Large-scale EXecution for Industry & Society

Deliverable D7.2

Architectural Requirements and System Design for Interchange of In-situ Unstructured Weather & Environmental Observations

Co-funded by the Horizon 2020 Framework Programme of the European Union

Grant Agreement Number 825532

ICT-11-2018-2019 (IA - Innovation Action)

DELIVERABLE ID | TITLE	D7.2 | Architectural Requirements and System Design for Interchange of In-situ Unstructured Weather & Environmental Observations
RESPONSIBLE AUTHOR	Antonio Parodi (CIMA)
WORKPACKAGE ID | TITLE	WP7 | Weather and Climate Large-scale Pilot
WORKPACKAGE LEADER	CIMA
DATE OF DELIVERY (CONTRACTUAL)	31/03/2019 (M03)
DATE OF DELIVERY (SUBMITTED)	01/04/2019 (M04)
VERSION | STATUS	V1.2 | Final
TYPE OF DELIVERABLE	R (Report)
DISSEMINATION LEVEL	PU (Public)
AUTHORS (PARTNER)	Antonio Parodi (CIMA), Emanuele Danovaro (CIMA)
INTERNAL REVIEW	Rubén Jesús García-Hernández (LRZ); Jan Martinovič (IT4I)

Project Coordinator: Dr. Jan Martinovič – IT4Innovations, VSB – Technical University of Ostrava
E-mail: jan.martinovic@vsb.cz, **Phone:** +420 597 329 598, **Web:** https://lexis-project.eu

DOCUMENT VERSION

VERSION	MODIFICATION(S)	DATE	AUTHOR(S)
0.1	First Draft	02/03/2019	Antonio Parodi (CIMA), Emanuele Danovaro (CIMA)
0.2	Review	14/03/2019	Rubén Jesús García-Hernández (LRZ)
0.3	Feedback from internal review	19/03/2019	Antonio Parodi (CIMA), Emanuele Danovaro (CIMA)
1.0	First public version	27/03/2019	Antonio Parodi (CIMA), Emanuele Danovaro (CIMA)
1.1	Final review, final formatting	31/03/2019	Jan Martinovič (IT4I); Tomáš Martinovič (IT4I); Kateřina Slaninová (IT4I)
1.2	Final version	1/04/2019	Antonio Parodi (CIMA), Emanuele Danovaro (CIMA)

GLOSSARY

API	Application Program Interface
ASCII	American Standard Code for Information Interchange
BUFR	Binary Universal Form for the Representation of meteorological data
CSV	Comma-Separated Values
DIAS	Data and Information Access Services
EC	European Commission
ERDS	Extreme Rainfall Detection System
FTP	File Transfer Protocol
GeoTIFF	Geo-referenced Tagged Image File Format
GPM	Global Precipitation Measurement
GRIB	Gridded Binary or General Regularly-distributed Information in Binary form
HDF5	Hierarchical Data Format version 5
ICPD	Italian Civil Protection Department
IMERG	Integrated Multi-satellitE Retrievals for GPM
IR	Infrared
IoT	Internet of Things
JSON	JavaScript Object Notation
NetCDF	Network Common Data Form
NWP	Numerical Weather Prediction
ODB2	A data format for observations
OPeNDAP	Open-source Project for a Network Data Access Protocol
PMW	Passive Microwave
REST	Representational State Transfer
SOAP	Simple Object Access Protocol
SensorML	Sensor Model Language
TRMM	Tropical Rainfall Measuring Mission
WaterML	Water Model Language
WCDA	Weather and Climate Data API
WMO	World Meteorological Organization
WRF	Weather Research and Forecasting

