Technical challenges in SeaDataCloud

By
Dick M.A. Schaap – MARIS
Technical Coordinator SeaDataCloud

Riga – Latvia, 30 November 2016,
SeaDataCloud CG meeting no 1

SeaDataCloud general challenges

- SeaDataCloud is the successor to the SeaDataNet II project
- It is about updating and further developing standards
- It is about improving and innovating services & products
- It is about adopting and elaborating new technologies
- It is about giving more attention to users and putting the user experience in a central position
- Moreover, it is about implementing a strategic and operational cooperation between the SeaDataNet consortium of marine and ocean data centres and the EUDAT consortium of e-infrastructure service providers

SeaDataCloud overall objectives

- **Improve services to USERS:**
  - Improve the performance of the CDI data access services by utilising a cloud environment with high performance computing;
  - Provide online services to visualise and process data, in order to preview, subset, format, or analyse data of interest;
  - Apply and fully exploit the opportunities of the Linked Data concept as part of the semantic web to improve the discovery services;
  - Provide customised services (MySeaDataCloud services) to each user, to let the user save his/her search profile, receive alerts of new available data, and to ingest and manage their own data sets, complementing existing national arrangements with data centres;
  - Provide a Virtual Research Environment (VRE) to facilitate collaborative and individual research by users , by configuring and operating workflows of processing services, selected data and ingested own data, to generate value-added data products and visualisations;

SeaDataCloud overall objectives

- **Improve interoperability with other European and International networks to provide USERS overview and access also to these other data sources**

- **Improve services to DATA PROVIDERS:**

  - Provide online services for publishing finalised data collections, produced within the consortium or by other projects or scientific teams, for minting these collections with a DOI, and for safeguarding them with a long term perspective, interacting and tuning with national data centres;
  - Provide feedback to data providers via data usage reports and feedback on assessed quality of delivered data sets;

SeaDataCloud overall objectives

- **Optimise connecting DATA CENTRES to the infrastructure:**
  - Ease connecting data centres to the SeaDataNet infrastructure by revising and upgrading the existing components;
  - Provide an integrated package by means of a virtual application which will contain all the necessary software applications and an operating system;
- **Optimise connecting DATA STREAMS to the infrastructure:**
  - Facilitate connecting and ingesting data streams from operational observation networks, including metadata for sensors, instruments and platforms by making use of OGC Sensor Web Enablement standards

Technical Task Group (TTG)

