HISEA DELIVERABLE 3.7

INTEGRATION OF EXTERNAL DATA SOURCES

WORK PACKAGE NUMBER: 3

WORK PACKAGE TITLE: SERVICE SPECIFICATIONS AND USER REQUIREMENTS

HISea Project Information	
Project full title	High-Resolution Copernicus-Based Information Services at Sea for Ports and Aquaculture
Project acronym	HiSea
Grant agreement number	821934
Project coordinator	Dr. Ghada El Serafy
Project start date and duration	1st January 2019, 30 months
Project website	https://hiseaproject.com/

Deliverable Information	
Work package number	3
Work package title	Service Specifications And User Requirements
Deliverable number	3.7
Deliverable title	Integration of External Data Sources

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 821934

Description	The HiSea operational system makes use of these different "data sources". These data sources primarily include Copernicus data sources and other external data that is being collected locally or remotely or external model results. In order for the platform to be able to properly "talk" with these different data sources specific plugins for each data will be set up. A relevant and innovative aspect of the HiSea is its capability to import any type of data and use the different data sources to automatically build what is commonly called the "best available information". This aspect contributes to increase the service reliability and resilience once if for some unexpected circumstance a data source is not available, the system is capable to use an alternative source. Usually the systems have a rank of data source priority and it always tries to use the data from the highest rank. In case of unavailability, it tries the next one and so on.
Lead beneficiary	Hidromod
Lead Author(s)	Pedro Galvão
Contributor(s)	Sandra Gaytan Aguillar
Revision number	3
Revision Date	30/10/2019
Status (Final (F), Draft (D), Revised Draft (RV))	F
Dissemination level (Public (PU), Restricted to other program participants (PP), Restricted to a group specified by the consortium (RE), Confidential for consortium members only (CO))	PU

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 821934

Document History			
Revision	**Date**	**Modification**	**Author**
0.1	24/10/2019	Text updating	Sandra Gaytan
0.2	25/10/2019	Text updating	Anna Spinosa

Approvals				
	Name	**Organisation**	**Date**	**Signature (initials)**
Coordinator	Ghada El Serafy	Deltares		GES
WP Leaders	Adelio Silva	Hidromod		AS

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 821934

Table of contents and figures

Contents

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 821934

Figures

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 821934

1 Executive Summary

The core functionality of the HiSea operational system is its capacity to aggregate data from different sources with different processing levels (raw data, derived data, etc.) and combine them to provide useful information for the end-users.

The HiSea platform chain starts with the acquisition of data into the system, either by a copy process or by means of real-time calls into external data repositories. The collected data are then converted on-the-fly to known formats that can be later processed by other components of the platform.

The HiSea platform integrates data from external sources such as Copernicus data with data provided by local modeles, and operational systems such as FEWS, SAFI or AQUASAFE.

For the platform to be able to manage these different data sources properly, specific plugins for each data type will be set up. A relevant and innovative aspect of the HiSea platform is its capability to import different types of data and use them to automatically build an ensemble of the "best available information". This aspect contributes to increase the service reliability and resilience; and guarantees the functionality of the platform As an example, if for some unexpected circumstance, a data source is not available, the system is capable to use alternative sources.

Each data provider will have a rank of data source priority, and always tries to use the data from the highest level. In case of unavailability, it attempts the next one and so on.

This document focuses on strategies to access Copernicus data via the Data and Information Access Services (DIAS) platform. It also outlines the approach for other data sources, and how both can be combined.

2 Accessing Copernicus data

To facilitate and standardize access to data, the European Commission has deployed five cloud-based platforms providing centralized access to Copernicus data and information, as well as processing tools. These platforms are known asData and Information Access Services, or DIAS.

DIAS is an excellent fit for hosting the HiSea services since it would allow access to Copernicus data without the burden of maintaining data on the platform. Most of the DIAS platforms provide access to information

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 821934

either via a network shared to a virtual machine (VM) running on their infrastructure or via data services hosted by the DIAS providers.. More information and detailed comparison have been addressed in D3.1 "State of the Art".

At this stage, as a starting testing exercise, CREODIAS has been selected among the available DIAS. The latest, due to the possibility of easily setting up an account with free credit in CREODIAS when compared to the other providers. However, most of the technologies here described, can be implemented to other DIAS providers.