TABLE OF CONTENTS

EXECUTIVE SUMMARY 5

1 INTRODUCTION AND DATA SOURCES 6

1.1 USE CASE: PROVISION OF WEATHER STATION DATA 6

1.1.1 Meteonetwork 6

1.1.2 Italian Civil Protection Department 7

1.1.3 Air Quality Sensors 7

1.2 USE CASE: PROVISION OF IOT DATA 8

1.2.1 Smart Gateway IoT Data [TESEO] 8

1.2.2 Moji Weather Data [ECMWF] 8

1.3 OPERA RADAR DATA 9

1.4 SATELLITE DATA 10

1.4.1 ESA Sentinel remote sensing products 10

1.4.2 Integrated Multi-satellitE Retrievals 10

2 DATA ACCESS 11

2.1 INTEGRATION WITH LEXIS PORTAL 11

2.1.1 BUFR 11

2.1.2 SensorML 11

2.1.3 GRIB 11

2.1.4 HDF5/NetCDF 12

3 CONCLUSION 13

LIST OF FIGURES

Figure 1 Meteonetwork (and related partners) temperature stations spatial distributions 6
Figure 2 ICPD authoritative rain gauge stations spatial distributions 7
Figure 3 Moji weather app 8
Figure 4 OPERA Radar Network.....9
Figure 5 Example global precipitation map generated from IMERG early run data 10

EXECUTIVE SUMMARY

The Weather and Climate Data API (WCDA) is the data management layer for curated weather data in LEXIS. It is responsible for storing and organizing weather observations from a variety of sources (including in-situ unstructured observations), as well as numerical weather prediction outputs and intermediate weather data. This report focuses on collection and curation of weather-related observations. It complements D7.1 [1] by providing details on the data sources and related requirements. It depends primarily upon the workflows defined by the Weather and Climate Large-scale Pilot (WP7).

Position of the deliverable in the whole project context

This deliverable is part of Task 7.2 entitled "Global weather and climate: from global in-situ unstructured observations to forecast products on the cloud", which is the second task in WP7 (Weather and Climate Large-scale Pilot). Acquisition and curation of observational datasets is a key asset in weather & climate modelling, thus most WP7 models will benefit from the outcome of this deliverable. It is a precursor to D7.5 (M15), which is the first release and test-bed deployment of the WCDA.

Description of the deliverable

The deliverable begins with an introduction to the WCDA and gives more detail into the observation data sources, data heterogeneity, overall requirements of data curation, storage and access. Section 2 focuses on Weather Station data sources, provided by the Italian Civil Protection Department (ICPD), Meteonetwork and Airparif (air pollutant); Internet of Things (IoT) data sources (Moji pressure data and Teseo smart gateway for weather stations) and structured data sources such as EUMETNET OPERA Radar [1] and Sentinel satellite imagery [2].

1 INTRODUCTION AND DATA SOURCES

This deliverable complements D7.1 [3] in the collection of architectural requirements as for interchange of in-situ unstructured weather & environmental observations. In particular, D7.2 will describe the available in-situ unstructured and structured weather & environmental observations, as well as the corresponding access techniques. The following preliminary list of in-situ unstructured weather & environmental observations sources is considered: the personal weather stations data provided by the Meteonetwork association over Italy and continental Europe; selected authoritative in situ weather stations data from the Italian Civil Protection datasets; air-quality sensors over industrial and urban areas in France, provided by NUM; IoT data, namely the surface pressure data provided by the Moji Weather (the largest social weather app in China with more than 500 million downloads) and TESEO smart gateway; selected datasets from the ECMWF global observing systems; Copernicus remote sensing observational data [4] (e.g. Sentinel-1, Sentinel-2, Sentinel-3 and Sentinel-5P user products).

The second part of the deliverable will present the suggested standards for the provision of the available data. Final comments will be drafted in the conclusions section.

1.1 USE CASE: PROVISION OF WEATHER STATION DATA

Observations from weather stations usually include temperature, humidity, wind and precipitation measurements. Such data are relevant for Assimilation in Numerical Weather Predictions, i.e. Weather Research and Forecasting (WRF) predictions executed by CIMA; predictions and observations analysis, i.e. Extreme Rainfall Detection System (ERDS) executed by ITHACA; and for validation of WP7 models.

We plan to collect data from thousands of weather stations provided by Meteonetwork and ICPD. The data rate varies from one acquisition per minute to one acquisition per hour. Some data sources already provide WaterML data format [5], while others are plain ASCII text or CSV data.

