EOSC-hub

D7.1 First Thematic Service software release

Lead Partner:	Cineca
Version:	1
Status:	Approved by EC
Dissemination Level:	Public
Document Link:	https://documents.egi.eu/document/3411

Deliverable Abstract
This first release includes all the services, which have been integrated so far with EOSC-hub. The integration happens at various levels: from the usage of cloud resources (IaaS, PaaS), to the usage of EOSC-hub common services (for example data and computing services), to the usage of EOSC-hub core services, like the authentication or the monitoring tools, to the publication of the services in the Service Catalogue or in the Marketplace.

COPYRIGHT NOTICE

DELIVERY SLIP

Date	Name	Partner/Activity	Date
From:	Claudio Cacciari	Cineca/WP7	28/12/2018
Moderated by:	Malgorzata Krakowian		
Reviewed by:	Gergely Sipos Giacinto Donvito	EGI/WP8 INFN/WP10	
Approved by:	AMB		

DOCUMENT LOG

Issue	Date	Comment	Author
v.0.1	25/10/2018		Claudio Cacciari, Dieter Van Uytvanck, Willem Elbers, Daniele Spiga, Tobias Weigel, Sandro Fiore, Paolo Mazzetti, Mattia Santoro, Anabela Oliveira, Alberto Azevedo, Mário David, Alexandre Bonvin, Antonio Rosato, Brian Jimenez Garcia, Marco Verlato, Christian Briese, Michele Manunta, Marcin Gil, Simone Mantovani, Peter Baumann, Grega Milcinski, Fabrizio Pacini, Davor Davidovic
v. 0.2	04/11/2018		Claudio Cacciari
v. 0.3	12/11/2018		Claudio Cacciari
v. 0.4	16/11/2018		Claudio Cacciari
v. 0.5	28/11/2018	Reviewed by Gergely Sypos	Claudio Cacciari
v. 0.6	28/12/2018	Reviewed by Giacinto Donvito	Claudio Cacciari
v.1	09/01/2018	Final version	

TERMINOLOGY

https://wiki.eosc-hub.eu/display/EOSC/EOSC-hub+Glossary

Terminology/Acronym	*Definition*
Thematic services	Scientific services (incl. data) that provide discipline-specific capabilities for researchers. (e.g. browsing and download data and apps, workflow development, execution, online analytics, result visualisation, sharing of result data, publications, applications)

Contents

Executive summary

The research communities, which are partner in the project, are both service consumer and providers. They offer services to their users, which are called Thematic Services. As reported in the Terminology table above, Thematic Services are scientific services (incl. data) that provide discipline-specific capabilities for researchers (e.g. browsing and download data and apps, workflow development, execution, online analytics, result visualisation, sharing of result data, publications, applications). In some cases, in order to integrate those services with EOSC-hub, which is the main scope of the work package 7, they have modified the software code which implements them, extending existing components or developing new plugins. These set of software applications forms the current software release.

This first release includes all the services, which have been integrated so far with EOSC-hub. The integration happens at various levels: from the usage of cloud resources (IaaS, PaaS), to the usage of EOSC-hub common services (for example data and computing services), to the usage of EOSC-hub core services, like the authentication or the monitoring tools, to the publication of the services in the Service Catalogue or in the Marketplace. The following Thematic Services are included in this software release:

- CLARIN: only the Virtual Language Observatory (VLO) service is part of this release.
- DODAS: Dynamic On Demand Analysis Service.
- ECAS: ENES Climate Analytics Service.
- OPENCoastS: On-demand oPEratioNal Coastal circulation forecast Services.
- WeNMR: Worldwide e-Infrastructure for NMR and structural biology.
- EO Pillar: a set of services related to Earth science.
- DARIAH: only the Science Gateway service is part of this release.

Two Thematic Services, GEOSS (T7.4) and LifeWatch (T7.9) are not part of this release and will be included in the second release. The first is not included because according to its work plan they have just started the integration activities. The second postponed the integration activities because of an internal administrative issue, which has just been solved.

The software included in this release has been developed to enable the following integrations among Thematic Services and EOSC-hub services:

- **CLARIN**: the Virtual Language Observatory is published on the Marketplace and on the Service Catalogue.
- **DODAS**: it is currently integrated with several EOSC-hub service.
  - *Compute*: Infrastructure Manager, PaaS Orchestrator.
  - *Security*: Identity and Access Management (IAM) and Token Translation service.
  - *Data*: OneData, CVMFS stratum 0 and 1.
  - It is published on the Marketplace and on the Service Catalogue.
- **ECAS**:
  - *Security*: B2ACCESS, Indigo IAM and EGI Check-in.

- o   It is published on the Marketplace and on the Service Catalogue.
- **OPENCoastS**: it is published on the Marketplace and on the Service Catalogue.
- **WeNMR**:
    - o   *Compute*: DIRAC4EGI, EGI High-Throughput Compute.
    - o   *Security*: EGI Check-in.
    - o   *Data*: OneData, B2DROP
- **EO Pillar**: it is published on the Marketplace and on the Service Catalogue.
- **DARIAH**:
    - o   *Security*: Indigo IAM.
    - o   *Compute*: EGI FedCloud.
    - o   It is published on the Marketplace and on the Service Catalogue.

From the point of view of the software and services management procedures, the most common approach is to handle the software lifecycle, including the release management, internally to the community, providing the final code as public product and, in some cases, only the service. Often the defined procedures are quite simple in order to minimize the effort required to manage the software release.

1 Introduction

This document describes the Thematic Services software release. There is one chapter for each Thematic Service, which can be a group of services or just a single one. Each service has its own components and dependencies. The description of each service includes the references to the code, the documentation and the endpoints of the production instances. Moreover, the service providers introduced the software release description giving details about the context of the services in terms of purpose of the service, target users or resource providers. It seemed opportune to add those details, being this document the first deliverable related to the Thematic Services and hence, aside from the DoA, not existing any previous reference for the services' description. However, it is not intended to be an exhaustive explanation of the integration work done by the service providers and of their plans. That information will be part of the deliverable D7.2 (M12). This is a picture of the software already put in production and hence included in the release. All the activities related to prototypes and still ongoing developments will be described in the D7.2.

2 T7.1 CLARIN

The Virtual Language Observatory (VLO) is a service provided by CLARIN ERIC [1] offering uniform search and discovery functionality for (language) resources and tools. The metadata indexed is heterogeneous in terms of content and structure. This metadata is sourced regularly from over forty CLARIN centres that provide resources or tools of interest to scholars with an interest in language data. On top of that, CLARIN harvests metadata from a few dozen external providers that are also providing relevant content.

The VLO is openly accessible via the web to anyone. Researchers can freely enter a search term and/or use a number of pre-defined facets to refine the search results. This method of faceted browsing is as easy as using an online store and allows for quick filtering on basis of object language, nature of the resource, subject or organisations involved. Search results can be used to obtain a link to the associated resources and/or process the search result in the language resource switchboard. Repository administrators' interest in integration of their metadata into the VLO can contact the support team.

The VLO production and development instances are running on specialized cloud nodes provided by a commercial provider. The beta instance is running on an infrastructure provided by the MPCDF [2], an academic data centre. This VLO has specific requirements with respect to disk IOPS and memory, therefore the cloud node has been optimized for high IOPS and has an above average amount of memory.

2.1 Service architecture

The VLO is a java web application, using Apache Wicket [3] and spring for the front-end. For the backend there is a SOLR [4] database holding the metadata index and a harvester command line application running periodically to update the SOLR index with harvested metadata from the providers. This is shown in the following high-level architecture diagram (Fig. 1) of the VLO and metadata harvester:
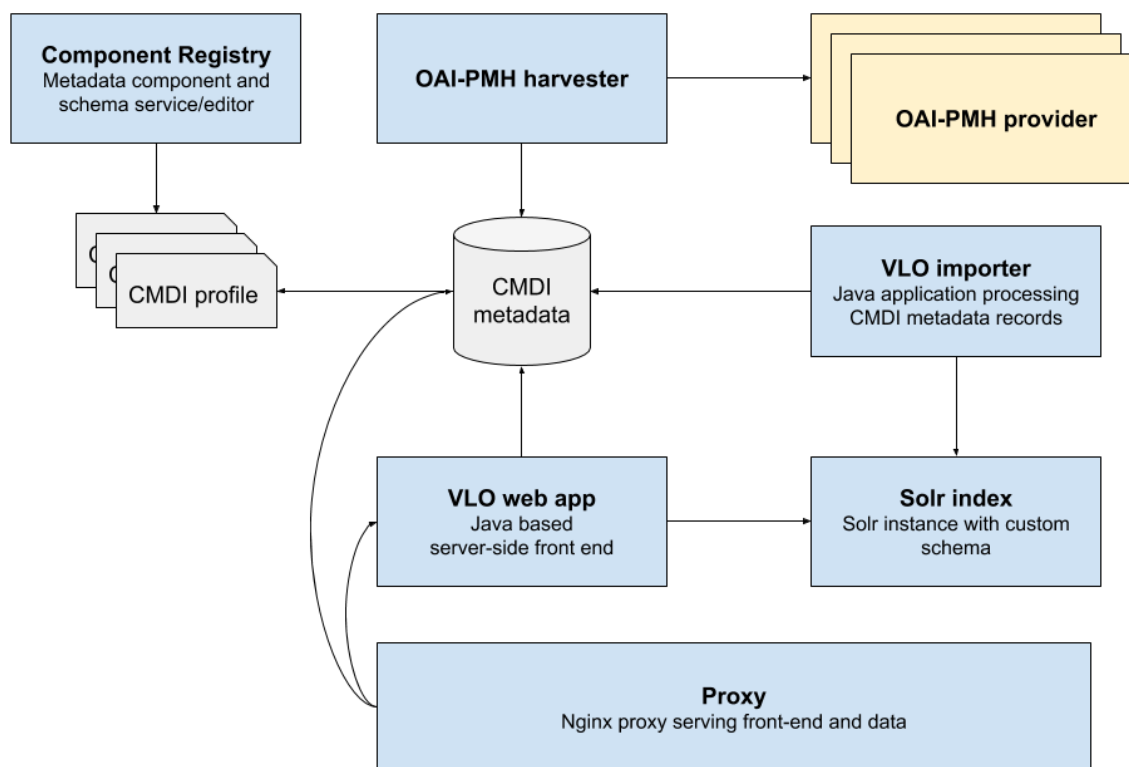
Fig. 1 - High level architecture diagram

Metadata is harvested using the OAI-PMH [5] protocol and is required to follow the CMDI [6], DC [7] or OLAC [8] specification.

2.2 Software release

Within CLARIN the general direction and prioritization of issues is decided internally, taking into account bug reports and user feedback. The concrete items resulting from this process are managed as issues and milestones on GitHub [9], resulting in a medium term roadmap.

Development versions are made available on https://alpha-vlo.clarin.eu. After completing a milestone, a release candidate is provided and made available on https://beta-vlo.clarin.eu. This version is tested and evaluated by a pool of testers from within the CLARIN community. Several iterations of bug fixes, grouped into one or more new release candidates, are made available until a stable instance is available. This version is then released into production and deployed on https://vlo.clarin.eu/.

All releases are managed with major, minor and path releases, where major releases can include breaking changes, minor version include new features and patch releases provide bug fixes. This approach is inspired by the semantic versioning [10] principles.

The entire release is packaged as Docker images and accompanied with a Docker-compose setup for easy deployment, either locally or in the CLARIN infrastructure

2.3 References

- Service endpoints:
    - https://vlo.clarin.eu
- Software code and releases:
    - Software code and release are managed in GitHub:
        - https://github.com/clarin-eric/VLO(/releases)
        - https://github.com/clarin-eric/oai-harvest-manager (/releases)
    - Docker images and compose packages are managed in gitlab:
        - https://gitlab.com/CLARIN-ERIC/docker-vlo-beta
        - https://gitlab.com/CLARIN-ERIC/compose_vlo
        - https://gitlab.com/CLARIN-ERIC/docker-oai-harvester
        - https://gitlab.com/CLARIN-ERIC/compose_oai_harvester
- Documentation: user oriented, administrator oriented and developer oriented:
    - https://trac.clarin.eu/wiki/CmdiVirtualLanguageObservatory
        - This is a closed wiki. On request we can provide an export of the relevant wiki pages

Instructions to request access are available on https://www.clarin.eu/dev

2.4 EOSC-hub Integration

The latest stable VLO version, the v.4.5.1, is published via EOSC-hub:

- on the Marketplace: https://marketplace.eosc-hub.eu/component-metadata-infrastructure/96-virtual-language-observatory.html;
- on the Service Catalogue: https://www.eosc-hub.eu/catalogue/Component%20MetaData%20Infrastructure

3 T7.2 DODAS

The main feature of DODAS can be summarised as follows:

- To provide a simple yet complete abstraction of hybrid cloud infrastructures
- To automate both virtual hardware provisioning and configuration
- To provide a cluster platform with a high level of self-healing and scalability
- To guarantee set-up and service customization to cope with specific scientific requirements.