2.1 Data access in CREODIAS

2.1.1 Object Data Access API (SWIFT/S3)

In CREODIAS data can be accessed via SWIFT or S3 API. Both SWIFT and S3 refer to implementations of cloud storage which usually consists of object storage[1] provided through a web service interface.

S3 and Swift are the most commonly used cloud object protocols. Amazon develops the S3 protocol and is available free for all third-party developers. The OpenStack Foundation manages the SWIFT protocol.

The OpenStack Foundation is a non-profit corporate entity established in September 2012 to promote OpenStack software and its community.

CREODIAS recommends Object Data Access API as the primary access method to the stored data (EOdata). Tenants (virtual machines in the CREODIAS infrastructure) can access Eart Observation (EO) data by sending HTTP requests to a standard SWIFT or S3 object interface.

Third-Party Services can use the Object API, but it is also the generic data access method used by other DIAS data access services. It may also be used by registered Users to download data directly to their workstation.

A test on a virtual machine running in CREODIAS infrastructure has been conduced and cmds3 has been successfully used to copy data via the S3 protocol from EOdata (Figure 1).

[1] Object storage is a computer data storage architecture that manages data as objects, as opposed to other storage architectures like file systems.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 821934

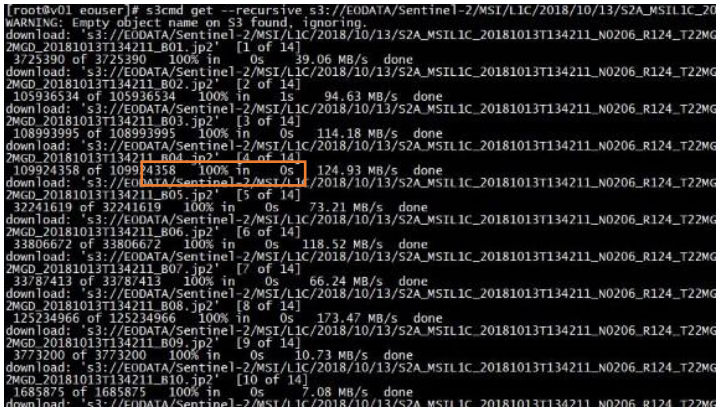

Figure 1- Using s3cmd to copy data to a virtual machine running in CREODIAS

2.1.2 Filesystem Interface

The file system interface allows access to EO data through a "regular file system." Data is materialized in what seams to be a regular folder in virtual machines running on the CREODIAS infrastructure. This approach is useful since most applications (ex: Sentinel Toolboxes) were designed to access data using a traditional POSIX filesystem interface (Figure 2).

Filesystem access may be achieved either through an emulation software that allows a Linux VM to 'mount' an object filestore as a filesystem or an NFS proxy. In both cases, the mounted repository is a read-only file tree mounted under the machine's filesystem.

CREODIAS does not recommend the usage of the filesystem interface. Whenever possible, direct object access (SWIFT/S3) is the preferred access method. Both emulation and NFS proxy have issues of performance and throughput of data access.

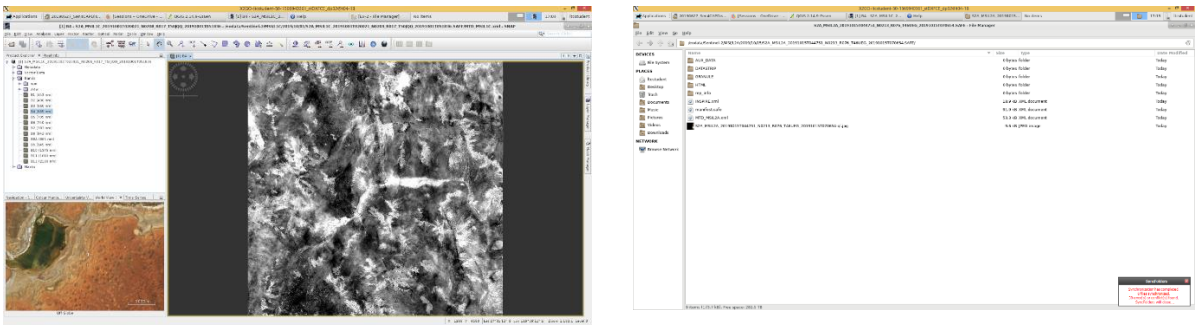

Figure 2 - Using SNAP over EOData in a virtual machine running in CREODIAS

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 821934

Some tests were conducted to test this sort of access with mixed results. Using SNAP over the mapped drive was successful, but listing directories and copying files via command-line resulted in errors or long execution times.