1.1.1 Meteonetwork

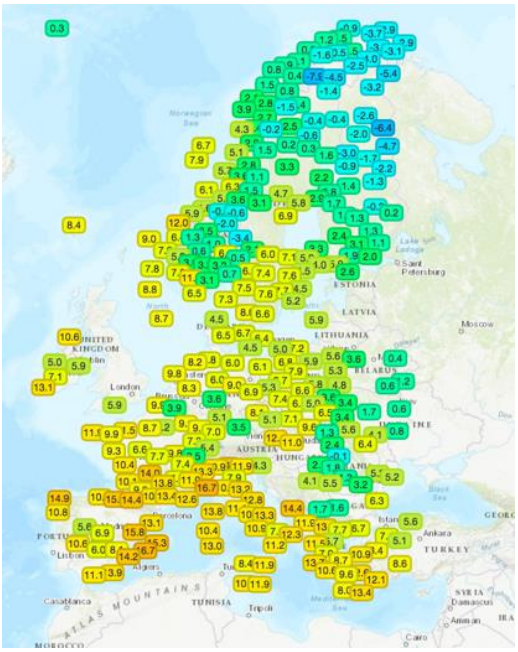

Figure 1 Meteonetwork (and related partners) temperature stations spatial distributions

Meteonetwork[1] was founded in 2002 in Lombardy (Italy), with the aim of raising public awareness about meteorological and climatological issues. Through the years the organization, besides the continuous holding of events such as meetings, conferences and talks, has been standing out because of its forum and its wide network of weather stations. Meteonetwork also cooperates with several public and private bodies, among which Centro

[1] http://www.meteonetwork.it/

Epson Meteo in Milan stands out. Since 2008 it has been officially recognized as a non-profit organization by the Province of Milan and at the same time the first regional branches developed with the aim of spreading even further the founding values.

Meteonetwork manages, in cooperation with its partner associations, more than 2 000 weather stations all over Europe, namely: thermometers (Figure 1 a de-cluttered representation of Meteonetwork thermometers), hygrometers (to measure humidity), anemometers (to measure wind speed and wind direction), and rain gauges. The weather stations provide their observations with temporal resolutions ranging from 10 minutes to 1 hour, with a dataflow around 50 MB/day. The data are available in ASCII format through a set of dedicated endpoints[2].

1.1.2 Italian Civil Protection Department

The Italian Civil Protection Department (ICPD) operates in designing and managing in real time risk reduction actions over the national territory, determined by high impact adverse weather, through its *Centro Funzionale Centrale* (Central Functional Center), and coordinating a Federated national Early Warning System, in collaboration with regional authorities. In this framework, ICPD manages a large number of in-situ authoritative weather stations: 6 059 rain gauges (Figure 2), 2 299 hydrometers, 4 373 thermometers, 1 270 barometers, about 2 500 anemometers, and finally 2 683 hygrometers. The corresponding dataflow, with data temporal resolution down to 1 minute, is about 100 MB/day. CIMA Foundation archives and curates the aforementioned data on behalf of ICPD. These data can be made available to the LEXIS project, for environmental monitoring and civil protection research activities, pending the formal approval of ICPD, via API based download services. The adopted format for the data dissemination is the Water Model Language (WaterML 2.0[3]), which is a standard model for the representation of observation data linked to the waters, with the aim of enabling and promoting the exchange of these data among different computer systems.

Figure 2 ICPD authoritative rain gauge stations spatial distributions

1.1.3 Air Quality Sensors

NUM will serve hourly air-quality measurements: Sulfur dioxide (SO_2), Ozone (O_3), Nitrogen dioxide (NO_2) and Coarse Particulate Matter (PM_{10}) on selected French sites, both in industrial and urban environment. Such data are freely available to all WCDA users (Open Licence) and are provided by Airparif[4], a local French Air quality agency active in Paris area.

[2] REST: http://api.meteonetwork.it/xml_rpc/public/rest
SOAP: http://api.meteonetwork.it/xml_rpc/public/soap
JSON-RPC 2.0: http://api.meteonetwork.it/xml_rpc/public/jsonrpc
[3] http://www.opengeospatial.org/resource/products/byspec/?specid=537
[4] https://www.airparif.asso.fr

Each station is able to sense a subset of the different pollutant, namely SO_2 - 5 stations, O_3 - 5 stations, NO_2 - 18 stations, PM_{10} - 9 stations. Raw data are encoded in CSV file format.