DODAS completely automates the process of provisioning, creating, managing and accessing a pool of heterogeneous computing and storage resources, thus drastically reducing the learning curve, as well as the operational cost of managing community-specific services running on distributed clouds.

Currently DODAS provides support to deploy:

- HTCondor[11]-based batch system as a Service
- Big Data platform for ML as a Service

DODAS is currently adopted as a solution for several use cases:

- Exploitation of opportunistic computing, intended as resources not necessarily or permanently dedicated to a single experiment and/or activity;
- Elastic extension of existing facilities, to absorb peaks of resource usage;
- Generation of on-demand batch systems for data processing.
- Instantiation of Spark cluster for data processing and reduction

As such DODAS targets resource provider which need a smart solution for elastic extension of their facilities e.g. to provide additional (opportunistic) resources to the supported collaborations. Moreover, DODAS target researcher of small scientific communities who needs a technical solution to effectively exploit cloud resources (both public and private). Typical use case is to be able to make use of a batch system out of a Cloud bunch of resources. Users who need to develop and train ML/DL models might need platforms. Unless there is dedicated IT support, it would be required to spend a lot of time and effort for setting up properly working setup. In such a case, DODAS allows to save user's time by reducing IT-specific effort in favour of research time/effort. Generally speaking, DODAS targets any user who needs to interact with a cloud environment to setup a complex computing scenario. The motto is: exploit any cloud with almost zero effort.

Within the EOSC-hub project, DODAS-Thematic Service is providing both the Core Services and an Enabling Facility. DODAS Core services can be used to exploit any cloud the user as granted access. Besides there is a freely accessible Enabling Facility where users can test a customisation and/or simply try out how DODAS behaves. Data and compute resources of this Enabling Facility are offered by two distinct providers at INFN: Cloud@CNAF and ReCaS@Bari. Both of them are based on Openstack [12] middleware. The Enabling Facility is freely accessible through DODAS PaaS core services, upon successful registration and authentication on https://dodas-iam.cloud.cnaf.infn.it/.

Note this is the same AuthN/Z user need to register to access the Core Services. The latter are provided by INFN and physically run at CNAF.

About 50 CPUs and a few TBs of storage have been dedicated by INFN in order to provide a Highly Available DODAS Core system deployment. Moreover, DODAS Thematic Service offers a distributed Enabling Facility. Regarding the latter INFN provides about 700 cores as compute nodes, integrated within the Cloud facility installed at INFN-CNAF in Bologna and the Cloud Facility installed at ReCaS, Bari. About 120 TB of disk storage and a total of about 1.4 TB of RAM will be reserved. At the time of writing roughly half of the above resource estimation is actually provided. Quota will increase up to the nominal values above mentioned in the upcoming months. Other resource providers grant access to resources via DODAS, Imperial College London provide resources via DODAS both on local Openstack facility and on AWS[13]. During 2018 additional resources provided by DODAS also come from the fruitful collaboration with HelixNebula Science Cloud [14] project.

3.1 Service architecture

DODAS has a highly modular architecture and the workflows are highly customisable. This is a key to the extensibility, spanning from software dependencies up to the integration of external services passing through the user tailored code management. There are four main pillars in the DODAS architecture, which can be summarized as:

- **Abstraction**: both in term of software application and dependency descriptions and in term of underlying IaaS.
- **Automation**: of software and application setup; to manage resources and orchestrate software applications.
- **Multi-cloud support**: to deal with multiple heterogeneous Cloud infrastructures.
- **Flexible AAI**: authentication, authorization, delegation and credential translation primitives providing a secure composition of the various services participating in the DODAS workflow.

The DODAS core system has been realized mostly using building blocks originally developed by the INDIGO-DataCloud project and currently part of the EOSC-hub's portfolio. One is the PaaS Orchestrator, which has the role of taking the requests related to application, or service deployment coming from the user expressed using TOSCA, another is the OASIS standard to specify the topology of services provisioned in IT infrastructures. Based on the user requirements (typically expressed in the TOSCA template), the Orchestrator has the role to identify the best infrastructure (IaaS) for the deployment taking into account information about user's SLAs the availability and the health status of the IaaS services. The actual interaction with the infrastructure is delegated to the Infrastructure Manager (IM).

The IM is in charge to deploy complex and customized virtual infrastructures on different IaaS Cloud deployment, providing an abstraction layer to define and provision resources in different clouds and virtualization platforms. IM enables computing resource orchestration using TOSCA protocol. Moreover, it eases the access and the usability of IaaS clouds by automating the VMI (Virtual Machine Image) selection, deployment, configuration, software installation, monitoring and update of the virtual infrastructure. The glue of the implemented flow is the Identity and

Access Management service (IAM). IAM provides a layer where identities, enrolment, group membership attributes and policies to access distributed resources and services can be managed in a homogeneous and interoperable way. It supports the federated authentication mechanisms behind the INDIGO AAI. The IAM service provides user identity and policy information to services so that consistent authorization decisions can be enforced across distributed services. Identity and Access Management is provided through multiple methods (SAML [15], OpenID Connect [16] and X.509 [17]) by leveraging on the credentials provided by the existing Identity Federations (i.e. IDEM [18], eduGAIN [19], etc.). The support to Distributed Authorization Policies and Token Translation Service will guarantee selected access to the resources as well as data protection and privacy.

The resource abstraction and the full automation, one of the pillar stone of DODAS, are thus implemented combining together the PaaS Orchestrator and the IM.
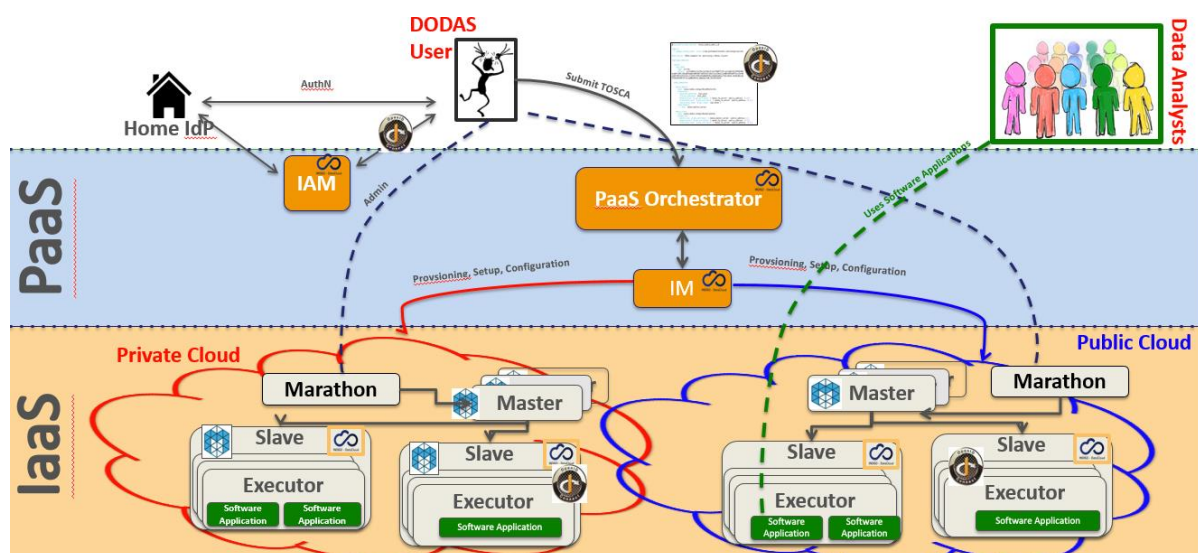

Fig. 2 - A high-level schema of the DODAS architecture. Three colours shown User Domain, DODAS PaaS core service layer and resource layer (from top to bottom).

3.2 Software release

Being a composition of services, there is not a single release, nor a single procedure. Also about the roadmap, not a single procedure is foreseen. Core services are kept aligned to the latest stable version (provided by sub services). Experiment/community software is not under DODAS responsibility. TOSCA Templates are those who actually represent the software, from the user perspectives. From a technical point of view TOSCA templates need to guarantee coherence between underlying services actually not directly exposed to the users. Said that, DODAS provides releases of the TOSCA templates, but a well-defined procedure is still under definition. This is an objective of the project for M18.

3.3 References

Link and references to Service endpoints:

- DODAS login page: https://dodas-iam.cloud.cnaf.infn.it/login.

Software code and releases:

- Community images: https://github.com/DODAS-TS.
- Tosca templates: https://github.com/indigo-dc/tosca-templates/blob/master/dodas.

Documentation:

- https://dodas-ts.github.io/dodas-doc

3.4 EOSC-hub integration

DODAS is currently integrated with several EOSC-hub service namely: Infrastructure Manager, PaaS Orchestrator, Identity and Access Management (IAM) and Token Translation service are part of the core. IAM has been further integrated also in the CMS workflow supported by DODAS. In the latest months, DODAS integrated also OneData as a possible solution for the data ingestion and for transparent data access. CVMFS stratum 0 and 1 have been also integrated in order to support newer uses cases coming from AMS-02 experiment requirements. Finally, DODAS stable version is published via EOSC-hub:

- On the Marketplace:
  https://marketplace.eosc-hub.eu/51-dynamic-on-demand-analysis-service.
- On the Service Catalogue:
  https://eosc-hub.eu/catalogue/Dynamic%20On%20Demand%20Analysis%20Service.

4 T7.3 ECAS

The ENES Climate Analytics Service (ECAS) enables scientific end-users to perform data analysis experiments on large volumes of multidimensional data (e.g. NetCDF data format), by exploiting a PID-enabled, server-side, and parallel approach. It aims at providing a paradigm shift for the ENES community with a strong focus on data intensive analysis, provenance management, and server-side approaches as opposed to the current ones mostly client-based, sequential and with limited/missing end-to-end analytics workflow/provenance capabilities. A dedicated goal of this service is to engage end-users directly, interact with them and induce a cultural change for the scientific workflow. The activities will also bring the ENES data infrastructure (IS-ENES, the European contribution to the Earth System Grid Federation) closer to an amalgamation with EUDAT, EGI and INDIGO.

ECAS can be used by the following general user groups:

1) Scientific users from the <u>core climate modelling community</u> who do not have access to sufficient computing resources for large climate data analysis experiments;
2) Scientific users from the <u>direct downstream usage communities</u>, such as climate impact studies, who are not as familiar with the data design and ESGF data distribution mechanics and will therefore, benefit both from easier accessibility of processing and transparent selection of input data without the need to understand data locality or arrange data transfers.
3) Scientific users from <u>other research communities</u> who are dealing with same data model (multidimensional data) and similar challenges (large scale data analysis). Based on the experience coming from INDIGO-DataCloud, additional communities that could benefit from the ENES Climate Analytics Service could be: (i) EMSO (European Multidisciplinary Seafloor and water-column Observatory) [20], (ii) LBT (Large Binocular Telescope) [21] and also (iii) LifeWatch.

In the first phase of the project, ECAS is primarily addressing ENES-related use cases (points 1 and 2).

The main service provider organizations are:

1. Deutsches Klimarechenzentrum GmbH (DKRZ)
2. Fondazione Centro Euro-Mediterraneo sui Cambiamenti Climatici (Fondazione CMCC)
3. EGI will represent the third option for service providers willing to deploy, manage and exploit their own instance of ECAS in the EGI cloud to serve more user groups. Such possibility will be available during the second part of the project, once the integration activities related to EGI will be completed (M18).

