Demonstration on iCUPE data and service contingency

Alexander Mahura, Tuukka Petäjä, Hanna K. Lappalainen, Ella-Maria Duplissy (University of Helsinki); Steffen M. Noe (Estonian University of Life Sciences); Stefano Nativi, Paolo Mazzetti (Consiglio Nazionale delle Ricerche)

Helsinki, Finland, 30 October 2019

WP6: Dissemination and strategic development

Task 6.4: Future strategies and contingency plans

Del 6.4.1: Demonstration on iCUPE data and service contingency

Version 1

1. Introduction

The iCUPE WP6 ("*Dissemination and strategic development*") facilitates dissemination of the iCUPE (www.atm.helsinki.fi/icupe) project results and data to end-users, decision-makers, and stakeholders and facilitates interactions with corresponding communities. This WP interacts within the ERA-PLANET four strands and plans for the legacy and continuity of iCUPE beyond the project period. In particular, Task 6.4.1 ("*Demonstration on iCUPE data and service contingency*") interacts with Task 5.1 ("*iCUPE Data management plan*") and Task 5.5 ("*Facilitating iCUPE data pilots, data and services towards ERA-PLANET community, GEO and Copernicus*") in order to assure the iCUPE legacy and prepare for the continuation of iCUPE work after the end of the project.

The datasets (in total about 20) as products for researchers, decision- and policy makers, stakeholders and end-users communities will be produced as part of the Integrative and Comprehensive Understanding on Polar Environments (iCUPE) project activities. All these datasets are expected to be publicly available for different applications. Focusing on the Arctic region territories, the planned datasets will include novel data on anthropogenic contaminants in snow and ice cores and organic contaminants in the air-snow-water; concentrations of different chemical species and aerosols as well as their characteristics including vertical profiles; various atmosphere-hydrosphere-cryosphere-etc. related parameters in the Arctic based on ground-airborne-satellite-etc. platforms; near-real time parameters of the Arctic Research Infrastructures; others. Some datasets will focus on selected areas in northern latitudes, others - on geographical locations (measurement sites). A list of expected

datasets is presented at https://www.atm.helsinki.fi/icupe/index.php/datasets/list-of-datasets-as-deliverables.

As the iCUPE project is approaching its final stage, it is important figure out on (i) how produced datasets can be linked with other larger data storages/ repositories and data services; (ii) how delivered datasets can be exploited and used afterwards; and (iii) how does it benefit the activities related to other services, such as Copernicus. Overall, in the continuation there are topical options for the contingency, e.g. for iCUPE datasets, the following can be applied.

For that, selected iCUPE datasets are also to be tested and integrated into several platforms. One of these platforms is linked to the Copernicus services. Copernicus is known to be the largest data provider. To facilitate and standardize access to Copernicus data, there are 5 cloud-based online platforms known as the Data and Information Access Services (DIAS). These will provide mass storage and handling of data as well as centralized access to data, processing tools, and relevant information. The DIAS platforms (CREODIAS - creodias.eu; SOBLOO - sobloo.eu; MUNDI - mundiwebservices.com; ONDA - www.onda-dias.eu/cms; and WEkEO - www.wekeo.eu) allow users to explore, process, and download Copernicus data and information as well as have ability to process and combine with data from other sources. It is also possible to develop and host new applications in the cloud.

Other tested platforms for pre/post-processing/analysis data include the Virtual Laboratory, VLab (https://vlab.geodab.org) Google Earth Engine (https://earthengine.google.com), Polar Thematic Exploitation Platform, Polar-TEP (https://portal.polartep.io), Global Earth Observation System of Systems, GEOSS Portal (https://www.geoportal.org). Moreover, datasets can be also interlinked with the INTAROS web-catalogue (https://catalog-intaros.nersc.no/dataset).