1.2 USE CASE: PROVISION OF IOT DATA

Smart devices are ubiquitous and are usually equipped with a large number of sensors (i.e. temperature, pressure) and direct connectivity to the Internet. We aim at collecting a massive amount of pressure data from consumer cellphones, as well as design and manage a Smart Gateway capable of connecting a wide range of high-quality sensors. Data collected from such IoT data sources will be curated and served through WCDA.

1.2.1 Smart Gateway IoT Data [TESEO]

One of the goals of WP7 is the development of an IoT Smart Gateway for acquisition, pre-filtering and delivery of weather data on the field. Smart Gateway is under active design phase, and the design document will be published as D7.3 [6]. According to preliminary requirements, the gateway will be able to deliver data in different data formats, with the notable inclusion of JSON, through REST APIs, and WaterML or SensorML[5] encoding, to guarantee standard and effective data representation.

The primary focus of SensorML is to provide a robust and semantically-tied means of defining processes and processing components associated with the measurement and post-measurement transformation of observations. This includes sensors and actuators as well as computational processes applied pre- and post-measurement.

The number of smart gateways and data rate has not yet defined and will be described in D7.3 [6].

1.2.2 Moji Weather Data [ECMWF]

Moji Weather is a social, crowd-sourced app generating real-time weather data from its user base around the world.

The Moji Weather app (see Figure 3), when the user opens it, will provide latitude, longitude, time-stamp and pressure. At the moment this app provides approximately ten million data points per day, corresponding to a data flow of 400 MB/day. At the time of the deliverable writing, the format of the data is yet to be determined, but they will be converted into some standard: among the candidate standards there are certainly BUFR and ODB2.

Figure 3 Moji weather app

[5] http://schemas.opengis.net/sensorML/

1.3 OPERA RADAR DATA

OPERA[6] is the Radar Programme of EUMETNET. OPERA main objectives are to provide a European platform wherein expertise on operationally-oriented weather radar issues are exchanged, as well as to develop, generate and distribute high-quality pan-European weather radar composite products on an operational basis. OPERA has been coordinating radar data exchange in Europe for 20 years, and its data centres have been operational for almost a decade. Odyssey, the OPERA Data Centre, generates and archives composite products from raw single site radar data, belonging to the OPERA Radar Network (Figure 3), using common pre-processing and compositing algorithms: instantaneous surface rain rate, instantaneous max reflectivity, and one-hour rainfall accumulation. The composites cover the whole of Europe in a Lambert Equal Area projection. They are updated every 15 minutes and issued ca. 15 minutes after data time.