They have allocated the following resources to the services:

- ECAS@CMCC: "fat nodes" analytics cluster with:

- 1 node client with 4 cores, 8GB memory, 12GB disk, for the client-side components, (e.g. the Ophidia CLI and PyOphidia module), and the services (e.g. JupyterHub);
- 1 server node with 8 cores, 16GB memory, 12GB disk, executing the Ophidia Server;
- 5 nodes, 20 cores/node, RAM 256GB/node, 60TB disk space shared via GFS + local disks.
- ECAS@DKRZ:
  - 1 node client (VM) with 4 cores, 15GB memory, 50GB local disk + 457 TB shared Beegfs disk space, for the services JupyterHub and Docker;
  - 1 server node (VM) with 4 cores, 4GB memory, 30GB disk, executing the Ophidia server;
  - 1 computing node (Blade) with 40 cores, RAM 256GB, 2TB local disk, executing Ophidia I/O server + 457 TB shared Beegfs disk space.
  - CMIP5 and CMIP6 data pool (read-only)

4.1 Service architecture

The following figure (Fig. 3) provides an overview of the initial ECAS architecture. The components included in the system can be characterized as core and integrated. The integrated components relate to the integration activities performed with regard to INDIGO-DataCloud, EGI and EUDAT services. Some key examples are B2DROP, IAM, OneData, B2SHARE, IM/Orchestrator (see paragraph 4.4).
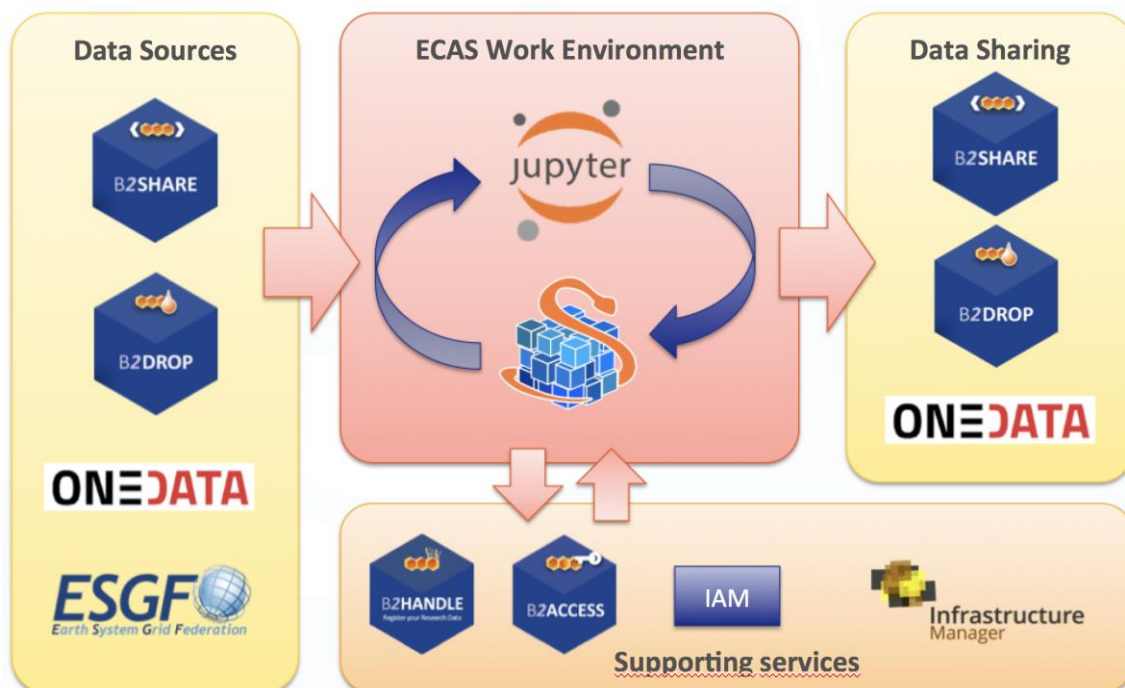

Fig. 3 - Overview of the initial ECAS architecture

1.1.1 Core Components

4.1.1.1 Ophidia

Ophidia represents the core component of ECAS. It is a big data analytics framework for scientific data. It provides declarative, server-side, and parallel data analysis, jointly with an internal storage model able to efficiently deal with multidimensional data and a hierarchical data organization to manage large data volumes. From a physical point of view, the data is partitioned in fragments consisting of multidimensional binary arrays and distributed over multiple nodes. Ophidia also provides a native analytics workflow engine, to define processing chains and workflows with tens to hundreds of data analytics operators to build real scientific use cases. It provides about 100 array-based functions and more than 50 datacube-based operators to enable OLAP tasks. Ophidia exposes several interfaces to address interoperability: WS-I, OGC-WPS, and GSI/VOMS. From the client-side perspective, a CLI, as well as a Python module, are provided to run interactive or batch experiment sessions.

4.1.1.2 WPS interface

ECAS offers the Web Processing Service (WPS) standard interface through the PyWPS [22] component. Such endpoint exposes a set of processes one for each operator, which enable WPS-based client-server interactions, and also facilitates interoperability with ESGF.

4.1.1.3 JupyterHub

JupyterHub [23] allows running a multi-user environment for the execution of multiple instances of the Jupyter Notebook service. It manages also authentication and authorization of users to their notebooks. A Jupyter Notebook allows creating, editing, running and sharing documents containing live code, equations and plots via a web browser. It supports several languages, such as Python, R, Scala. Additionally, it provides a dashboard (similar to a file manager) to manage the notebooks and run terminals that mimic an actual Linux shell. A full integration with IAM is ongoing.

In the context of ECAS, Jupyter Notebooks represents a very valuable tool for running interactive analysis and visualizations of the results taking advantage of the wide eco-system of Python modules already available for data science, including PyOphidia. To this end, notebooks specifically designed for ECAS users are being developed and frequently updated as training and demonstration material.

4.1.1.4 ECAS Web Portal

The ECAS web portal (see Fig. 3) provides an informative web site with the description of the ECAS service, the quick start guide for using the available JupyterHub instance and the Ophidia server, an introductory video, the full description of some real-world use cases in the climate change domain jointly with the GitHub link to the associated workflow description, and the registration form. It also includes the link to access to the JupyterHub endpoint.

Fig. 4 – Jupyter notebook example

4.2 Software release

ECAS is based on the different components described in the previous architectural section; for this reason in this section we will keep providing a component-wise view, with respect to the release process. ECAS release notes provides a pointer to the Ophidia, PyWPS and JupyterHub releases.

Ophidia is released as open source software (GPLv3). All the components are available on GitHub on the Ophidia project page: https://github.com/OphidiaBigData/. Releases of the components can be found in the single code repositories. For each source code release, RPMs and DEBs packages, respectively for CentOS7 and Debian 14, as well as a ready-to-use VM image, are also created. Additionally, the Ophidia Python bindings (PyOphidia), is released also on Conda Forge (https://anaconda.org/conda-forge/pyophidia) and on the Python Package Index (https://pypi.org/project/PyOphidia/). Releases are announced, usually, as news on the Ophidia website (http://ophidia.cmcc.it/) and as "tweets" on the Ophidia's Twitter page

(https://twitter.com/OphidiaBigData). Finally, a complete administrator and end-user documentation is available for each new release at http://ophidia.cmcc.it/documentation/.

Additional components required for integration that are not existing services or software (e.g. JupyterHub, B2DROP) maintained by ECAS include Docker images and associated documentation. These are released under an Apache Open Source license.

Source code and binary releases are managed through a procedure based on a private (on premise) Jenkins service. A release manager, appointed internally at CMCC, is responsible for updating, implementing and guaranteeing the correct execution of this release procedure. In particular, the code in the "devel" branches of the GitHub repositories undergoes, on a daily-basis, some continuous integration steps in order to:

- check code quality in terms of code style;
- identify, as soon as possible, issues in code building;
- verify the unit tests and the code coverage (only for critical components).

When the code is ready for a new release, in addition to the previous steps, a software rollout procedure is applied. This includes the automated building of the code and creation of the binaries (RPMs and DEBs), as well as the configuration of the full ECAS software stack and the executions of some functional tests on both CentOS7 and Ubuntu14 systems. The whole process exploits a flexible containerized (Docker-based) approach to perform Jenkins automated tasks. Reports are available for each run of the QA suite.

The source code and building of the customized Jupyter Docker image for ECAS@DKRZ are managed through Cloud-based services: GitHub and Docker hub. For each new Docker image, a new tag is locally assigned and then pushed to Github. Next, a built is triggered automatically on Docker hub and made available with the assigned tag.

The next major software release of ECAS is expected at M22 with a possible pre-release at the opening of the service for virtual access at M18. ECAS components such as Ophidia may have a different release cycle. In particular, Ophidia releases are usually provided every 3 months or when key milestones are released or major bugs are fixed.

The ECAS releases include multiple components with their own release cycles. Full ECAS software stack releases are coordinated and agreed by the ECAS team at DKRZ and CMCC, taking project deadlines and planning into account.

4.3 References

- Service endpoints:
    - CMCC: https://ecaslab.cmcc.it/web/home.html
    - DKRZ: https://ecaslab.dkrz.de/home.html
- Software code and releases:
    - ECAS Web site: https://github.com/ECAS-Lab/ecas-web
    - Ophidia platform: https://github.com/OphidiaBigData

- JupyterHub: https://github.com/jupyterhub/jupyterhub
- ECAS (Jupyter) Notebooks: https://github.com/ECAS-Lab/ecas-notebooks
- Documentation, user oriented, administrator oriented and developer oriented:
  - Ophidia (quickstart, user guides, administrator manuals): http://ophidia.cmcc.it/documentation/index.html
  - JupyterHub (installation and usage guides): https://jupyterhub.readthedocs.io/en/stable/
  - The ECASLab web interface with workflow experiments description, quick start guide and registration form: https://ecaslab.cmcc.it/web/home.html

4.4 EOSC-hub integration

The current version of ECAS is integrated in EOSC-hub both at component level and at level of access channels.

1.1.2 Integrated components

There is an additional ECAS component called ECASLab Auth (test instance), which includes the integration of ECAS with the three major EOSC-hub AAI providers: B2ACCESS, Indigo IAM and EGI Check-in. This component is able to authenticate new users to the JupyterHub against the above mentioned AAI providers and it is also part of the testing/release process. Its role consists of:

- Redirecting users from JupyterHub to ECAS Auth Portal.
- Brokering the authentication/authorization.
- Managing users on premise.

1.1.3 Access channels

The latest stable ECAS service is published via EOSC-hub:

- on the Marketplace: https://marketplace.eosc-hub.eu/45-enes;
- on the Service Catalogue: https://eosc-hub.eu/catalogue/ENES%20Climate%20Analytics%20Service;

5 T7.4 GEOSS

The GEOSS (Global Earth Observation System of Systems) services support the implementation of the Sustainable Development Goals (SDGs) defined by the United Nations. Services scope is to help SDG monitoring and assessing by providing the necessary Indicators and Essential Variables (EVs) defined by the Community. The core of the service, GEO DAB (Discovery and Access Broker), will be able to access via open APIs the virtual IaaS and PaaS provided by EGI for the following objectives:

- EGI Federated Cloud: to access computing and storage resources currently not available to end-users, who currently need to procure and manager their own compute infrastructure for data exploitation;
- EGI Data Hub for advanced data management services and "Cloud Workload optimisation".
- EGI Core Infrastructure platform services (e.g. monitoring, AAI, etc.).

5.1 EOSC-hub integration

The GEOSS service integration in EOSC-hub is a work in progress, not finalized at production level yet. Therefore, it is not included in this first software release. The last GEOSS software release is expected at the task's end (M30).

6 T7.5 OPENCoastS

The OPENCoastS service builds on-demand circulation forecast systems for user-selected sections of the North Atlantic coast and maintain them running operationally for the timeframe defined by the user. This service generates daily forecasts of water levels and 2D velocities over the spatial region of interest for periods of 48 hours, based on numerical simulations of the relevant physical processes. Forcing conditions at the boundaries and over the domain are defined by the user from global forecast databases (such as NOAA's GFS, FES2014 and Meteo France predictions). Automatic comparison with real-time in-situ sensor data can be provided for a number of user specified locations, taking advantage of the European Marine Observation and Data Network (EMODNET). Generic operational services, available at the service Web interface, include:

1) extract and download time series at user-defined locations using standard formats (virtual sensors);
2) download specified model outputs using standard formats;
3) visualize model outputs of water levels and velocity;
4) access automatic data/model comparisons at the selected data stations.