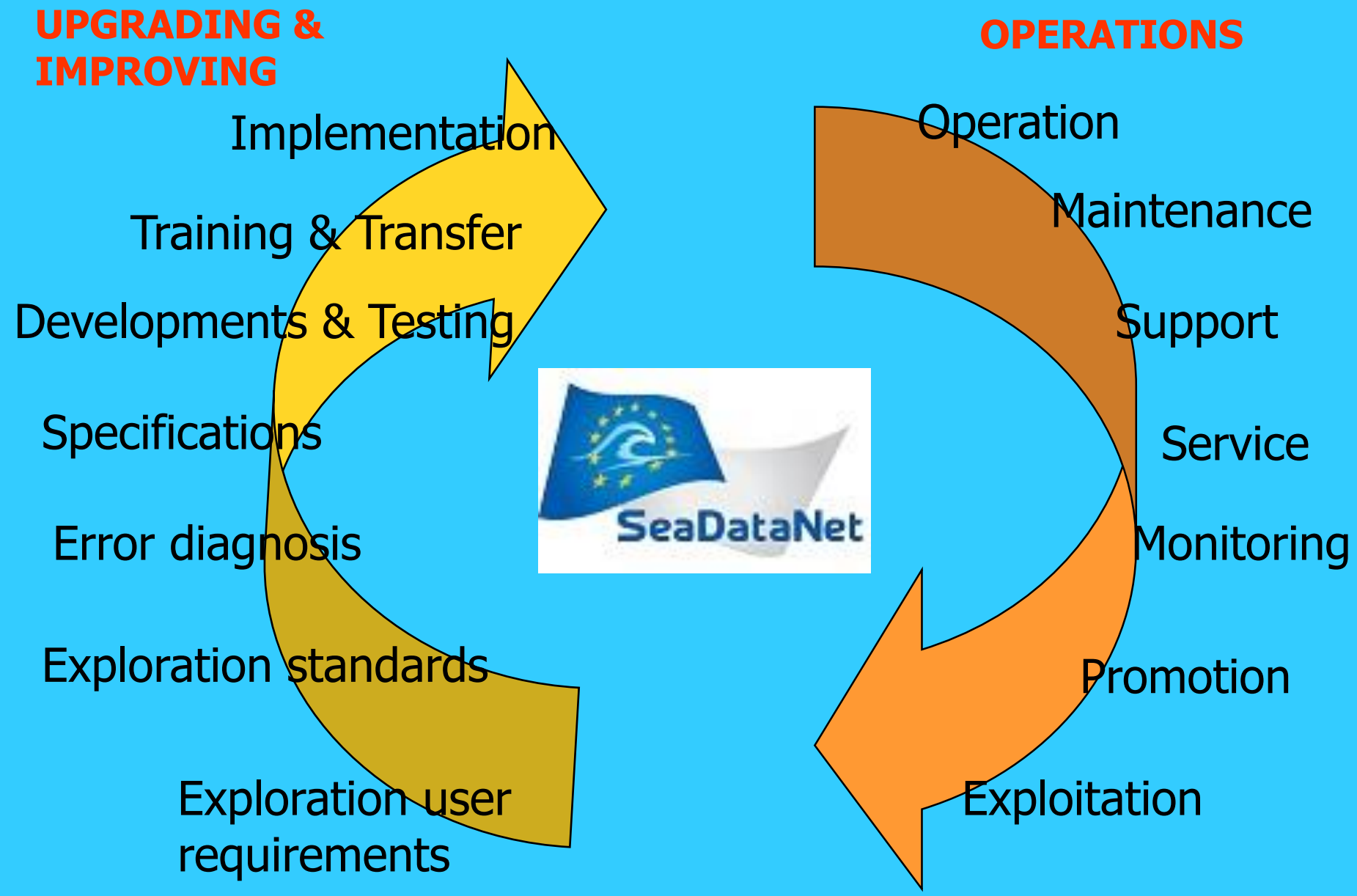
- The technical development represents a critical part of the project. It will be managed by the **Technical Coordinator (TC)** assisted by the **Technical Task Group (TTG)**

- The **TTG** is composed of the leaders of the technical development tasks, various experts in standards (such as ISO, OGC, W3C, oceanographic standards) and involved in the INSPIRE process, and the EUDAT experts in e-infrastructure, High Performance Computing and cloud storage.