2. iCUPE Datasets vs. Data Storages/ Platforms/ Services

2.1. iCUPE Repository

The iCUPE planned datasets are promoted to larger science and public communities through so-called dataset “teasers” (www.atm.helsinki.fi/icupe/index.php/submitted-datasets). These include basic information on planned datasets with contacts and short description (including where it is applicable illustrations and references).
For the Arctic regions, these include “promotional” materials on fractional snow cover area in selected sites of Svalbard islands; proxies for mixing layer height, condensation sink and gross primary production; dataset for ground-validation of precipitation measurements in high-latitudes;

atmospheric mercury speciation and isotope observations; time series of lake size changes; concentration of organic contaminants, mercury and other heavy metals in annual snow and shallow core records; source apportionment of organic aerosols including source regions; occurrence, transport and exchange fluxes of emerging organic contaminants; small-scale vertical and horizontal variability of the atmospheric boundary layer aerosol using unmanned aerial systems; absorption coefficient/ equivalent black carbon standardized dataset for long term impacts; continuous vertical observation of aerosol and cloud properties; and others. These also include those from the iCUPE Russian collaborators for the Russian Arctic: atmospheric mercury measurements at Amderma station; elemental and organic carbon over the north-western coast of the Kandalaksha Bay of the White Sea; micro-climatic features and Urban Heat Island intensity in cities of Arctic region; and others.

Since December 2018, so far, datasets on emerging organic contaminants in air/snow/water, anthropogenic contaminants in snow/ice cores, near-real time aerosol absorption measurements from selected regions/ locations of the Arctic regions were delivered and hosted by the University of Helsinki (www.atm.helsinki.fi/icupe/index.php/datasets/delivered-datasets), and more datasets are expected during upcoming 2019-2020.
Note that majority of archived datasets (as products) are linked (and directly downloadable) at the iCUEP website, and corresponding summary "Read-Me" files are available with detailed description and metadata information included. Following the iCUPE data management plan, the project coordinator team will maintain these datasets accessibility, and the raw data are to be hosted and maintained by the datasets providers.

For example, for the iCUPE dataset on emerging organic contaminants in air from the Arctic, such "Ream-Me" file contains the following detailed metadata description-information (https://www.atm.helsinki.fi/icupe/images/Datasets/Readme_iCUPE_DatasetReleased_DS-241_ver1.pdf):
"*The produced dataset (in MS Excel format) contains concentrations of neutral PFASs in vapor and particle phases of air samples (in total 44) collected at Ny-Alesund (Norway) (78.917° N, 11.933° E) from 27th September 2011 until 29th August 2012. The concentrations for 12 different PFASs are given in pg/m3. For each of 12 PFASs, the values for average blank (n=10), standard deviation, and method detection limit (MDL) are included. As 8:2 FTA, EtFOSA, MeFBSA and MeFBSE are not detectable in all filter blanks, their MDLs are set as 0 for particle phase in this work. In addition, an annex (Table) contains information on full names, acronyms, Chemical Abstract System (CAS) numbers, molecular weight (MW), chemical structure, method detection limits (MDL) and compound descriptors of the 12 PFASs*."
In case when measurements are carried out at different geographical locations during, for example, ship cruise, fieldwork, expedition, etc. additional information-map is provided within "Read-Me" file.

Note that for all expected datasets there is contact information of persons responsible for delivering the datasets (see in Annex for the iCUPE Data Management Report), and this information is available only at the internal iCUPE project website.

2.2. INTAROS Web-Catalogue

The INTAROS (Integrated Arctic Observation System; https://intaros.nersc.no) has strong focus on in-situ measurements (including fieldwork) contributing to the Arctic observing system. The INTAROS web-based data catalogue is focused on collection and making publicly available and shared various datasets linked to Arctic territories. At the current moment, this catalogue (https://catalog-intaros.nersc.no/dataset) includes about 50 datasets, and the number will be growing in 2020-2021 (with possibility of project extension for longer period), as new ones to be uploaded. Structurally, the catalogue has information on organizations contributed, related research themes/ topics, data formats, and types of licenses (majority under the Creative Common Attribution). Themes in the catalogue includes atmosphere, ocean, sea ice, marine ecosystem, terrestrial, glaciology, natural hazards, and community-based monitoring.
In addition to the iCUPE's hosted repository, the iCUPE selected datasets (fx. time-series of short- and long-term observations at specific geographical locations) can be also uploaded into the INTAROS data web-catalogue, making them more widely distributed for both the research and stakeholder communities.

2.3. GEOSS Portal

The GEOSS (Global Earth Observation System of Systems; www.earthobservations.org/geoss.php) is a set of observation, information and processing systems that interact and provide access to diverse information. GEOSS links these systems to strengthen monitoring. It shares environmental data collected from various observing systems and ensures that these data are findable, accessible, identifiable, reusable and interoperable to support development of tools and services. The GEOSS Portal (https://www.geoportal.org) allows to search data, imagery and analytical software packages. It connects users to existing databases and portals, provides reliable and up-to-date information. Information on descriptive keywords, protocol, format, source, and responsible organization are included as well as search and attribution by geolocation and thematic areas is also included. Information, metadata, and corresponding links to iCUPE datasets can be also added to the GEOSS portal.