The OPENCoastS service is available freely to anyone who plays a role in coastal areas, from coastal authorities to the general public. The OPENCoastS platform is dedicated to all entities with activity on coastal regions across Europe. It targets coastal managers, public institutions, research groups and private companies with responsibilities in emergency and monitoring purposes across Europe. National, regional or local coastal managers from the public and private sector with responsibilities in emergency and monitoring purposes need forecast systems to anticipate hazardous events and prepare emergency response. At the same time, these systems can support planning activities, from daily tasks to strategic interventions. Being able to reproduce the operational behaviour of coastal engineering interventions (even before they are implemented in the coast), the OPENCoastS service is a valuable tool for consultancy companies working in the field of coastal engineering to support engineering projects and their implementation (e.g. study the impact of maritime structure building and dredging interventions in coastal regions). Given the flexibility and generic nature of the OPENCoastS service, research groups extend their limited use in specific sites to broad geographical scope studies of coastal processes, of the climate change and anthropogenic impacts in the coastal zone among other topics. This platform will also facilitate the access to circulation forecasts to research groups with little experience in numerical physical modelling of oceanic and coastal zones such as biologists, geologists and biogeochemists, which have strong needs in understanding the impact of water dynamics in water quality, ecology and sediments dynamics. By making the service available for deployment in any European coastal regions, OPENCoastS leverages the conditions for any entity to develop their responsibilities in a faster, efficient and high accuracy way.

The main service providers' organizations are:

- Laboratório Nacional de Engenharia Civil (LNEC), Portugal
- Infraestrutura Nacional de Computação Distribuída (INCD), Portugal

- Instituto de Física de Cantabria (CSIC/IFCA), Spain

In addition, the hardware resources used by the services are:

- OPENCoastS@INCD:
  - Cloud resources: 8 VMs, 32 VCPUs, 65GB RAM, 2.5TB Storage
  - The services are:
    - 1 haproxy node (to be deployed a second one): exposes https of several services,
    - https://opencoasts.ncg.ingrid.pt - Opencoasts main portal,
    - https://opencoasts.ncg.ingrid.pt:9443 - Minio [24] object store,
    - https://opencoasts.ncg.ingrid.pt:8443/ncWMS - ncWMS [25] (NetCDF files),
    - 1 node: application frontend and backend,
    - 1 Minio object store: this will be substituted in the future by EOSC data management service,
    - 1 node for the ncWMS (NetCDF files),
    - 1 node for the PostGIS [26] service,
    - 1 node for NFS server,
    - fat compute nodes: via direct access to Openstack, EGI Grid computing through DIRAC4EGI or EGI Federated Cloud,
    - Data repository: to be deployed as part of the EUDAT infrastructure.

OPENCoastS@IFCA: To be deployed.

6.1 Service architecture

The OPENCoastS service architecture includes a frontend with a user interaction component for forecast systems configuration and management, via a web application, a backend where models and mapping services run and a storage tier for preservation. The following figure illustrates this architecture with more detail.
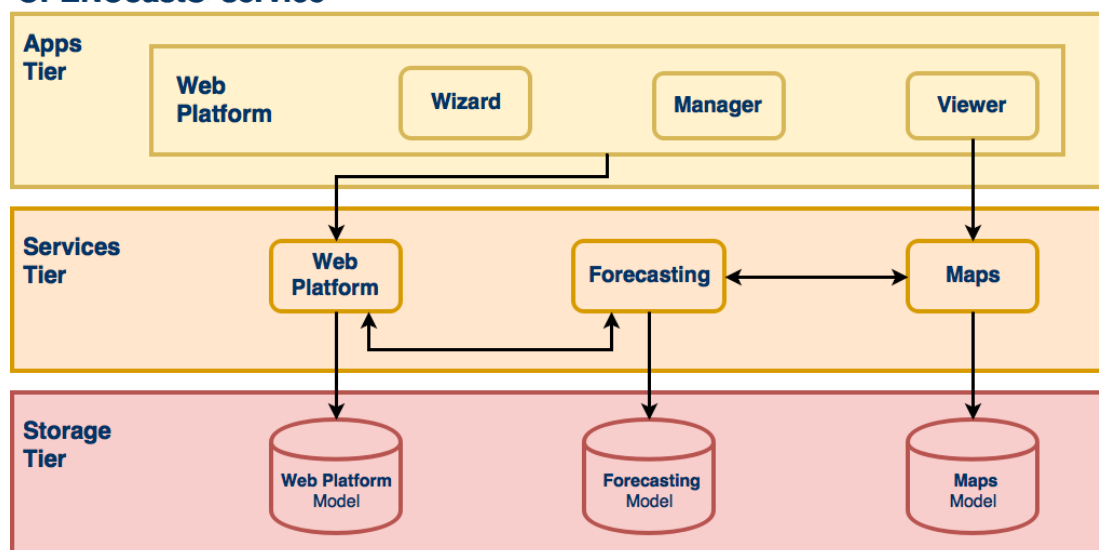
Fig. 5 - OPENCoastS service architecture

6.1.1 Web application

This component provides access to the service, through web pages hosting wizard, manager and viewer applications (apps). Each of these apps allows the users to interact with the different aspects of the service while keep them independent of each other. The Django Web Framework is used as the development basis of this component, which follows Django's design philosophy of having a project composed by applications, each with a set of concerns and functionalities.

6.1.2 Forecasting

While the web application, as the user facing component, allows the users to interact with the service and manage their operational real-time forecast systems, it is the forecasting component that is responsible for producing the forecast results. As a central piece, it interacts with all other components, directly or indirectly, to be able to gather all the necessary information to run simulations and make their results available.

6.1.3 Mapping

The mapping services complement the web application ones, by providing WMS services, which are then consumed by the viewer application. The ncWMS2 server is used to publish the spatial forecasting results on the web, sourced from netCDF files.

6.1.4 Storage

The storage component keeps the state of all services and is a requirement for all of them. The storage technologies range from typical relational databases servers, to lower level shared file systems. The relational database software used is the PostgreSQL, with PostGIS support, storing most structured information about the service. The object storage service, currently provided by Minio, makes available files used in and resulting from the simulations. Shared file systems, in this case NFS, are used to share folders and files among the computation and mapping host resources.

6.2 Software release

OPENCoastS integration was planned considering a set of milestones, starting from the simplest version, which integrates the basic functionalities necessary to configure and deploy a forecast system on demand, to more complex versions that add more choices for users during the configuration process, such as different model settings and forcing sources to apply. These milestones are set by our scientific team members that are simultaneously part of the testers' team and also end users.

OPENCoastS development considers three environments:

- local development environment;
- local production environment;
- official environment.

Newly developed features, along with bug fixing and feature improvements, are always published first to the local development environment where they are fully tested and validated by developers and testers. As soon as the testers feedback is positive these changes are published and tested in the local production environment, that simulates more accurately the official environment; only then, and with the project's manager approval, updates are published to all end users.

6.3 References

Service endpoints:

- https://opencoasts.ncg.ingrid.pt - Opencoasts main portal
- https://opencoasts.ncg.ingrid.pt:9443 - Minio object store
- https://opencoasts.ncg.ingrid.pt:8443/ncWMS - ncWMS (NetCDF files)

Software code and releases: the project's versioning is managed by a Git service, more specifically using Gogs. At the moment, the service is hosted on local servers. However, it is planned to move it to a public server in the future.

The user manual is available at:

- https://opencoasts.ncg.ingrid.pt/static/OPENCoastS_manual.pdf
- http://opencoasts.lnec.pt/OPENCoastS_manual.htm

Additional information for users is available at:

- http://opencoasts.lnec.pt/pdfs/IMUM_AFortunato.pdf
- http://opencoasts.lnec.pt/pdfs/aoliveira_final_imum_hands_on.pdf

Documentation for administrators and developers will be made available when the software code is placed in a public repository.

6.4 EOSC-hub integration

The latest stable version of the service is published via EOSC-hub:

- on the Marketplace: https://marketplace.eosc-hub.eu/44-opencoasts;
- on the Service Catalogue: https://eosc-hub.eu/catalogue/OPENCoastS.

7 T7.6 WeNMR

The WeNMR thematic services consist of a suite of computational tools serving the structural biology community at large. Structural biology studies the functions and interactions of proteins, nucleic acids and other biomolecules using experimental methods such as X-ray crystallography, Nuclear Magnetic Resonance (NMR) or cryo-electron microscopy (cryo-EM). All these methods generate data that needs to be processed, analyzed and finally converted into three dimensional (3D) structures (or models) of biomolecules using a variety of computational tools and techniques.

Gaining access to 3D structures of biomolecules, their dynamics, and their interactions with other molecules is key to a proper understanding of their function. It also allows you, for example, to rationalise the effect of disease-causing mutations, to engineer better molecules for material, health or food applications and to obtain a starting point for drug design to combat disease. As such, structural biology has a strong socio-economical impact on many application fields from health, to food, to materials.

The WeNMR suite is composed of seven individual platforms (refer to section 7.1 for details):

- AMPS-NMR, a web portal for Nuclear Magnetic Resonance (NMR) structures
- CS-ROSETTA, to model the 3D structure of proteins
- DISVIS, to visualise and quantify the accessible interaction space in macromolecular complexes
- FANTEN, for multiple alignment of nucleic acid and protein sequences
- HADDOCK, to model complexes of proteins and other biomolecules
- POWERFIT, for rigid body fitting of atomic structures into cryo-EM density maps
- SPOTON, to identify and classify interfacial residues as Hot-Spots (HS) in protein-protein complexes

7.1 Service architecture

Most of the WeNMR portals have been in operation since several years. They are geographically located in Utrecht in the Netherlands and at the University of Florence in Italy. All portals are web-based, built on a variety of technological solutions (e.g. Python, Flask, Apache, …), but all present a unified and well-recognizable front end to users. They make use of the EGI HTC resources to distribute jobs to the sites supporting the enmr.eu VO using in most cases DIRAC4EGI for job submission, but also gLite in some specific cases, e.g. like the GPGPU-grid enabled DISVIS and POWERFIT portals that are sending jobs to specific CEs in Florence. These two applications make also use of udocker, a basic user tool to execute simple Docker containers in user space without requiring root privileges, developed by the INDIGO-DataCloud project and currently supported by EOSC-Hub.
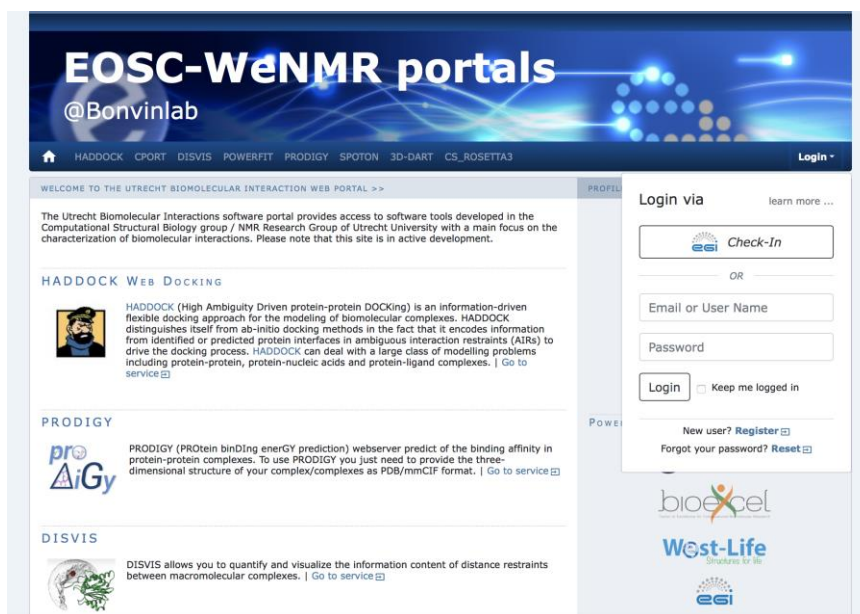
Fig. 6 - Front end of the Utrecht WeNMR portals

7.1.1   Description of the WeNMR portals

7.1.1.1   AMPS-NMR

AMPS-NMR, (AMBER-based Portal Server for NMR structures) is a web interface to set up and run calculations with the AMBER package. The interface allows the refinement of NMR structures of biological macromolecules through restrained Molecular Dynamics (rMD). Some predefined protocols are provided for this purpose, which can be personalized; it is also possible to create an entirely new protocol. AMPS-NMR can handle various restraint types. As an ancillary service, it provides access to a web interface to AnteChamber, enabling the calculation of force field parameters for organic molecules such as ligands in protein–ligand adducts.