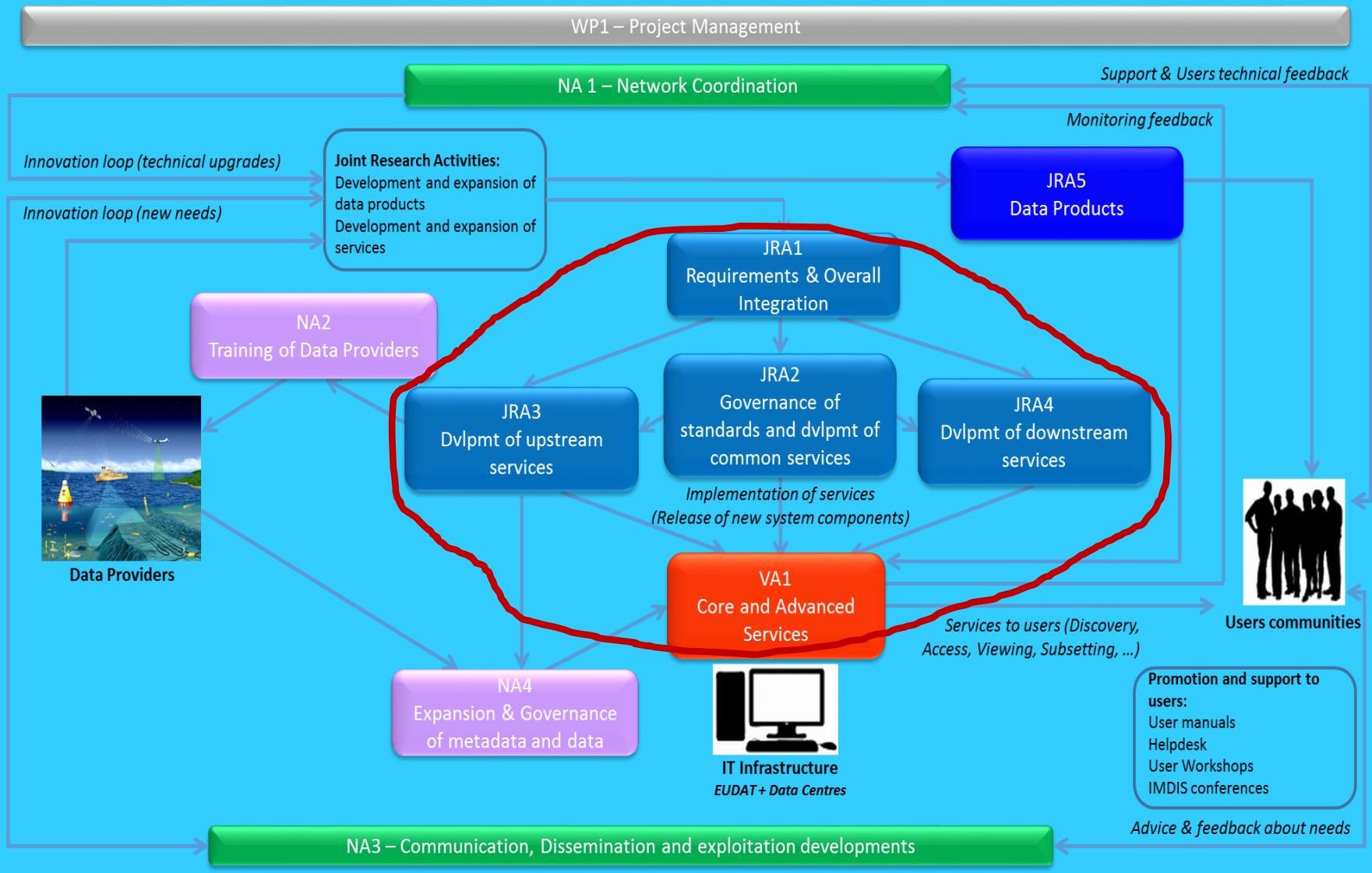
- The TTG will be the forum to coordinate and monitor all standards and technical developments.

- The TC and technical WP leaders will join the meetings of the **Scientific Committee** for interaction on specifications, developments and evaluations