2.4. Polar-TEP

The Polar Thematic Exploitation Platform (https://portal.polartep.io) provides working environment for users to remotely access algorithms and data, computing resources and tools in the cloud. It allows to avoid transferring of large observational datasets around the world. The platform focus is on Arctic regions. The Polar TEP platform includes data, information, software and community services. It has data sharing through discovery, access, and transportation capabilities. It has ability to provide standardized methods to create, manage, exchange, and extract useful information from data. It has software, applications, and tools (including opportunity to develop own), which can be used remotely. It provides infrastructure for computing resources, storage and networking capabilities. It also provides tools to publish, share and discuss data/results on the platform. To access the platform, register at: https://portal.polartep.io/help-sso.
Selected iCUPE datasets can be also uploaded and analyzed in the Poler-TEP platform environment.

2.5. Google Earth Engine

The Google Earth Engine (https://earthengine.google.com) combines a catalog of satellite imagery and geospatial datasets with multi-scale analysis capabilities. It is cloud-based geospatial processing platform as well as “big data” and visualization platform. It is a useful tool to detect changes, map trends, and quantify differences on the Earth's surface. This platform can be used for visualization and analyses of geospatial datasets. In particular, it has satellite imagery and stores it in a public data archive that includes historical (> 40 years) images, and these are added daily and available for multi-scale data mining. More than 10 PB of such data are available including imagery, geophysical data, etc. with more than 300 analysis ready datasets. It also provides Javascript and Python APIs and other tools to enable analysis of large datasets.
The Earth Engine provides a web-based access to, primarily, satellite imagery and other geospatial data in an analysis-ready format. There is searchable data catalog (https://developers.google.com/earth-engine/datasets), which includes Landsat, MODIS, Sentinel-1, NAIP, precipitation, sea surface temperature, CHIRPS climate, elevation, etc. data. Moreover, it is also possible to upload/ import your own imagery and data into the Earth Engine for analysis. Examples of several realized case studies employing Google Engine are given at https://earthengine.google.com/case_studies. For access, you need to apply through signup.earthengine.google.com. Documentation is available at https://developers.google.com/earth-engine. In addition, the Earth Engine Explorer is a user-friendly graphical interface (https://explorer.earthengine.google.com/#workspace), which has already a large collection of various datasets (mostly remote sensing data type) publicly available for downloading, visualization and analysis/ interpretation.
For demonstration iCUPE datasets available as satellite data or time-series of observations, already developed operations through importing/exporting and processing arrays, matrices, time-series,

tables, vector data can be applied (https://developers.google.com/earth-engine/tutorials). The Earth Engine Code Editor (https://code.earthengine.google.com) is a primary online integrated development environment for a user.
Considering iCUPE datasets, both "Image" and "Charts"-types can be applied. In particular, there are possibilities to represent and analyze data as histogram, profile, spectra, time-series, etc. and combined variants.

2.6. Virtual Laboratory VLab

The Virtual Laboratory (VLab) platform is a tool for invocation of scientific workflow generating new knowledge from observation and simulation data as well as models. Using VLab, it is possible to access various remote datasets in repositories of external systems, make data harmonization, apply selected coordinate reference system and chose resolution, work in different programming environments. Through the VLab, it is possible to create applications such as dashboards, mobile apps, accessing data and workflows through different interfaces with different level of complexity. For each workflow created, description of workflow and its scheme, information on developers and organization are generated. VLab provides also functionalities for discovering workflows, publishing and running models on the platform. The VLab documentation and user guide are available at https://confluence.geodab.eu. In order to access platform, registration is required (https://vlab.geodab.org). Installation of both GitHub (https://github.com) and Docker (https://docs.docker.com/toolbox/overview) is also needed.
The VLab processing capabilities were tested with iCUPE remote sensing data and post-processing of modelling results.