The AMPS-NMR grid-enabled web server, originally developed under the WeNMR e-Infrastructure project (www.wenmr.eu) uses resources provided by the EGI (www.egi.eu) and the associated National Grid Initiatives (NGIs).

- Bertini, I., Case, D. A., Ferella, L., Giachetti, A. & Rosato, A. A Grid-enabled web portal for NMR structure refinement with AMBER. *Bioinformatics* **27,** 2384–2390 (2011).

7.1.1.2   CS-Rosetta3

CS-ROSETTA is a protocol which generates 3D models of proteins, using only the 13CA, 13CB, 13C', 15N, 1HA and 1HN NMR chemical shifts as input. Based on these parameters, CS ROSETTA uses a SPARTA-based selection procedure to select a set of fragments from a fragment-library (where the chemical shifts and the 3D structure of the fragments are known). The fragments are assembled using the Rosetta protocol. The generated models are rescored based on the difference between the back-calculated chemical shifts of the generated models and the input chemical shifts, and when available, with a post-scoring procedure based on unassigned NOE lists.

The [CS-Rosetta3 grid-enabled web server](), originally developed under the WeNMR e-Infrastructure project ([www.wenmr.eu]()) uses resources provided by the EGI ([www.egi.eu]()) and the associated National Grid Initiatives (NGIs).

- Wassenaar, T. A. et al. WeNMR: Structural Biology on the Grid. *J Grid Computing* **10**, 743–767 (2012).
- Schot, G. & Bonvin, A. M. J. J. Performance of the WeNMR CS-Rosetta3 web server in CASD-NMR. *Journal of Biomolecular NMR* **62**, 497–502 (2015).

7.1.1.3 DisVis

[DisVis]() allows visualizing and quantifying the information content of distance restraints between macromolecular complexes. It performs a full and systematic 6-dimensional search of the three translational and rotational degrees of freedom to determine the number of complexes consistent with the restraints. In addition, it outputs the percentage of restraints being violated and a density that represents the center-of-mass position of the scanning chain corresponding to the highest number of consistent restraints at every position in space.

- G.C.P. van Zundert, M. Trellet, J. Schaarschmidt, Z. Kurkcuoglu, M. David, M. Verlato, A. Rosato and A.M.J.J. Bonvin. The DisVis and PowerFit web servers: Explorative and Integrative Modeling of Biomolecular Complexes. *J. Mol. Biol.*, **429**, 399-407 (2017).

7.1.1.4 FANTEN

Pseudocontact shifts (PCSs) and residual dipolar couplings (RDCs) arising from the presence of paramagnetic metal ions in proteins as well as RDCs due to partial orientation induced by external orienting media are nowadays routinely measured as a part of the NMR characterization of biologically relevant systems. PCSs and RDCs can be used: 1) to determine and/or refine protein structures in solution, 2) to monitor the extent of conformational heterogeneity in systems composed of rigid domains which can reorient with respect to one another, and 3) to obtain structural information in protein-protein complexes. The use of both PCSs and RDCs proceeds through the determination of the anisotropy tensors which are at the origin of these NMR observables. A new user-friendly web tool, called FANTEN (Finding ANisotropy TENsors), has been developed for the determination of the anisotropy tensors related to PCSs and RDCs and has been made freely available through the WeNMR ([http://fanten-enmr.cerm.unifi.it:8080]()) gateway. The program has many features not available in other existing programs, among which the possibility of a joint analysis of several sets of PCS and RDC data and the possibility to perform rigid body minimizations.

- Rinaldelli, M., Carlon, A., Ravera, E., Parigi, G. & Luchinat, C. FANTEN: a new web-based interface for the analysis of magnetic anisotropy-induced NMR data. *J Biomol NMR* **61**, 21–34 (2014).

7.1.1.5 HADDOCK

HADDOCK2.2 (High Ambiguity Driven protein-protein DOCKing) is an integrative, information-driven flexible docking approach for the modeling of biomolecular complexes. HADDOCK distinguishes itself from ab-initio docking methods in the fact that it encodes information from identified or predicted protein interfaces in ambiguous interaction restraints (AIRs) to drive the docking process. HADDOCK can deal with a large class of modeling problems including protein-protein, protein-nucleic acids and protein-ligand complexes.

The HADDOCK2.2 grid-enabled web server, originally developed under the WeNMR e-Infrastructure project (www.wenmr.eu) uses resources provided by the EGI (www.egi.eu) and the associated National Grid Initiatives (NGIs).

- van Zundert, G. C. P. et al. The HADDOCK2.2 Web Server: User-Friendly Integrative Modeling of Biomolecular Complexes. *J Mol Biol* **428**, 720–725 (2015).

7.1.1.6 PowerFit

[PowerFit](#) automatically fits high-resolution atomic structures into cryo-EM densities. To this end it performs a full-exhaustive 6-dimensional cross-correlation search between the atomic structure and the density. It takes as input an atomic structure in PDB- or mmCIF-format and a cryo-EM density with its resolution and outputs positions and rotations of the atomic structure corresponding to high correlation values. PowerFit uses the local cross-correlation function as its base score. The score is by default enhanced with an optional Laplace pre-filter and a core-weighted version to minimize overlapping densities from neighbouring subunits.

- G.C.P. van Zundert, M. Trellet, J. Schaarschmidt, Z. Kurkcuoglu, M. David, M. Verlato, A. Rosato and A.M.J.J. Bonvin. The DisVis and PowerFit web servers: Explorative and Integrative Modeling of Biomolecular Complexes. *J. Mol. Biol.*, **429**, 399-407 (2017).

7.1.1.7 SpotON

[SpotOn](#) is a robust algorithm developed to identify and classify the interfacial residues as Hot-Spots (HS) and Null-Spots (NS) with a final accuracy of 0.95 and a sensitivity of 0.95 on an independent test set. The predictor was developed using an ensemble learning algorithm with up-sampling of the minor class and was trained on a large number of complexes and on a high number of different structural- and evolutionary sequence-based features.

- I.S. Moreira, P.I. Koukos, R. Melo, J.G. Almeida, A.J. Preto, J. Schaarschmidt, M. Trellet, Z.H. Gümüş, J. Costa and A.M.J.J. Bonvin. SpotOn: High Accuracy Identification of Protein-Protein Interface Hot-Spots. *Sci. Reports*. **7**:8007 (2017).

7.1.2 General architecture of the WeNMR portals

All WeNMR portals are built on the same philosophy of shielding the end user from the complexity of accessing/using HTC (grid or cloud) resources. From a user perspective, a user only interacts with web-based portals, filling in forms and uploading data. Upon successful submission those data are processed through complex workflows calling typically a variety of software and using a combination of both local and EOSC HTC resources. Finally, the results are post-processed and presented to the user in a user-friendly manner, facilitating their interpretation. The general architecture is illustrated in the following figure.
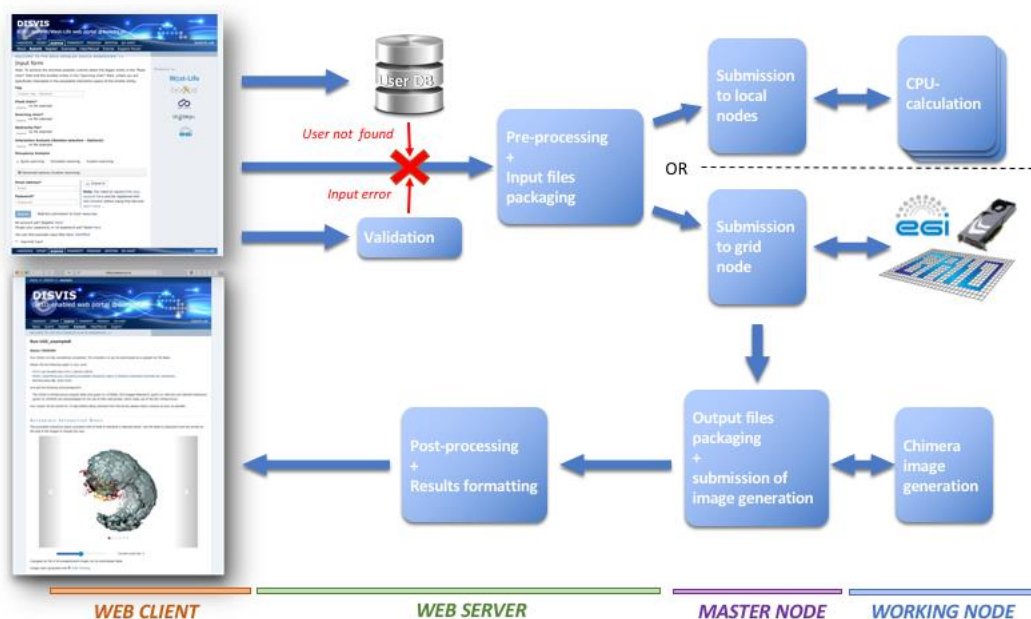
Fig. 7 - Illustration of the general workflow behind the WeNMR web portals.

7.2 Software release

There is no general software release associated with the WeNMR thematic services at this time. In some cases, the software run behind the portal is from third parties and not developed within the WeNMR partners. In other cases, WeNMR partner are the developers of the software behind the portals and those are usually freely available and downloadable directly from GitHub. Other do require a specific license. Note that none of that software has been developed in the context of EOSC-hub (or previous related projects). They are the result of research work financed by a variety of national and international project.

Freely downloadable software:

- DISVIS   https://github.com/haddocking/disvis
  also available as a Docker container from the INDIGO Docker hub:
    o   https://hub.docker.com/r/indigodatacloudapps/disvis/

- POWERFIT   https://github.com/haddocking/powerfit
  also available as a Docker container from the INDIGO Docker hub:
    o   https://hub.docker.com/r/indigodatacloudapps/powerfit/

Software requiring a license (free for non-profit):

- HADDOCK   http://www.bonvinlab.org/software/haddock2.2

Third party software:

- AMBER   http://ambermd.org
- CS-Rosetta   https://www.rosettacommons.org/docs/latest/CS-Rosetta

7.3 References

7.3.1 Portal-related publications:

- Bertini, I., Case, D. A., Ferella, L., Giachetti, A. & Rosato, A. A Grid-enabled web portal for NMR structure refinement with AMBER. *Bioinformatics* **27,** 2384–2390 (2011).
- Wassenaar, T. A. et al. WeNMR: Structural Biology on the Grid. *J Grid Computing* **10**, 743–767 (2012).
- Schot, G. & Bonvin, A. M. J. J. Performance of the WeNMR CS-Rosetta3 web server in CASD-NMR. *Journal of Biomolecular NMR* **62**, 497–502 (2015).
- G.C.P. van Zundert, M. Trellet, J. Schaarschmidt, Z. Kurkcuoglu, M. David, M. Verlato, A. Rosato and A.M.J.J. Bonvin. The DisVis and PowerFit web servers: Explorative and Integrative Modeling of Biomolecular Complexes. *J. Mol. Biol.*, **429**, 399-407 (2017).
- van Zundert, G. C. P. et al. The HADDOCK2.2 Web Server: User-Friendly Integrative Modeling of Biomolecular Complexes. *J Mol Biol* **428**, 720–725 (2015). I.S. Moreira, P.I. Koukos, R. Melo, J.G. Almeida, A.J. Preto, J. Schaarschmidt, M. Trellet, Z.H. Gümüş, J. Costa and A.M.J.J. Bonvin. SpotOn: High Accuracy Identification of Protein-Protein Interface Hot-Spots. *Sci. Reports*. **7**:8007 (2017).