SeaDataCloud – Overall approach - cycle



SeaDataCloud Pert diagram



WP8 - Governance of standards and development of common services - Led by BODC

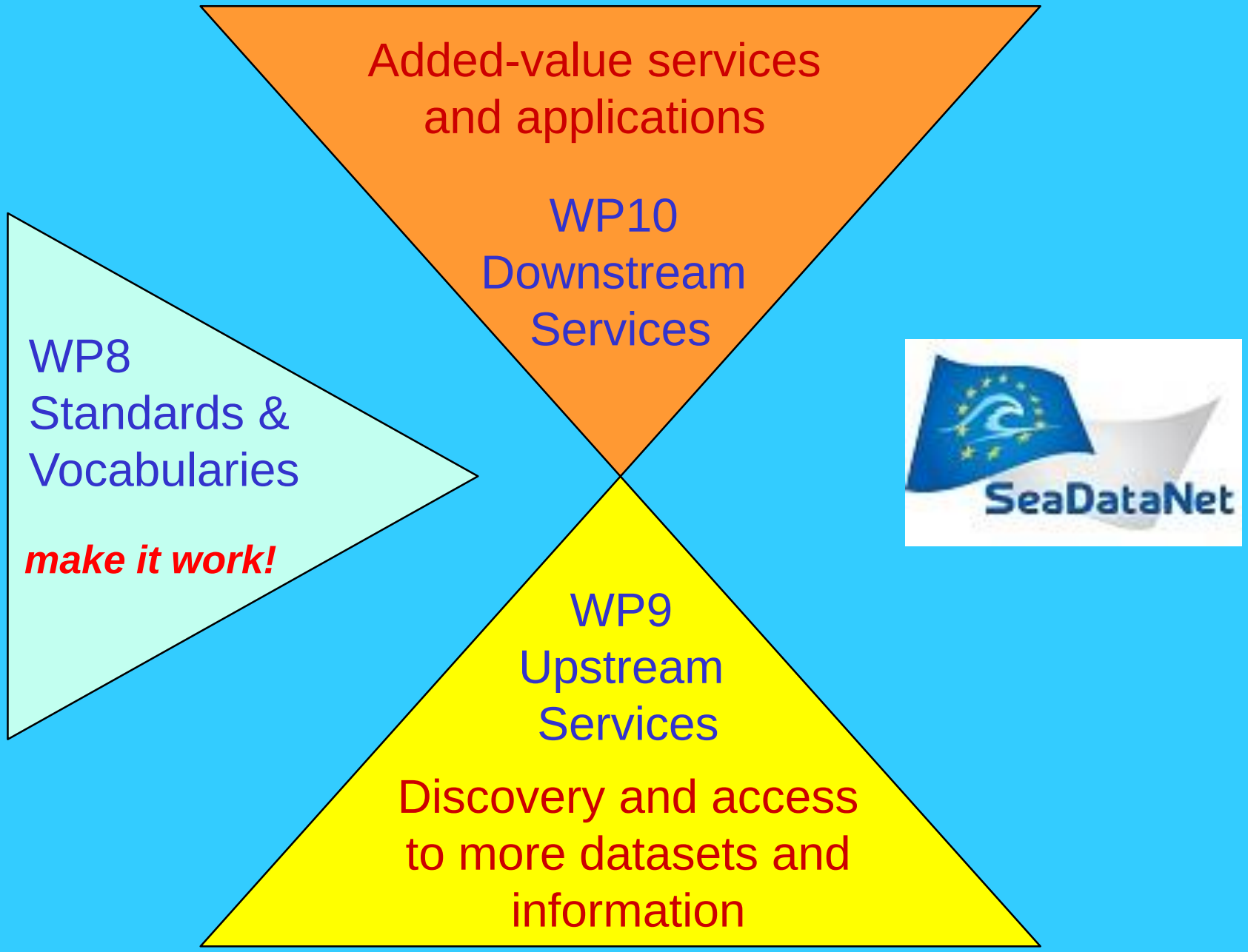
- To develop further the SeaDataNet controlled vocabularies and related services,
- To analyse and deploy a pilot for adopting the Linked Data principle for SeaDataNet directories,
- To review and expand the SeaDataNet data formats for achieving INSPIRE compliance,
- To integrate the SeaDataNet AAI services with GEANT/eduGAIN and social networks,
- To upgrade the SeaDataCloud monitoring service.