2.7. Copernicus Services

The Copernicus (https://www.copernicus.eu/en) is the European Union initiative to provide free and open access to near-real-time data, models and forecasts based on data collected from satellites and in-situ observations. There are 6 thematic services provided by Copernicus. For several selected of these, the iCUPE datasets could represent practical interest to be considered and integrated, and especially for those using observations in the Arctic region domain for monitoring and forecasting purposes as well as historical data analyses.
The Copernicus Climate Change Service (C3S; https://www.copernicus.eu/en/services/climate-change) provides knowledge base to support adaptation and mitigation policies. It provides access to information for monitoring and predicting climate change. It also includes climate related indicators and indices. Access to datasets is at: https://cds.climate.copernicus.eu. The Copernicus Atmospheric Monitoring Service (CAMS; https://www.copernicus.eu/en/services/atmosphere) provides information on global and European atmospheric composition. It includes monitoring and forecasting components,

analyses of historical data. Access to datasets is at: https://atmosphere.copernicus.eu/data. For 3 other services data are available at: https://www.copernicus.eu/en/access-data/conventional-data-access-hubs and also linked to the Copernicus Data and Information Access Services (DIAS). Moreover, the Copernicus Land Monitoring Service (CLMS; https://www.copernicus.eu/en/services/land) provides geographical information on land cover, vegetation and water cycle. The Copernicus Marine Environment Monitoring Service (CMEMS; https://www.copernicus.eu/en/services/marine) provides information on physical state and variability, ocean and marine ecosystems dynamics. The Copernicus Emergency Management Service (CEMS; https://www.copernicus.eu/en/services/emergency) provides information based on satellite and in-situ data for crisis management, civil protection, etc. communities. It includes disasters caused by natural and man-made hazards.
The areas of application of the Copernicus services are very wide and include health, environmental monitoring, forecasting, weather, climate, renewable energy, spatial planning, forest and water management, agriculture, etc. Although datasets are freely available, but registration is also required in addition to access the services.

2.8. Copernicus DIAS Platforms

The Copernicus is the largest data provider, and especially, the provider of various satellite data. It became possible due to technological developments and advances in providing access to available data, which are rapidly growing. Due to improved access capabilities, the number of potential users is also growing and allowing to more new stakeholders and citizens to be involved. To facilitate and standardize access to Copernicus data, there are 5 cloud-based online platforms known as the Data and Information Access Services (DIAS). These will provide mass storage and handling of data as well as centralized access to data, processing tools, and relevant information. The DIAS platforms allow users to explore, process, and download Copernicus data and information as well as have ability to process and combine with data from other sources. It is also possible to develop and host new applications in the cloud.

The **CREODIAS** (https://creodias.eu) platform is a cloud infrastructure (including storage cluster and services) adapted for processing big amounts of Earth Observation (EO) data and support scientific, operational and commercial applications. This consists of repository, processing platform and service tools. Currently, there are more than 30 datasets available (https://discovery.creodias.eu/dataset) and mostly these are satellite/ remote sensing data. The data can be accessed various interfaces.
The processing (https://creodias.eu/processing) covers full set of virtual resources available: Virtual Machines, VMs with operating systems available (such as CentOS, Ubuntu, Debian, Linux, etc.). The virtual storage (https://creodias.eu/storage) volumes can be mounted to VMs together with object storage solution, virtual networks, virtual appliances, physical servers, etc. The data discovery, indexing

and processing solution is an element of the CREODIAS platform. The Finder Tool (https://finder.creodias.eu) allows to find data stored, obtained or processed at selected times with selected coverage levels and other selection criteria. The Browser Tool (https://browser.creodias.eu) allows to visualize and process selected collections of data. Extended search capabilities are provided through SPARQL interface by linking metadata of all products stored in the CREODIAS repository with various information from the Internet.
Selected ICUPE datasets, and especially the satellite/ remote sensing data, can be also uploaded/ integrated into the CREODIAS repository, although details of pricing suggestions (https://creodias.eu/price-list) should be evaluated in case when long-term storage is considered. To register and create CREODIAS account, follow instructions at https://portal.creodias.eu/register.php.

