environmental assessments in MSFD context after many years, and for scientific papers.

Benefits for data centres

- Data centres will have a **Replication Manager** module and an **Import Dashboard** to trigger and control themselves the import of new and updated metadata and data sets (unrestricted) into the CDI service

- Data providers can oversee all relevant transactions for their data centre in the upgraded and much faster RSM system and generate relevant reports

- The system will also support handling restricted data sets

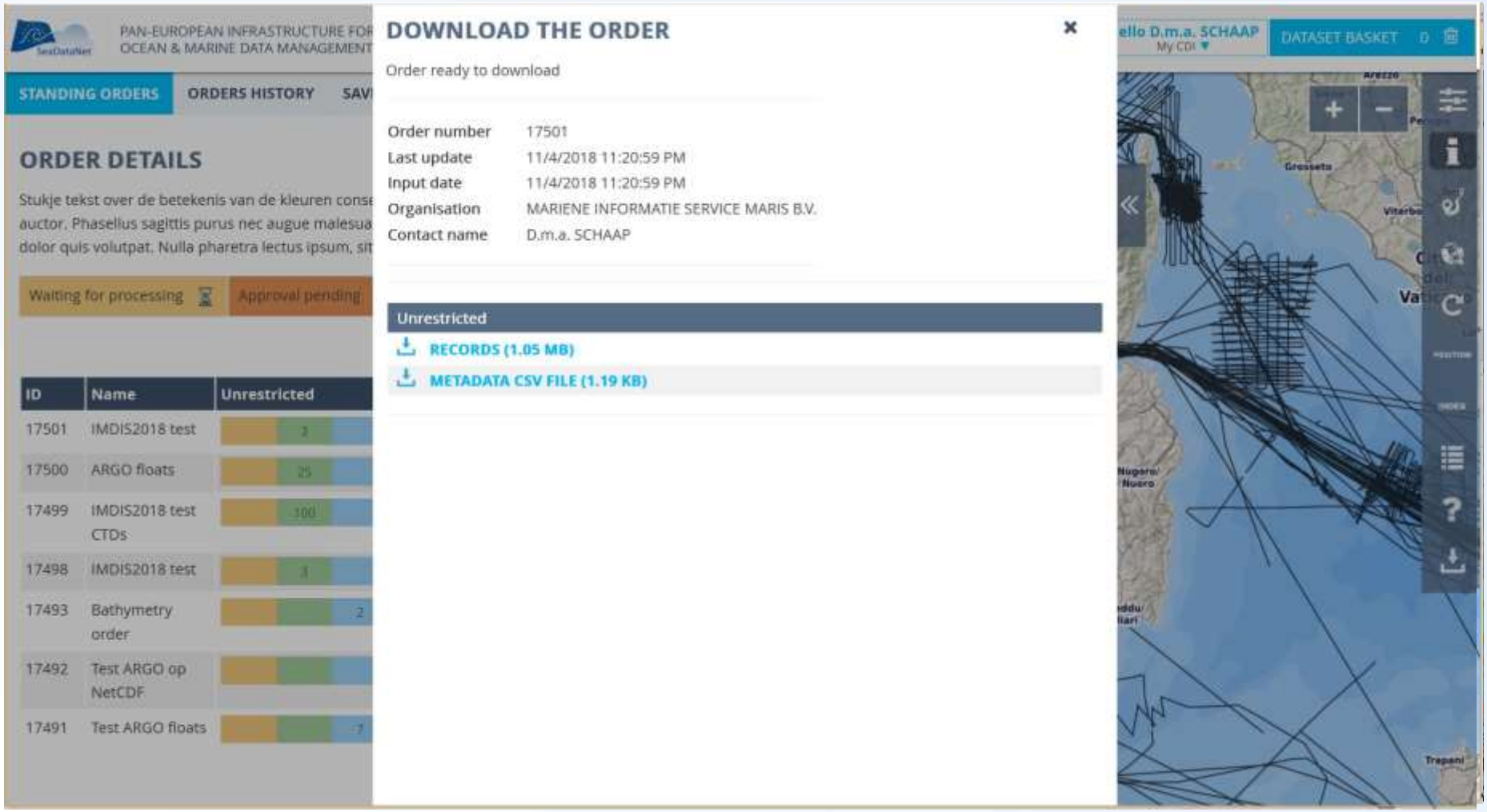
- Data centres will be outfitted with a Replication Manager (RM) replacing the Download Manager. The RM has less complexity and is easier to configure.

- Alternatively, Data centres can make use of the '**indirect solution**' which will be provided with improved functionality, handling both unrestricted and restricted data sets

New GUI on top of a new system

- Faster, using Elastic Search for search and indexing, and GeoServer for latest mapping technology
- Easier, including full text search next to controlled terms
- Modern design with large map and sliding windows
- **MySeaDataNet** integrated in GUI for customized services, such as SSO, RSM access, search profiles, prepared for VRE data pooling
- Developed and refined as prototype (http://sdc.maris.nl) interacting with users and data managers, following the SeaDataCloud Training Workshops in June 2018

New GUI impressions



SeaDataCloud

www.seadatanet.org