WP9 - Developments of upstream services - Led by MARIS

- To upgrade the CDI service making use of the cloud,
- To develop and deploy additional services in the cloud for ensuring integrity and conformity of the CDI related cloud data resources,
- To upgrade existing directory maintenance services and tools,
- To develop an online SWE ingestion service for operational observing systems,
- To expand SeaDataNet capability for handling different data types,
- To integrate external datasets from international programmes and organisations,
- To develop a preconfigured and pre-built virtual appliance system as complete solution to new data centres to connect to the CDI service,
- To develop a solution for a coordinated distributed DataCite DOI minting service.

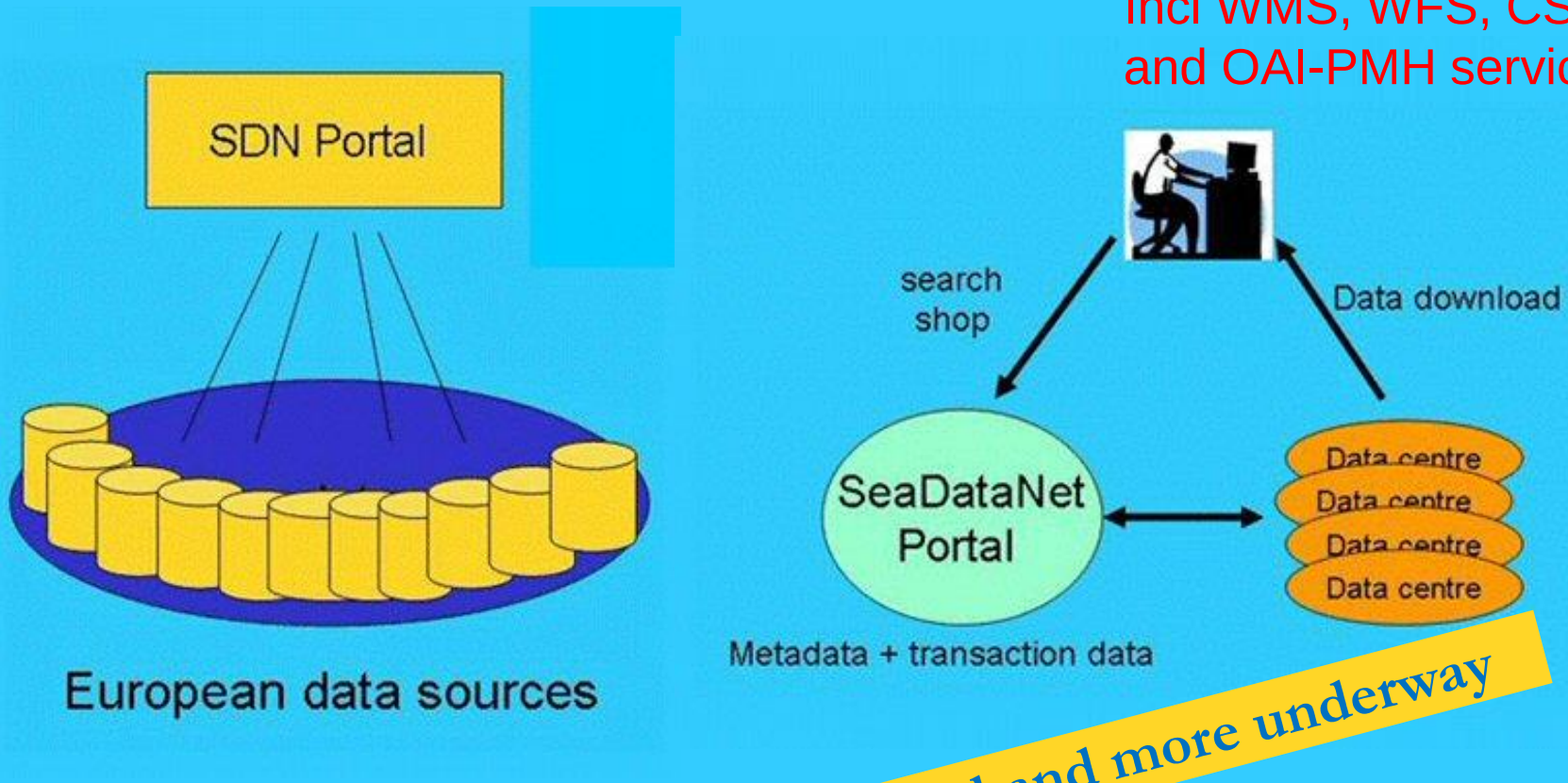
WP10 - Developments of downstream services - Led by IFREMER