7.3.2 WeNMR web site with services, description, support and tutorials:

- http://www.wenmr.eu

7.3.3 EOSC-Hub related pages:

- https://marketplace.eosc-portal.eu/46-wenmr-suite-for-structural-biology
- https://eosc-hub.eu/eosc-in-practice-wenmr

7.4 EOSC-hub integration

From day 1 of the project, the WeNMR thematic services have been in operation, sending over the first eight months of the project over 5 million jobs to the EOSC HTC resources, most of which through the DIRAC4EGI service. According with the EGI Accounting Portal, these account for over 15 million HS06 CPU Time hours consumed. Although the WeNMR services operate making use of opportunistic computing resource, the access to the federated resources of EGI has been formalized by a Service Level Agreement (SLA) between EGI.eu and the enmr.eu VO (represented by the Faculty of Science – Department of Chemistry of Utrecht University). This SLA was established in 2016 and has been renewed in 2018, granting to enmr.eu VO until 31/12/2020 an amount of opportunistic computing time up to 53 Million of normalized CPU hours and opportunistic storage capacity up to 54 TB [27]. Five resource centres signed this last version of the SLA: INFN-PADOVA (Italy), TW-NCHC (Taiwan), SURFSara and NIKHEF (The Netherlands), NCG-INGRID-PT (Portugal).

During the first nine month of the project, a number of WeNMR portals have been migrated from the old gLite-based job submission to the EOSC-hub **DIRAC4EGI** service. Further, all portals are now

offering Single Sign On (SSO), either through the West-Life SSO which connects to both ARIA (the access management solution of the Structural Biology infrastructure INSTRUCT-ERIC) and the old legacy WeNMR SSO [28], or directly through the **EGI Check-in**. Users can now register and use the WeNMR services using the Check-in, allowing them to use a variety of identity providers. AMPS-NMR is currently using only the West-Life SSO, but is in process of including EGI-Check-in too.

7.4.1   West-Life SSO integration

INFN hosts an instance of the EOSC-hub OneData service, offering up to 10 TB of storage space to WeNMR community. WeNMR users can request a storage space at the Oneprovider service hosted at INFN-Padova data center by connecting to the Onezone server (onezone.cloud.cnaf.infn.it) hosted at INFN-CNAF data center (located in Bologna). INFN developed a new plugin to enable West-Life SSO as authentication method for Onedata.

7.4.2   West-Life Virtual Folder integration with Onedata

Virtual Folder (VF) is a tool developed in the context of West-Life project and now maintained by INSTRUCT-ERIC [29]. Currently integrated in several WeNMR portals, it acts as a gateway for many storage systems, such as Dropbox, EOSC B2Drop and any other system accessible through the WebDAV protocol. Each protocol requires a specific plugin loaded by the engine of the VF framework for retrieving metadata for files and directories of the storage back-end. INFN developed a plugin for integrating VF with Onedata, i.e. to enable Onedata storage system as additional back-end. The current integration of VF with Onedata can only be considered as a prototype. It will become "production ready" only when a final release of Onedata is available.

8 T7.7 EO Pillar

The EO-Pillar service provides access to different services in the field of Earth Observation (EO). The services are categorised into three main classes: data access and computing services, data exploitation services, general user services:

- Geohazards Exploitation Platform (GEP)
  - High-Resolution Change Monitoring for the Alpine Region (TerraDue)
  - EO Services for Earthquake Response and Landslides Analysis (TerraDue)
- EODC JupyterHub for global Copernicus data(EODC)
  - Provides a Jupiter hub service to explore the data provided by EODC
    - http://jupyterhub.eodc.eu/hub/login
- EODC Data Catalogue Service (EODC)
  - EODC's data catalogue service, see
    - https://eomex.eodc.eu/
    - https://csw.eodc.eu/
- Sentinel Hub (Sinergise)
  - OGC compliant WMS, WCS and WMTS access to global archives of Sentinel-1 GRD, Sentinel-2 L1C, Sentinel-2 L2A, Sentinel-3 OLCI, Sentinel-5P, Envisat MERIS, Landsat-8, Landsat-5 (Europe archive), Landsat-7 (Europe archive)
  - Statistical API providing statistical data over long time-series (e.g. min/max/median value over point or polygon)
  - Configuration utility to expose various configurations over WMS/WCS/WMTS
  - Custom scripting for ad-hoc definition of algorithms
  - More info: https://www.sentinel-hub.com/develop/documentation/api
  - OSX-Sentinel Hub (https://sentinel.eosc.grnet.gr)
- MEA Platform (Data access and exploitation service) (MEEO)
  - OGC-standardised discovery (Open Search)
  - OGC-standardised access (WCS 2.0)
- rasdaman EO Datacube (RASDAMAN)
  - standards-based access to Petascale DIAS data, offered as space-time datacubes for flexible extraction, visualization, and analytics/statistics
- CloudFerro Data Collections Catalog (CloudFerro)
  - based on CKAN [30], user friendly web interface for all activities associated with data publication and subscription and EO Browser – specialized catalogue application for robust satellite imagery discovery and quick georeferenced preview
- CloudFerro Infrastructure(CloudFerro)
  - cloud infrastructure adapted to the processing of big amounts of EO data, including an EO Data storage cluster and a dedicated IaaS cloud infrastructure for the platform's Users.
- CloudFerro Data Related Services - EO Finder(CloudFerro)

○ tool, which allows finding data products stored in the repository, obtained or processed at, selected times with selected cloud coverage levels and with other selection criteria.
- CloudFerro Data Related Services - EO Browser (CloudFerro)
  ○ allows visualization and basic processing of selected data collections (like Sentinel-1 L1 GRD or Sentinel-2 L1C)
- EPOSAR Service of EPOS infrastructure (IREA-CNR)
  ○ Systematic generation of displacement maps and time series with Sentinel-1 data

The services have a strong focus on the Earth Observation (EO) community, but aim to reach out to further research disciplines.

List of service providers:

- IREA CNR: http://irea.cnr.it
- CloudFerro: http://cloudferro.com
- EODC: https://www.eodc.eu
- GRNET: https://grnet.gr
- MEEO: http://www.meeo.it
- RASDAMAN: http://www.rasdaman.org
- Sinergise: https://sinergise.com
- Terradue: https://www.terradue.com

The different services provide different resources, i.e. data access and computing services, data exploitation services, general user services.

The resource providers are EODC, GRNET and CloudFerro.

8.1 Service architecture

The different services utilise different architectures, protocols and standards.

8.1.1 Multi-sensor Evolution Analysis (MEA) platform

MEA platform implements the concept of Digital Earth making global environmental geospatial data Findable, Accessible, Interoperable and Reusable (FAIR). MEA platform provides an effective sub-setting functionality that accesses the data only when requested and serves to the client only the data amount that is really needed. It is a cross-domain application that supports various application fields: climate health, land cover, pollution, agriculture, weather, oceanography.

The MEA platform is data exploitation (via GUI) and data discovery and access (via DAS component) services:

- Graphic User interface: https://eodataservice.org. This UI provides access to data available through DAS components distributed across several infrastructures. Three types of user interfaces are provided:
  o web based graphic user interface

        o   Jupyter notebook (python console with possible extension to 40 different programming languages)

        o   APIs: OGC-compliant (OpenSearch, WCS2.x) RESTful interfaces

- Data Access System (DAS), this component enables discovery and access service via OGC standard interfaces (OpenSearch, WCS):
  - http://snf-823526.vm.okeanos.grnet.gr/(access to Regional data(cubes): ground measurements, numerical simulations, ...)
  - https://sentinel.eodataservice.org/(access to Regional and Global data(cubes): Sentinel-1, Sentinel-2, Sentinel-3, Sentinel-5P)
  - https://landsat.eodataservice.org/(access to European data(cube): Landsat-8)
  - http://dar05.eodataservice.org(access to Global thematic data(cube): MODEL Atmospheric Level 2 and Level 3 products, …
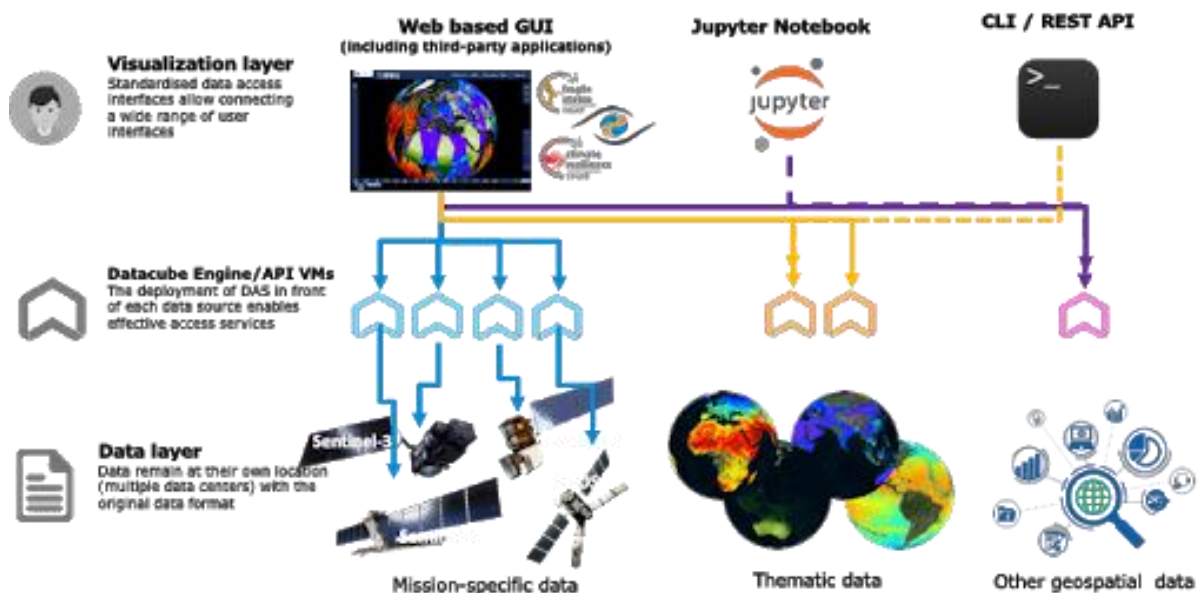

Fig. 8 – MEA platform

8.1.2 Geohazards Exploitation Platform (GEP)

The Geohazards TEP (GEP) is an enhancement of the precursor platforms (G-POD, SSEP), and is designed to support the Geohazard Supersites (GSNL) and the Geohazards community via the CEOS WG Disasters. GEP is an ESA originated R&D activity on the EO ground segment to demonstrate the benefit of new technologies for large scale processing of EO data. Its goal is to apply a complementary operations concept, i.e. moving User activities to the Data: users access a work environment containing the data and resources required, as opposed to downloading and replicating the data 'at home'. The Geohazards TEP platform is a complex system composed by the following main components:

- Web portal
- Data Agency including
  - Catalogue service

    ○ Data Gateway
- Production Center
- Development environment

8.1.3 EODC Data Catalogue Service

The EODC Data Catalogue service allows querying the Copernicus Sentinel satellite data hosted at EODC. The service is available through a simple Web GUI, eomEX+, as well as an API. The back-end of eomEX+ is the EODC pycsw server, an implementation of an OGC CSW server. As a consequence, the eomEX+ API is accompanied by an expert level API provided by the EODC CSW server, located at https://csw.eodc.eu. Further details can be found here: https://eomex.eodc.eu/manual.

8.1.4 EODC JupyterHub for global Copernicus data

This service is based on an implementation of JupyterHub at EODC. The service functions as a starting point to get free access to Copernicus Sentinel satellite data hosted at EODC. In addition, the service facilitates the development and realizations of algorithms. Functions and algorithms can be executed directly on the data by utilizing well-known environments such as Jupyter notebooks and a simple terminal interface. It can be accessed via https://jupyterhub.eodc.eu/hub/login. Support is provided via support@eodc.eu.

8.1.5 Sentinel Hub: Real-time processing of EO data

Sentinel Hub OGC API is optimized for on-demand on the fly processing of raw (unchanged) EO data. The following steps are typically performed within a standard request:

- query Catalogue for chosen AOI, time range, cloud coverage, mission, etc.
- download necessary data from on-line storage
- decompress data
- application of pre-mosaic filters (e.g. DOS-1, statistical atmospheric correction, etc.)
- creation of mosaic based on priority (e.g. most recent data on top), cloud replacement, etc.
- compositing relevant bands to chosen EO product (true color, false color, NDVI, etc.) using chosen style (greyscale, RGB, red temperature, etc.)
- application of post-mosaic filters (color balance, HDR, midtone, gamma correction)
- re-projection to chosen CRS (e.g. Popular Web Mercator, WGS 84, national CRS systems).
- output creation in chosen file type (JPG, PNG, JP2, JPX, GeoTiff, etc.)
- compression of the output for faster download, based on user settings

Note that user can choose based on the parameters, in case some steps should be performed or not. A typical scenario described above takes about 1-2 second for an area of interest with a size of 1000*1000 px (depending on the chosen scale this may represent either 100 sq.km up to 25.000 sq.km).