2.1.3 EO Data Processing/Access HUB

Eo Data processing and access hub is a set of OGC[2] compatible services hosted by CREODIAS. The Access Hub is focused on Copernicus satellites but also supporting other sources such as Land sat, PlanetLabs, and others. Image data from the EO Data Repository is processed in real-time using the CREODIAS cloud infrastructure and presented to the users in the form of customizable WMS/WMTS/WCS/WFS web services.

The Basic EO Data Access Hub services are free of charge. These include

- WFS access to catalogue data
- WMS, not for integration in web applications.
- EO Browser
- Mosaic Generator

More advanced functions are available via subscription with specific pricing. However, this sort of access seams to limiting for the goals of the HiSea platform.

3 Accessing other data sources

The main problem of accessing external sources is the diversity of transport protocols and file formats that exist. OGC standards such as sensor observation service (SOS) and Web coverage service (WCF) are still only partially adopted by most data providers.

Most gridded data are accessible via OPENDAP[3] protocol; however, other formats such as time-series and even satellite images are usually more scattered regarding access protocols.

[2] *The Open Geospatial Consortium (OGC) is a consensus standards organization, originated in 1994. In the OGC, more than 500 commercial, governmental, nonprofit and research organizations worldwide collaborate in a consensus process encouraging development and implementation of open standards for geospatial content and services, sensor web and Internet of Things, GIS data processing and data sharing.*

[3] *OPeNDAP is an acronym for "Open-source Project for a Network Data Access Protocol," an endeavor focused on enhancing the retrieval of remote, structured data through a Web-based architecture and a discipline-neutral Data Access Protocol (DAP). Widely used, especially in Earth science, the protocol is layered on HTTP, and its current specification is DAP4,[1] though the previous DAP2 version remains broadly used*

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 821934

For instance, the Global Sea Level Observing System (GLOSS), has a global set of 290 tide gauges, most of them with close to real-time data. However, even though other access channels have been tested, the most reliable is the direct access to an HTML page for each sensor.

The global forecast system GFS[4] is accessible via an OPENDAP server, FTP access, or cloud storage via Amazon. It is the same data but with three different data access protocols.

4 How the HiSea platform serves data

The HiSea platform supplies data via two services (*HiSea Data Service*):

- Time Series service – Comparable to OGC SOS, it provides time-series data from fixed or moving sensors. Only one coordinate in space sampled at any given time.
- Grid Data service – Comparable to OGC WCS, serves up to 4D data, multiple points in space sampled at the same time.

Other services in the platform (model, calculations, etc.) call on *HiSea Data Service*s to obtain input data. Once completed, data is sent back into local storage so that HiSea Data Services can again serve it (Figure 3).