- To expand the range of services of the SeaDataNet infrastructure by specifying, developing and deploying a Virtual Research Environment (VRE) with advanced e-services to facilitate individual and collaborative research by using, handling, curating, quality controlling, transforming and processing marine and ocean data into value-added analyses, harmonised data collections, and data products which can be integrated, visualised and published using OGC and high level visualisation services.


Added-value services
and applications
WP10
Downstream
Services
WP8
Standards &
Vocabularies
make it work!
WP9
Upstream
Services
Discovery and access
to more datasets and
information
SeaDataNet

Example: Common Data Index (CDI) service for discovery and unified access of data

Incl WMS, WFS, CSW and OAI-PMH services



Already > 100 data centres connected and more underway

Issues with present CDI service

- Population and uptake by Data Centres of the CDI service is very successful)> 100 Data Centres and > 1.8 million CDIs)
- However:
  - usage of the discovery and access services lags behind expectation; main reason is that users consider the access and delivery of datasets time-consuming and unattractive. Major obstacle is that users have to undertake multiple download transactions in case of shopping baskets with data from multiple data centres.
  - there are performance issues that data centres are not always online, operational and different machine capacities ; this gives extra delays
  - there are quality issues concerning formats of data files (ODV + NetCDF) and their consistency with CDI metadata
  - Installation and configuration of Download Manager software can be challenging due to different configurations, firewalls etc => there are different versions installed, because upgrading can give issues