8.1.6 Rasdaman datacube

The rasdaman datacube engine is a multi-parallel, federated array database system optimized for flexibility, performance, and horizontal/vertical scalability. Its interfaces consist of easy-to-use

OGC WMS and WCS APIs plus high-level, declarative, standardized query languages (OGC WCPS and ISO Array SQL) allowing any query, any time, on any size.

8.1.7   Sentinel Data Hub

GRNET offers a dedicated instance for the Sentinel Data hub (OSX – Sentinel, https://sentinel.eosc.grnet.gr) that retrieves Sentinel Products from the National Sentinel Data Mirror Site (https://sentinels.space.noa.gr/) which is a web based system designed to provide EO data users with Search - Cataloguing - Order and Dissemination capabilities for the Sentinel products. Furthermore, GRNET offers upon request cloud resources (VMs & Storage) from its primary IAAS facility ~Okeanos. GRNET is currently working to integrate OSX-Sentinel with the EOSC-hub AAI ecosystem via EGI Check-in. This will allow users to have seamless access to the service using their credential from their home institution.

8.1.8   CloudFerro Data Collections Catalog (CloudFerro)

The Catalogue is based on CKAN open source software, which is widely used for open data publications like e.g. European Data Portal, data.org.uk or danepubliczne.gov.pl. CKAN provides a user-friendly web interface for all activities associated with data publication and subscription. It is capable of advanced data management. All datasets are organized and described with metadata, which allows it to be easily discoverable, with the use of search phrases and customizable filters (e.g.: tags, categories, data formats). It is possible to publish one dataset in different data formats, not only as downloadable files but also as links to web service, web API or links to external WWW resources. Datasets can be stored along with version history and dataset statistics, which allows to monitor the interest in datasets. CKAN also provides functionalities for collaboration, community participation and providing feedback, such as comments, ratings and sharing. It is highly customizable in both terms of Look&Feel and functionalities. Besides, it provides very rich RESTful JSON API, which allows other applications to discover and access the datasets, and it can be integrated easily with Semantic Web technologies such as RDF data model and SPARQL.

8.1.9   CloudFerro Infrastructure (CloudFerro)

CloudFerro infrastructure covers full set of virtual resources available in the solution:
* VM – Virtual Machines (or virtual computing servers) with several operating systems available (both free like CentOS, Ubuntu, Debian, Scientific Linux, and commercial like RedHat, SUSE, Microsoft Windows Server),
* virtual storage volumes that can be easily mounted to the VMs together with object storage solution,
* virtual networks,
* virtual appliances like firewalls (FWaaS) and VPN concentrators (VPNaaS),
* physical servers (baremetal) that can be integrated to the virtual world,
* Single Server VMs – full physical server with a single VM and very fast pass-through NVMe storage – a combination of advantages of a dedicated server and a cloud VM (high capacity, storage speed, no noisy neighbour problem).

8.1.10 CloudFerro Data Related Services - EO Finder (CloudFerro)

Catalog and a search engine dedicated to Earth Observation products. It is main purpose it to handle EO satellite imagery but it can be used to store any kind of geospatialized data.

EO Finder search API is compliant with the CEOS OpenSearch Best Practice Document and is mentioned in ESA's "Exploitation Platform Common Core Components" as the closest implementation of a catalogue component according to the requirements specified in ESA's "Exploitation Platform Open Architecture". It provides API-s in OpenSearch [31], GEOJson [32] and ATOM standards allowing for easy data search. The interface exposes the operational metadata catalogue.

8.1.11 CloudFerro Data Related Services - EO Browser

https://github.com/sentinel-hub/EOBrowser

8.1.12 EPOSAR service

EPOSAR is one of the services of the EPOS infrastructure (www.epos-ip.org). EPOSA, based on the Small BAseline Subset (SBAS) DInSAR technique, is targeted to generate Earth surface displacement maps and time series with sub-centimetric accuracy by exploiting Sentinel-1 (S1) data of the Copernicus Programme. The service implements the whole processing chain, from SAR data retrieval until the generation of geocoded products. It is able to efficiently manage very large SAR datasets (hundreds or thousands of acquisitions) in very large computing environments (it has been used with 280 AWS instances). The products are generated in automatic and systematic way over several selected areas of the Earth and are continuously updated when new S1 acquisitions are available. The achieved products can be discovered, visualized and downloaded by users through the web interface of the EPOS infrastructure.

8.2 Software release

The different services utilise different release mechanisms. No common procedure is defined for the different services. We can expect individual roadmaps. This decision is performed by the individual provider.

8.3 References

- https://geohazards-tep.eo.esa.int/#!
- https://geohazards-tep.eo.esa.int/geobrowser/?id=eoschub-alpsmonitoring-app
- https://geohazards-tep.eo.esa.int/geobrowser/?id=eoschub-landslide-app
- https://jupyterhub.eodc.eu
- https://csw.eodc.eu/
- http://sentinel-hub.com/
- https://eodataservice.org
- https://discovery.creodias.eu/dataset
- https://creodias.eu
- https://finder.creodias.eu/www

- https://browser.creodias.eu
- https://github.com/cloudferro
- https://github.com/sentinel-hub/EOBrowser
- rasdaman: geo-services guide; WCS, WCPS and WMS tutorial; webinars
- https://sentinel.eosc.grnet.gr

8.4 EOSC-hub integration

Everything should be ready for the inclusion of the first EO-Pillar services in the catalogue and the marketplace. Currently, the final steps are performed and they should be finalized by the end of the year.

9 T7.8 DARIAH

The DARIAH (Digital Research Infrastructure for the Arts and Humanities) Thematic Service (TS) aims to enhance and improve the usage of the cloud-based services and technologies in the domain of the digital arts and humanities research. It will enable end-users coming from the digital arts and humanities domains to seamlessly store, describe (metadata) and share their datasets, discover, browse and reuse datasets shared by the others and to perform analysis on various data volumes.

The DARIAH TS is providing the following services

- DARIAH Science Gateway,
- Invenio-based repository in the cloud,
- DARIAH repository (based on CDSTAR).

The **DARIAH Science Gateway** is a web-oriented portal, developed during the EGI-Engage project (DARIAH Competence Centre) and is especially tailored for the researchers coming from digital arts and humanities disciplines. It currently offers several cloud-based services and applications: Semantic and Parallel Semantic Search Engines, DBO@Cloud, Workflow Development and supports several file transfers protocols.

The **Invenio-based repository** is a service that enables researchers and scholars to easily create, deploy and configure their own Invenio-based repository instance and host it on the cloud infrastructure (Federated Cloud). The service aims to a smaller research groups lacking in adequate technical support and budget to acquire their own infrastructure for hosting data repositories.

DARIAH repository is a new service based on the *C*ommon *D*ata *St*orage *AR*chitecture (CDSTAR), a system for storing and searching objects in research projects. The repository is operated by DARIAH-DE as a digital long-term archive for human and cultural-scientific research data. The DARIAH repository is a central component of the DARIAH-DE Research Data Federation Infrastructure, which aggregates various services and applications and can be used comfortably by DARIAH users.

The DARIAH TS address the needs of the researchers and scholars from the DARIAH (Digital Research Infrastructure for the Arts and Humanities) community but also the digital arts and humanities communities in general.

The services are provided by the following institutes:
- **Rudjer Boskovic Institute** (RBI), Zagreb, Croatia - DARIAH Science Gateway, Invenio-based repository
- **Hungarian Academy of Sciences Institute for Computer Science and Control** (MTA SZTAKI), Budapest, Hungary - DARIAH Science Gateway
- **Gesellschaft für wissenschaftliche Datenverarbeitung mbH Göttingen** (GWDG), Göttingen, Germany - DARIAH repository

The resources for the DARIAH TS services are provided by the DARIAH Virtual organization (vo.dariah.eu) and GWDG. Currently, the DARIAH VO has an SLA with the INFN BARI cloud site which provides the following resources for DARIAH Science Gateway and Invenio-based repository: 24 vCPUs, 1 TB storage, 40 GB memory and 3 public IP addresses.

9.1 Service architecture

9.1.1 DARIAH Science Gateway

DARIAH Science Gateway is based on the WS-PGRADE/gUSE, an open source science gateway platform. The gateway consists of a set of portlets that provide specific functionality or services integrated into the portal. The gateway currently provides 5 services: DBO@Cloud, Semantic Search Engine, Parallel Semantic Search Engine, Simple Cloud Access and workflow management. Each of the services is integrated as a portlet and is based on different underlying frameworks:
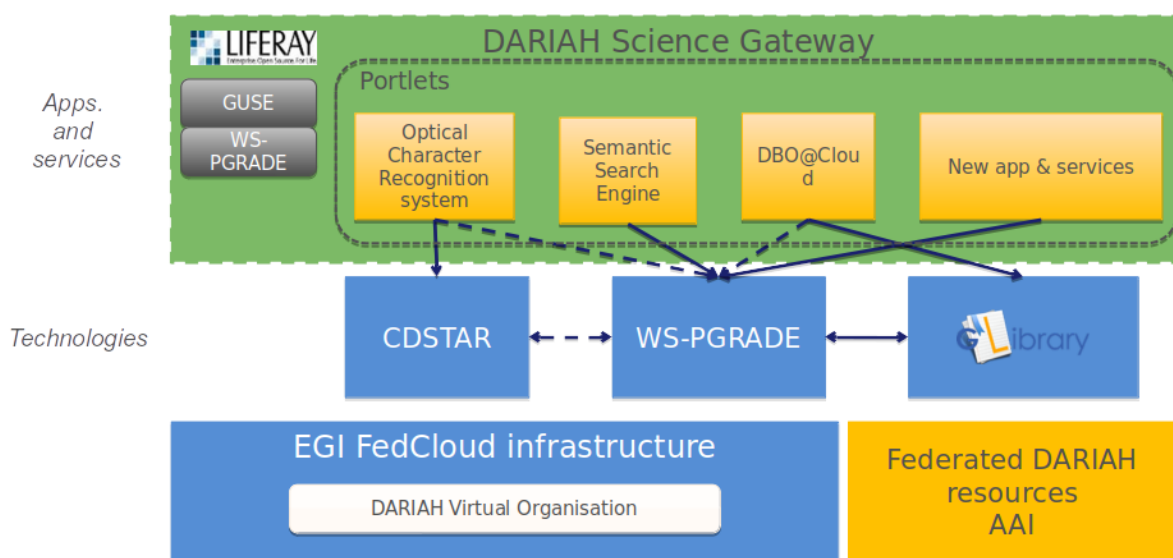

Fig. 9 – DARIAH Architecture

WS-PGRADE, gLibrary and CDSTAR.