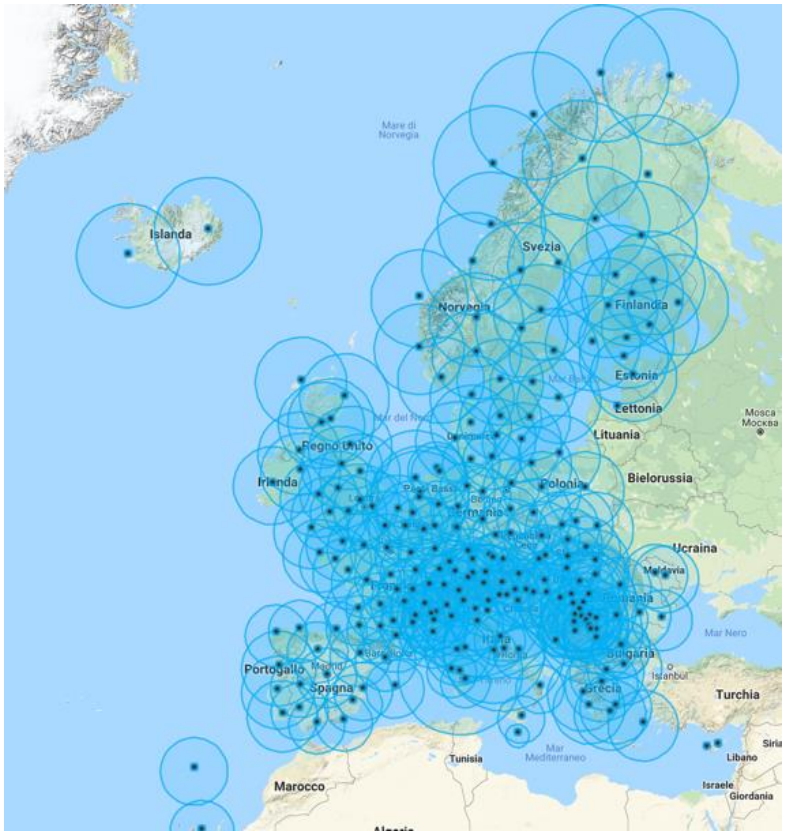

Figure 4 OPERA Radar Network

In the surface rain rate composite, each composite pixel is a weighted average of the lowest valid pixels of the contributing radars, weighted by the inverse of the beam altitude. Polar cells within a search radius of 2.5 km of the composite pixel are considered. Data measured below 200 m altitude are not used. The measured reflectivity values are converted to rainfall (mm/h) using the Marshall-Palmer equation [7]. Rainfall accumulation is simply the sum of the previous four 15-minute surface rain-rate products. In the Maximum reflectivity composite each composite pixel, at 2 km grid spacing, contains the maximum of all polar cell values of the contributing radars at that location. Composites are available in two formats: BUFR and HDF5. Each file has two fields: the data field and the quality field. The members of OPERA and EUMETNET may use the composites for their official duties without a separate license. The OPERA products are also available under license to 3rd parties: LEXIS project is working on finding an agreement under a research and education license.

[6] http://eumetnet.eu/activities/observations-programme/current-activities/opera/

1.4 SATELLITE DATA

1.4.1 ESA Sentinel remote sensing products

When all Sentinel satellites are operational (Sentinel-1A and 1B, Sentinel-2A, and Sentinel-3A are at the time of writing providing data like land surface temperature, soil moisture, sea surface temperature, and wind speed at sea surface), they will deliver in excess of 10 petabytes of data each year. Information from the Copernicus services, derived from the Sentinels, other satellite data as well as information from the Copernicus' in situ component, add to the total amount of geospatial data generated or made available by the Copernicus programme. This makes Copernicus the third largest data provider in the world, creating great opportunities, but also presenting great challenges. The European Commission (EC) has ambitious plans to tackle these challenges in a big-data enabled environment, and for that purpose, has decided to follow two different approaches: the Open Access Hub and the Copernicus Data and Information Access Services (DIAS).

On one side, the Open Access Hub provides complete, free and open access to Sentinel-1, Sentinel-2, Sentinel-3 and Sentinel-5P user products[7]. The Data Hub exposes two APIs for browsing and accessing the Earth Observation data stored in the rolling archive: OData and Open Search.

OData[8] is a data access protocol exposing REST services. WCDA will exploit OData to serve with an unified interface all Sentinel products. Open Search[9] is a set of technologies that allow publishing of search results in a standard and accessible format. It can be used to quickly identify the required resource which can then be downloaded by using OData. It is a RESTful technology and the Data Hub implementation uses the Apache Solr search engine.