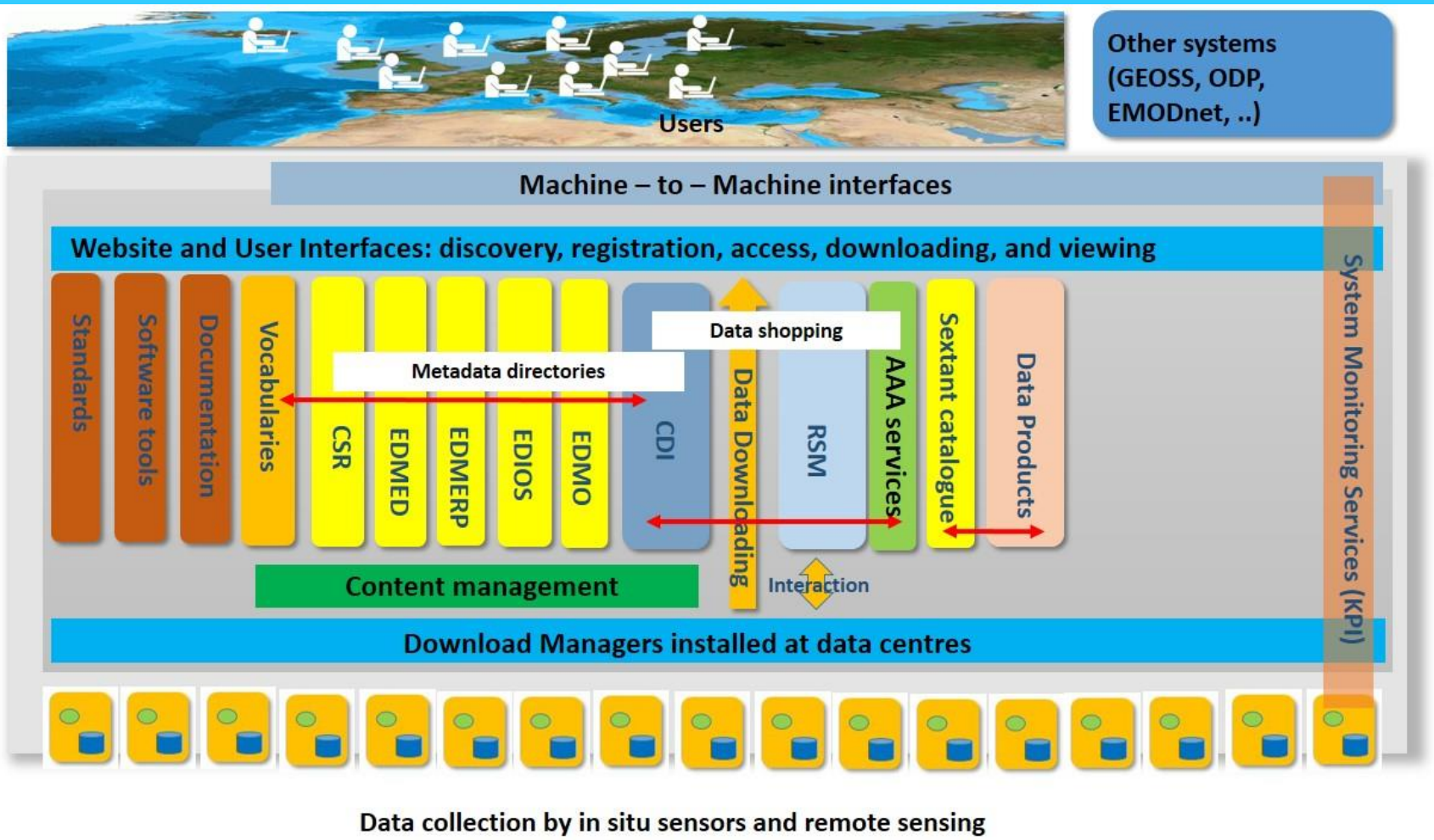
Upgrading CDI service using the cloud and HPC

- To configure and maintain a cloud environment with High Performance Computing (HPC) facilities to host **copies of all data resources**
- Exchange by dynamic **replication** from the individual data centres, following their updating of the CDI catalogue service
- In the cloud buffer:
  - checking possible duplicates
  - Checking overall quality of formats
  - 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 harmonizing data sets to common parameters and units, and converting data sets to other required output formats such as SeaDataNet NetCDF and relevant INSPIRE data models.