9.1.2 Invenio-based repository

The Invenio-based repository services main functionality is to provide an easy deployment of an invenio-based repository in the Cloud. The user interface is based on FutureGateway, which provides functionality such as basic configuration of the repository, e.g. required resources in terms of the number of CPUs and storage size and fine-tuned (optionally) configuration of the Mesos cluster size and per-sub-service resource allocation. The user then starts process of the repository creation in the cloud. The user request is submitted as a TOSCA template to the PaaS Orchestrator that coordinates resource provisioning and virtual infrastructure deployment. Once deployed virtual machines are automatically configured running dedicated Ansible roles shared on Ansible Galaxy. The Invenio components are installed via Ansible and executed as long-running services on top of a Mesos cluster through its framework, Marathon, to ensure fault tolerance and

scalability. The IP address of the newly created repository instance is returned to the user via
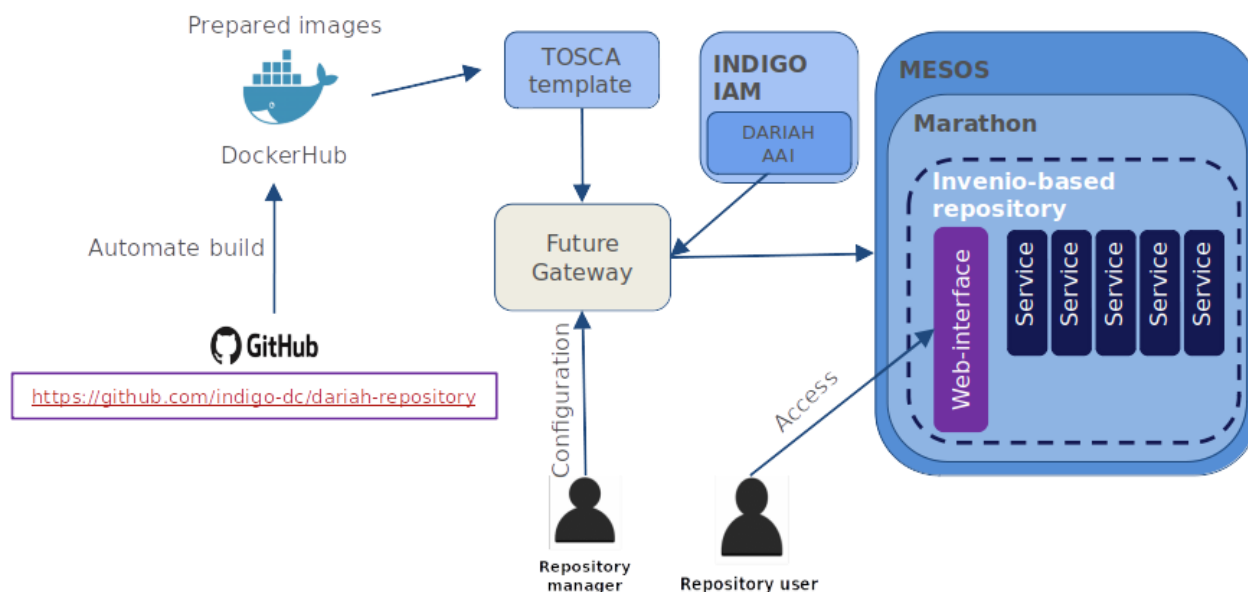

Fig. 10 - : Invenio-based repository in the cloud

FutureGateway. Afterwards, the end-user can access the repository using the provided IP.

9.1.3   DARIAH repository

DARIAH-DE repository is a digital long-term archive for human and cultural-scientific research data. The repository is a central component of the DARIAH-DE Research Data Federation Infrastructure, which aggregates various services and applications and can be used comfortably by DARIAH users. The repository allows users to save their research data in a sustainable and secure way, to provide it with metadata and to publish it. The collections as well as each individual file is available in the DARIAH-DE Repository in the long term and gets a unique and permanently valid Digital Object Identifier (DOI) with which the data can be permanently referenced and cited. In addition, one can register his collections within the Collection Registry, which are then also found in the Generic Search. The overall architecture of the DARIAH-DE repository is depicted in figure 10. An in-depth technical description can be found here: https://repository.de.dariah.eu/doc/services .
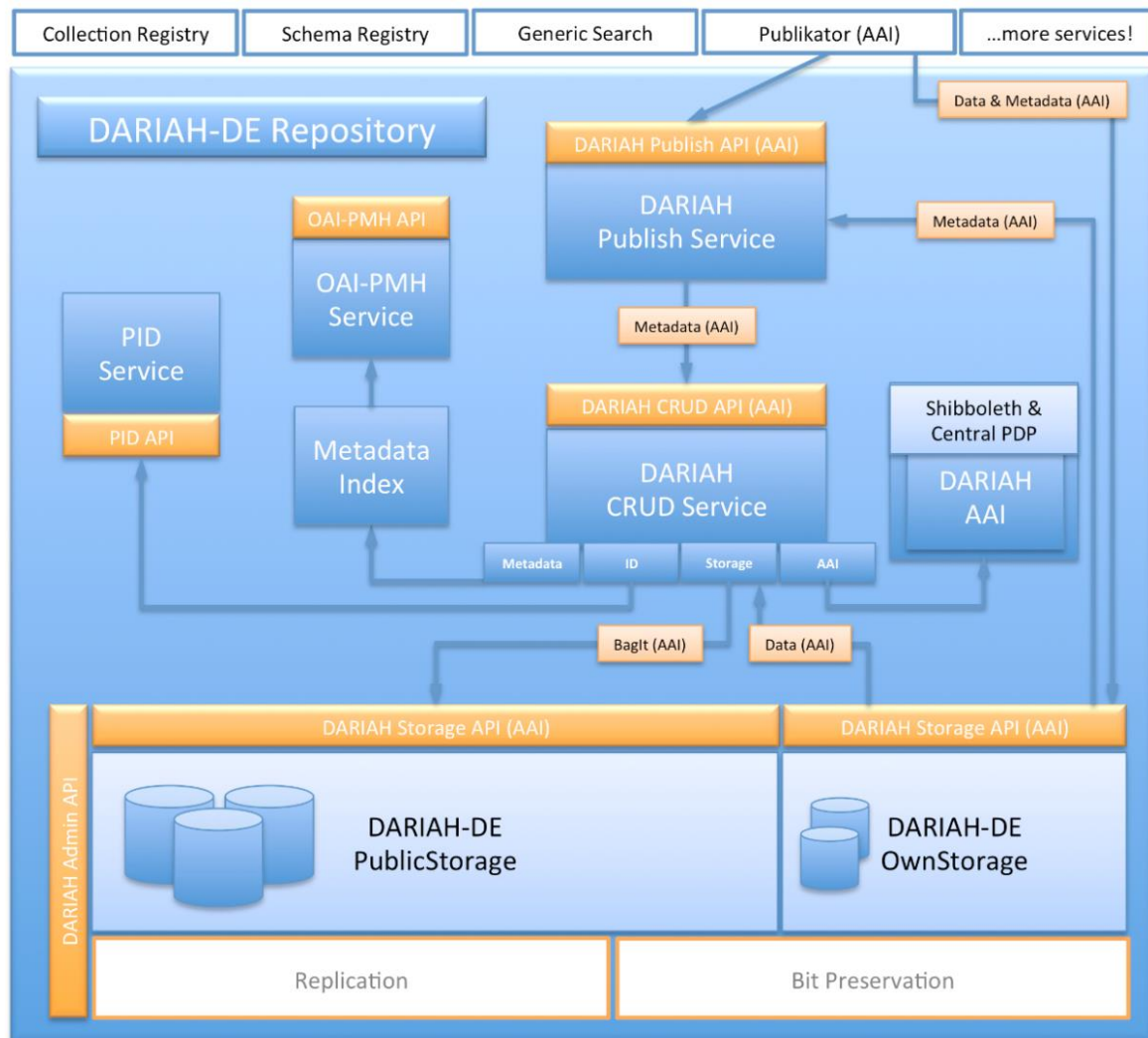
Fig. 11 - Overall scheme of the DARIAH repository

9.2 Software release

The current EOSC-hub thematic software release includes the DARIAH Science Gateway, while the Invenio-based repository and DARIAH repository will be part of the second EOSC-hub software release.

Currently, there is no a strict concept of releases in DARIAH TS and each service has its own release policy, update and versioning policy. The DARIAH Science Gateway updates are connected with new releases of the gUSE/WS-PGRADE system, though, the Gateway does not have any code versions. The code changes are verified manually on local machines before applying them into the production gateway instance. The Invenio-based service (not Invenio itself) does not have versioning.

Invenio-based repository also does not have any code versioning, although with a new Invenio framework release, the repository service can be updated (based on the users' request). In this particular case, the new software release does not affect the already running repository instances,

only the new ones that will be launched based on the new Invenio release. In that case, the deployment process has to be validated and tested manually, to test for the correctness of the deploying particular Invenio sub-components and its inter-dependencies.

The DARIAH repository is developed, tested, maintained and released based on the release-management processes defined by DARIAH-DE. Technically, the CI-system is based on Jenkins and Puppet. The complete process is documented in the DARIAH-DE Wiki.

9.3 References

- Service endpoints:
  - DARIAH Science Gateway - **https://dariah-gateway.lpds.sztaki.hu/** (a new web address will be provided soon)
  - Invenio-based repository - under deployment
  - DARIAH repository - https://repository.de.dariah.eu/publikator

- Software code and releases:
  - **DARIAH Science Gateway:** guse/WSPGrade portal:
    https://sourceforge.net/projects/guse/
  - **Invenio-based repository:**
    - Invenio-based repository source code: https://github.com/indigo-dc/dariah repository
    - Ansible role to start Dariah Repository containers on Marathon:
      https://github.com/indigo-dc/ansible-role-dariahrepo
  - **GWDG DARIAH repository:**
    - CDSTAR: https://github.com/gwdg/cdstar

- Documentation, user oriented, administrator oriented and developer oriented:
  - **DARIAH Science Gateway:**
    - WSPGRADE usage and examples for end-users:
      https://sourceforge.net/projects/guse/files/WS-PGRADE-Cookbook.pdf/download
    - Documentation for developers:
      https://sourceforge.net/projects/guse/files/Cookbook_for_Science_Gateway_Developers.pdf/download
  - **Invenio-based repository in the cloud: l**ocal repository installation for administrators/developers: https://github.com/indigo-dc/dariah-repository/blob/master/INSTALL.rst
  - **GWDG DARIAH repository:**
    - CDSTAR (end-user documentation):
      https://info.gwdg.de/docs/doku.php?id=en:services:storage_services:gwdg_cdstar:start
    - DARIAH repository (architecture):
      https://repository.de.dariah.eu/doc/services

■ DARIAH repository (end-user documentation):
https://wiki.de.dariah.eu/display/publicde/Das+DARIAH-DE+Repository

9.4 EOSC-hub integration

The current EOSC-hub thematic software release includes the DARIAH Science Gateway, while the Invenio-based repository and DARIAH repository will be part of the second EOSC-hub software release. In fact, the current version of the DARIAH Science Gateway is integrated with the EOSC-hub at resource level and access channels level.

9.1.4 Integrated resources

The DARIAH Science Gateway relies on the resources provided by the EGI FedCloud.

9.1.5 Access channels

The latest stable Science Gateway service is published via EOSC-hub:

- on the Marketplace: https://marketplace.eosc-hub.eu/49-dariah
- on the Service Catalogue: https://eosc-hub.eu/catalogue/DARIAH%20Science%20Gateway

10 T7.9 LifeWatch

The LifeWatch ERIC services are:

- PAIRQURS (Life+ RESPIRA),
- Citizen Science Services,
- GBIF data access under biogeographic context,
- Digital Knowledge Preservation Framework,
- Remote Monitoring and Smart Sensing.

The majority of the services are already linked to EGI, EUDAT or INDIGO initiatives. They will be extended to other scientific domains during the project, like for example Genetics or Earth Observation, but also on generic techniques, and in particular analytics, like support on HPC resources for R and Python. Moreover they will rely on the EGI Federated Cloud to access computing and storage resources currently not available to end-users, who currently need to procure and manager their own compute infrastructure for data exploitation.

10.1 EOSC-hub integration

The LifeWatch service integration in EOSC-hub is a work in progress, not finalized at production level yet. Therefore it is not included in this first software release. Although the initial version of the services, developed previously within the EGI-Engage project are published through the Marketplace: https://marketplace.eosc-hub.eu/48-lifewatch-eric-plants-identification-app

11 References

No	Description/Link
1	https://www.clarin.eu
2	https://www.mpcdf.mpg.de
3	https://wicket.apache.org
4	http://lucene.apache.org/solr
5	https://www.openarchives.org/pmh
6	https://www.clarin.eu/content/component-metadata
7	http://dublincore.org
8	http://www.language-archives.org
9	https://github.com/clarin-eric/VLO/milestones
10	https://semver.org
11	https://research.cs.wisc.edu/htcondor
12	https://www.openstack.org/
13	https://aws.amazon.com
14	https://www.hnscicloud.eu
15	https://wiki.oasis-open.org/security/FrontPage
16	https://openid.net/connect
17	https://tools.ietf.org/html/rfc4158
18	https://www.idem.garr.it/en/
19	https://edugain.org/
20	http://emso.eu/
21	http://www.lbto.org/
22	http://pywps.org/
23	https://github.com/jupyterhub/jupyterhub
24	https://www.minio.io/
25	https://reading-escience-centre.github.io/ncwms/
26	https://postgis.net/
27	Updated enmr.eu VO SLA and OLAs, https://documents.egi.eu/document/2751
28	WeNMR Thematic Service AAI report, EOSC-Hub Week, Malaga, 16-20 April 2018: https://indico.egi.eu/indico/event/3903/session/32/material/0/4.pdf
29	https://instruct-eric.eu/

30	https://ckan.org/
31	http://www.opensearch.org
32	http://geojson.org/