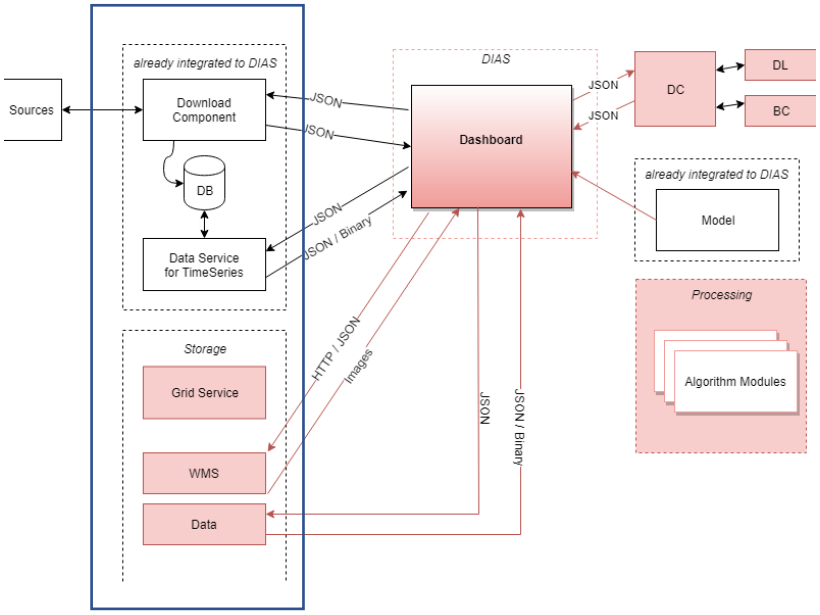

Figure 3- HiSea Data Services

[4] *GFS is a global numerical weather prediction system containing a global computer model and variational analysis run by the United States' National Weather Service (NWS).*

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 821934

These services are set up to run directly over local storage or act as a proxy to data stored in other services. When working as a proxy to external data sources is not possible (due to high latency, security, etc.), the HiSea download service is responsible for periodically downloading data into local storage accessed by the Data Services.

5 Connecting HiSea Data services to EoData

As described in previous deliverables, D4.1: HiSea backend architecture guidelines, the HiSea platform uses a Microservice architecture. Two of the core components were described in the section above, *How the HiSea platform serves data*.

A plugin architecture is the selected pattern for the design of both services. This option allows swapping the local file system with a proxy for the remote data source, without changing the service definition.

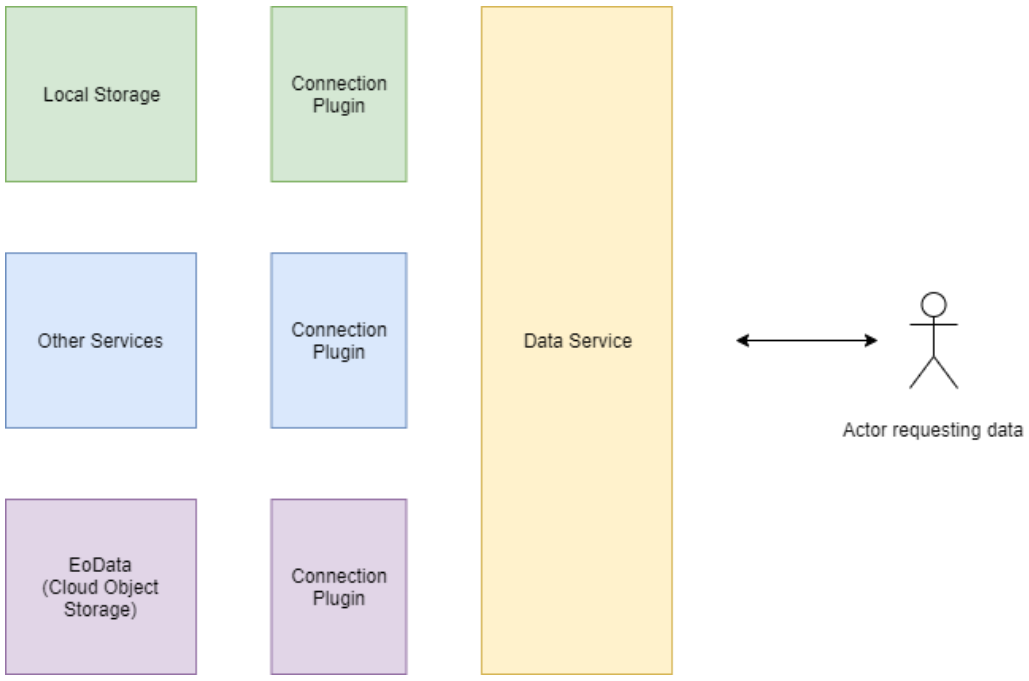

Figure 4- HiSea data services plugin architecture

As described in Figure 4, when a service or interface request data from the Data Service, the request will be mapped to the correct data repository and use the right plug-in to obtain data on the fly from that specific location.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 821934

The advantage of this approach is that the service can be developed independently of the available sources. At any time, it can be expanded by simply adding a plug-in that correctly maps the new source.