1.4.2 Integrated Multi-satellitE Retrievals

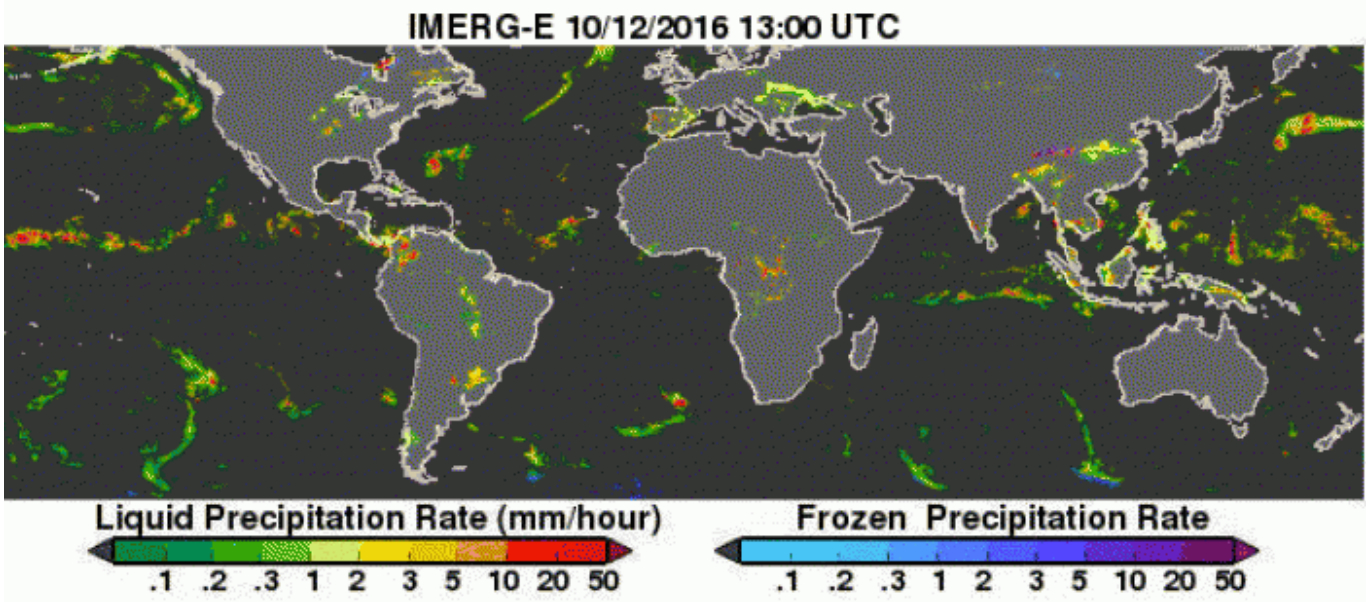

Figure 5 Example global precipitation map generated from IMERG early run data

The Integrated Multi- satellitE Retrievals for GPM (IMERG) is the unified U.S. algorithm that provides the Day-1 multi-satellite precipitation product for the U.S. GPM team. The precipitation estimates from the various precipitation-relevant satellite passive microwave (PMW) sensors comprising the GPM constellation are computed using the 2014 version of the Goddard Profiling Algorithm [8], then gridded, intercalibrated to the GPM Combined Instrument product, and combined into half-hourly 0.1°x0.1° fields.

[7] https://scihub.copernicus.eu/userguide/
[8] https://scihub.copernicus.eu/userguide/ODataAPI
[9] https://scihub.copernicus.eu/userguide/OpenSearchAPI

This algorithm is intended to intercalibrate, merge, and interpolate satellite microwave precipitation estimates, together with microwave-calibrated infrared (IR) satellite estimates, precipitation gauge analyses, and potentially other precipitation estimators at fine time and space scales for the TRMM and GPM eras over the entire globe. The system is run several times for each observation time, first giving a quick estimate and successively providing better estimates as more data arrive. The final step uses monthly gauge data to create research-level products.

Original data are provided in HDF5 and GeoTIFF[10] file format through FTP and OPeNDAP[11] service.

2 DATA ACCESS

2.1 INTEGRATION WITH LEXIS PORTAL

The LEXIS Portal (WP8) would like to make parts of the WCDA available to third-party users. The main requirements from this perspective concern authorization and accounting, and managing the curated data in the WCDA in such a way that limited access can be provided if necessary (shielding non-public data and limiting the number and size of requests to third-party users). Data Format

The data requests to the WCDA range from a few megabytes (in-situ unstructured data sources) to a few terabytes per day. WCDA will provide a harmonized data interface to personal weather stations, IoT Data, Satellite, Radar data and model products. Unstructured data will be exposed by converting the native format in standard data format (BUFR), while structured observations (Sentinel, OPERA Radar) will be provided in the native data format (HDF5, GRIB). We are also considering SensorML for the configuration, management and integration of the new SmartGateway.