The **SOBLOO** (https://sobloo.eu) platform provides data processing on-a-fly in order to produce needed geo-information. Data can be merged from different sources by automatic data fusion. In case of image (fx., satellite data) processing radiometric and geometric corrections can be performed to make more efficient analysis and interpretation of data. In perspective, analytical information (fx. such as economic activity indicators based on observation data) will be included. All these are current cloud services provided by SOBLOO. All uploaded datasets are available though (https://sobloo.eu/wui) searching and visualization tools. The SOBLOO's infrastructure provides cloud services (for computing, storage, networking – https://sobloo.eu/cloud-services) and managed services (operational systems, middleware, databases - https://sobloo.eu/managed-services). In particular, the services for the later include also optimization, upgrade, backups, and real-time monitoring of the database.
Selected ICUPE datasets, and especially the satellite/ remote sensing data, can be also uploaded/ integrated into the SOBLOO repository. In case of commercial services (not free), details on billing should be taken into account (https://sobloo.eu/marketplace). For registration and access of the SOBLOO services there is a need to create an account (https://sobloo.eu/help/getting-started-soblooeu).

The **ONDA** (https://www.onda-dias.eu/cms) platform allowing users to host their data and to build their applications in the cloud. At current moment, it has about 20 PB of data, with more than 23.6 mln products. In addition to collection of satellite data, through ONDA platform the access to the Copernicus's Marine, Land, and Atmosphere monitoring services is provided through a developed web-based interface. Data of very high resolution are available commercially (https://www.onda-dias.eu/cms/data/catalogue/very-high-resolution-data). There are virtual servers and environments, software tools for data processing and development of own applications (https://www.onda-dias.eu/cms/services/catalogues/tools). The ONDA pricing list is given at: https://www.onda-dias.eu/cms/services/catalogues/virtual-servers/#general-purpose.

Selected iCUPE datasets can be also uploaded and analyzed using the ONDA platform. Registration is required through https://www.onda-dias.eu/crm.

The **MUNDI** (https://mundiwebservices.com) platform provides, first of all, access to satellite data. It provides data, services and tools such as image processing, virtual computing servers and networks, storage capacities, methods and analysis functions. For selected the customized pricing list is available at: https://mundiwebservices.com/offer. MundiWebServices is targeted towards services for agriculture, urban, tourism and transport, energy, environment.
The iCUPE satellite related data could be integrated into such platform as for demonstration capabilities. Registration is required through https://mundiwebservices.com/register.

The **WEkEO** (https://www.wekeo.eu) platform provides service for environmental data and virtual processing environments in the cloud. It has about 200 datasets available to download and analyze. These include satellite data, modelling results, and derived products. It is also possible to transform Copernicus data into own specific services as well as upload and share data or infrastructure with WEkEO. From June 2018 it is open to large data and resources providers and subject to negotiations.
Selected iCUPE datasets can be also uploaded and analyzed using the WEkEO platform. Registration is required through https://www.wekeo.eu/documentation/registration.

2.9. Others Possibilities

In addition, information on suitable repositories for datasets produced by iCUPE can be also located using the Registry of Research Data Repositories (www.re3data.org) and Zenodo (zenodo.org) with providing tools to link publications and data. Information on research data management is also available from the Digital Curation Centre (www.dcc.ac.uk/dmponline) and ScienceMatters (www.sciencematters.io). In addition, the Research Data Alliance provides the Metadata Standards Directory (rd-alliance.github.io/metadata-directory) that can be searched for discipline-specific standards and associated tools, and the EUDAT B2SHARE (b2share.eudat.eu) tool includes a built-in license wizard that facilitates the selection of an adequate license for research data.

3. Concluding Remarks

As the iCUPE project is approaching its final stage, it is important to evaluate on how produced datasets can be linked with other larger data storages/ repositories, how in future datasets can be exploited; and how these can benefit other services.

The produced iCUPE datasets have possibilities to be integrated into several platforms considered in this report, and in particular, those platforms linked to the Copernicus services. Five cloud-based online platforms known as the Copernicus Data and Information Access Services can be potentially used. These provide storage, handling and centralized access to data, processing tools, and relevant information. These cloud-based platforms are CREODIAS - creodias.eu; SOBLOO - sobloo.eu; MUNDI - mundiwebservices.com; ONDA - www.onda-dias.eu/cms; and WEkEO - www.wekeo.eu. These allow users to explore, process, and download data and information as well as to have ability to process and combine with data from other sources, to develop and host new applications.

The iCUPE dataset in addition to own project repository can be also interlinked with the INTAROS web-catalogue (catalog-intaros.nersc.no/dataset). Moreover, platforms for pre/post-processing/analysis of iCUPE data include the Virtual Laboratory (vlab.geodab.org), Google Earth Engine (earthengine.google.com), Polar Thematic Exploitation Platform (portal.polartep.io), Global Earth Observation System of Systems's Portal (www.geoportal.org).