Taking into account the EoData test performed, the object storage is the preferred pathway to access the COPERNICUS data from the EoData repository available in the DIAS platforms. The file system interface would provide a more comfortable development model, but its limitations are not compatible with the performance requirements of the HiSea platform. The EO Data connection plugin must work over the S3 of Swift protocols.

It is relevant to notice that this service architecture allows for the same services that access the external data sources to be used to connect to the project partner sources by merely implementing an appropriate plugin. It is the responsibility of the connection plugin to use a caching mechanism if necessary.

6 Connecting HiSea Data services to other data sources

Other data sources that cannot be accessed using HiSea Data Services proxy will require the HiSea download service to create a local copy of the data. As described in *"Accessing other data sources"* section, the problem with the variety of formats and protocols used to access this data. The data retrieval process, as shown in Figure 5, is splitted into the following steps:

- Data enumeration – decide which "files" to download
- Data acquisition – transfer data from the remote system to local storage.
- Data conversion – transform the acquired data into HiSea standard formats
- Sending data to storage – Signal HiSea Data services that the dataset was updated

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 821934

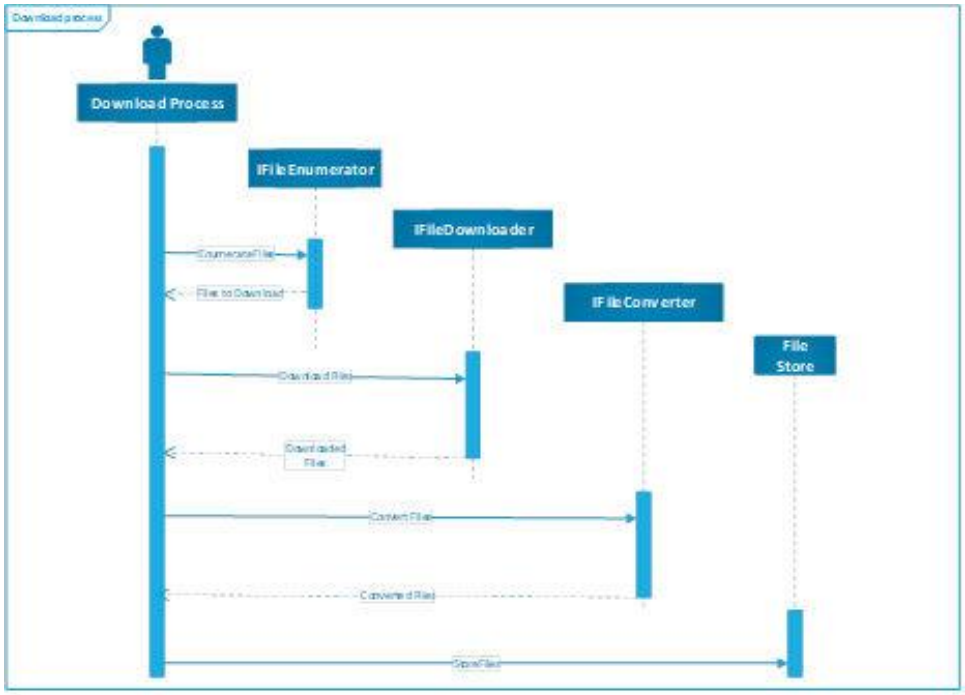

Figure 5- Data Collection Process

The Download Job entity encapsulates the configuration of each download process. Each download Job contains the description of:

- From where to retrieve data (RemoteAdress, Remote Credentials, etc.)
- how often to retrieve data (via a Cron expression)
- time period to be retrieved (hindcast, forecast, and ZeroTime)
- what data to retrieve (IFileEnumerator, DownloadProduct)
- how to retrieve it (IFIleDownloader)

Each enumerator plugin must implement the IFileEnumerator is the interface, ex:

- A MOTU plugin will generate the download command line for the selected products in the selected period
- An FTP plugin will generate the filenames for a specific FTP directory for the selected products in the selected period (e.g., Files with the dates on the filename, or files grouped in a directory by dates, etc.)

The IFIleDownloader will receive the list of "files" to download and perform the download ex:

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 821934

- OPENDAP downloader
- FTP client
- HTTP client

The period where to retrieve files from is calculated by:

- Start date = Current day at midnight + Zero time – hindcast
- End Date = Current day at midnight + Zero time + forecast

The DownloadProduct entity encapsulates the description of which product to download along with the necessary conversions (Figure 6).