2.1.1 BUFR

The Binary Universal Form for the Representation of meteorological data (BUFR) is a binary data format maintained by the World Meteorological Organization (WMO). The most recent version is BUFR Edition 4. BUFR Edition 3 is also considered current for operational use. The BUFR design goals are portability, compactness, and universality. Any kind of meteorological data can be represented, along with its specific spatiotemporal context and any other associated metadata. BUFR is characterized by table-driven code forms, where the meaning of data elements is determined by referring to a set of tables that are kept and maintained separately from the message itself.

2.1.2 SensorML

SensorML is an approved Open Geospatial Consortium standard. It provides standard models and an XML encoding for describing sensors and measurement processes. SensorML can be used to describe a wide range of sensors, including both dynamic and stationary platforms and both in-situ and remote sensors.

SensorML supports a wide range of use cases: description of sensor specification sheets; discovery of sensor, sensor systems, and processes; observation representations; on-demand processing of observations; support for tasking and alert services; support autonomous sensor networks and it provides services for archiving of sensor parameters. We focus on observation representations.

2.1.3 GRIB

GRIB (General Regularly-distributed Information in Binary form) is a concise data format, used in meteorology to store historical and forecast weather data. It is standardized by the World Meteorological Organization's Commission for Basic Systems. GRIB files are a collection of self-contained records of 2D data, and the individual records stand alone as meaningful data, with no references to other records or to an overall schema. Each GRIB record has two components - the part that describes the record (the header), and the actual binary data itself.

[10] http://geotiff.osgeo.org/
[11] OPeNDAP website: https://www.opendap.org

The data in GRIB-1 are typically converted to integers using scale and offset, and then bit-packed. GRIB-2 also has the possibility of compression.

Both the first and the second edition are widely used operationally worldwide by most meteorological centres, for Numerical Weather Prediction output (NWP).

2.1.4 HDF5/NetCDF

Hierarchical Data Format version 5 (HDF5) is a file format designed to store and organize large amounts of structured data.

HDF5 file structure includes two major types of object: Datasets, which are multidimensional arrays of a homogeneous type; Groups, which are container structures that can hold datasets and other groups.

This results in a truly hierarchical, filesystem-like data format. In fact, resources in an HDF5 file can be accessed using the POSIX-like syntax /path/to/resource. Metadata is stored in the form of user-defined, named attributes attached to groups and datasets. More complex storage APIs representing images and tables can then be built up using datasets, groups and attributes.

HDF5 includes a type system, and dataspace objects that represent selections over dataset regions. The API is also object-oriented with respect to datasets, groups, attributes, types, dataspaces and property lists.

NetCDF (Network Common Data Form) is a self-describing, machine-independent data formats that support the creation, access, and sharing of array-oriented scientific data. Version 4.0 (released in 2008) allows the use of the HDF5 data file format.

3 CONCLUSION

The Weather and Climate Data API has the goal to provide a clean and effective interface for interchange of in-situ unstructured weather & environmental observations and weather products. This document focuses on environmental observations that will be made available through WCDA, analyzing available datasets and related metadata (data rate, geographical distribution of the observations). Moreover, we provide a brief description of the supported data formats that will be made available through the WCDA.

REFERENCES

[1] “OPERA, the Radar Programme of EUMETNET”, [Online]. Available: http://eumetnet.eu/activities/observations-programme/current-activities/opera/.

[2] “Sentinel mission,” [Online]. Available: https://sentinel.esa.int/.

[3] LEXIS Deliverable, *D7.1 Architectural Requirements and System Design for Interchange of Weather & Climate Model Output between HPC and Cloud Environments.*

[4] “Copernicus Open Access Hub”, [Online]. Available: https://scihub.copernicus.eu/.

[5] “OGC WaterML Standard”, [Online]. Available: https://www.opengeospatial.org/standards/waterml .

[6] LEXIS Deliverable, *D7.3 Design of a Smart Gateway for Collecting, Pre-processing and Transmitting In-situ Observations.*

[7] J. Marshall and W. Palmer, “The distribution of raindrops with size,” *J. Meteor,* vol. 5, no. 4, p. 165–166, 1948.

[8] C.D. Kummerow et al., *The Evolution of the Goddard Profiling Algorithm to a Fully Parametric Scheme*. J. Atmos. Oceanic Technol., 32, 2265–2280, https://doi.org/10.1175/JTECH-D-15-0039.1