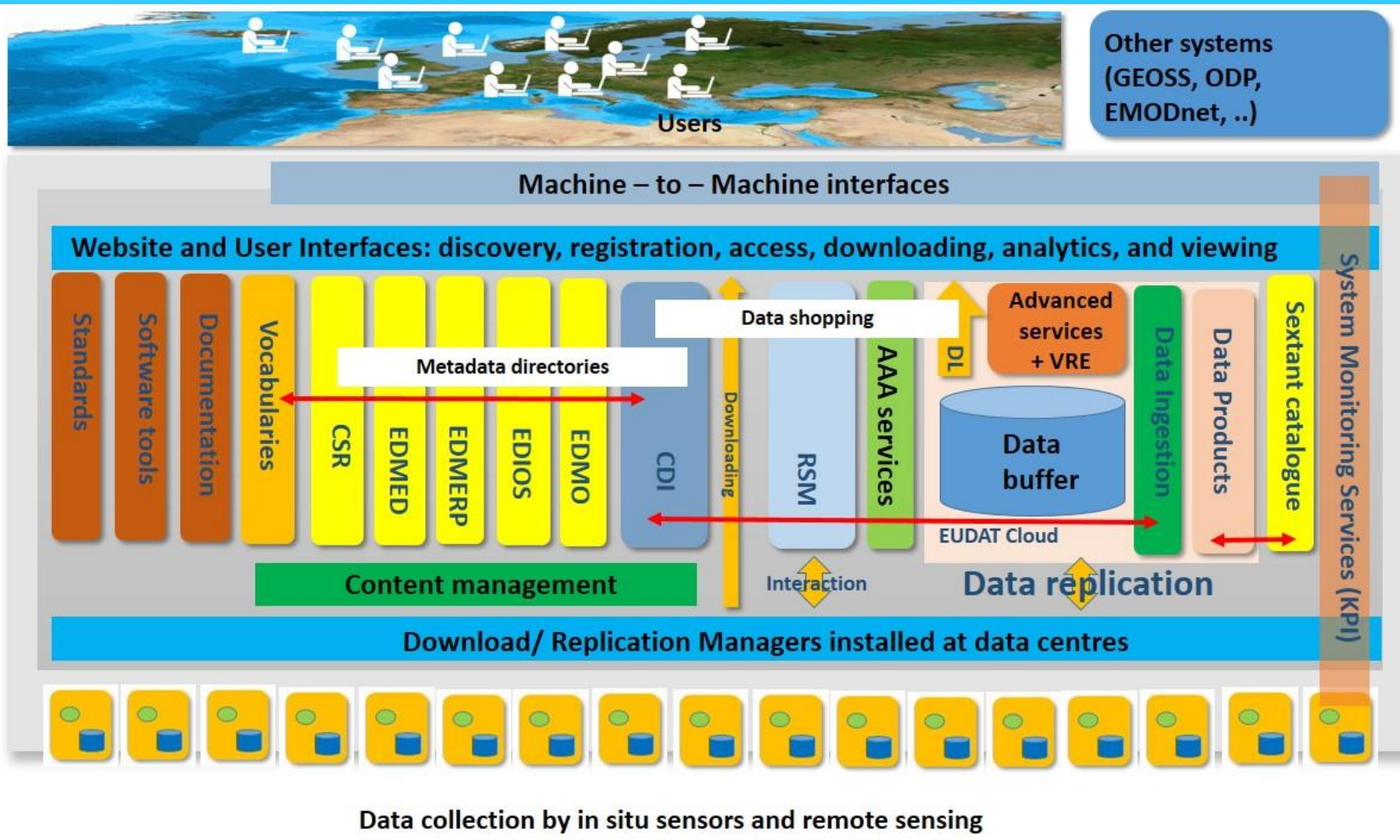
Potential benefits for CDI service and its users

- The cloud buffer in combination with the CDI service will speed up the performance and ease of use of the data access and downloading and will provide users with one integrated download package instead of multiple packages from multiple data centres.
- Overall quality and coherence (data – metadata) will improve
- Tracking and tracing of transactions will continue to be administered by the upgraded RSM service, so that data providers can oversee all relevant transactions for their data centre and users can oversee their requests and deliveries.
- 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.

Present SeaDataNet infrastructure



Upgraded SeaDataNet infrastructure



SeaDataCloud expected impacts

- Will improve access services for users considerably by
  - using the cloud for CDI service
  - applying Linked Data for semantic web
  - providing customised services (MySeaDataCloud)
- Will make many more data resources available to users by
  - connecting more data centres
  - more population from connected data centres
  - interoperability with international data systems (US-NODC, IMOS, ..)
  - SWE service for ingesting data streams from autonomous observatory platforms, making these available by means of catalogue and SOS services
  - using the EMODnet Ingestion service for external datasets from international programmes and organisations such as EuroGOOS ROOSes, Argo, WOD, a.o.

SeaDataCloud expected impacts

- Will facilitate users to optimally exploit the data resources for their research by:
  - highly efficient services for discovery, access, and transformation of data resources
  - a Virtual Research Environment (VRE) with a packaged set of advanced services for analysis, quality control, sub-setting, and visualisation of retrieved datasets and generation and publication of their data products

www.seadatanet.org