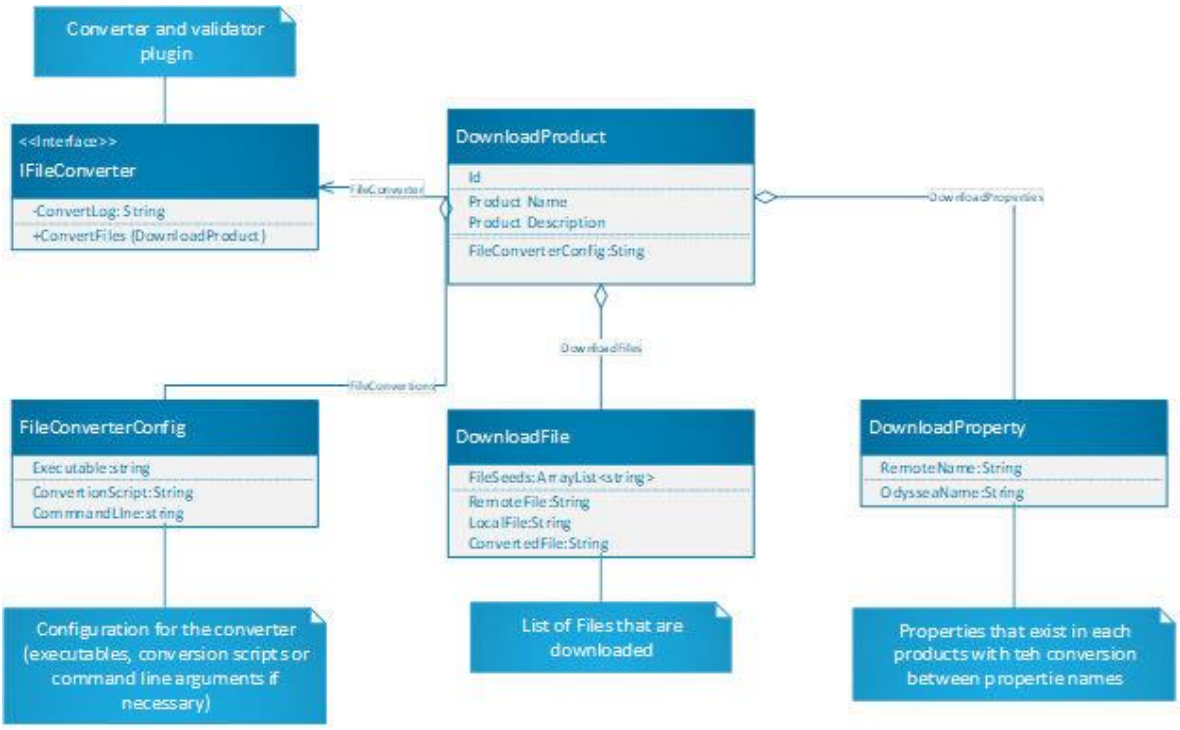

Figure 6 - DownloadProduct entity

Each product represents one or more files in the remote server. The DownloadFile entity describes this situation. This class contains the file seeds that are necessary for the IFIleEnumerator to populate the RemoteFile property. Once the download executes, the LocalFile property will provide the path to the file. Finally, after conversion, the ConvertedFile will contain the way to the converted data.

Each DownloadProduct contains a list of DownloadProperty. The DownloadProperty acts as a dictionary between the names of properties in the remote files and the conventions used in HiSea.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 821934

The DownloadProduct also references a plugin to manage the conversion from the original format into one of the HiSea standard formats. The FileConverterConfig will enable the users to configure how the transformation occurs.

7 Conclusions

The DIAS platforms provide multiple methods of accessing Copernicus data. After some testing, we concluded that the process that best suits the HiSea platform is cloud object storage either through S3 or SWIFT.

The HiSea data services use a plugin architecture to abstract the local or remote location of data. This architecture allows the service to display the same behavior independently of the location of the data.

Due to the dispersed nature of formats available on other sources that COPERNICUS, a flexible download component must be introduced into HiSea. The download component will use a plugin architecture to accommodate the multiple protocols and formats involved.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 821934