Environmental Data Platform (EDP), a Solution to Work with Big Data, Standardized and Reproducible

Andrea Vianello
Eurac Research
Viale Druso 1
Bolzano, Italy
andrea.vianello@eurac.edu

Armin Costa
Eurac Research
Viale Druso 1
Bolzano, Italy
armin.costa@eurac.edu

Bartolomeo Ventura
Eurac Research
Viale Druso 1
Bolzano, Italy
bartolomeo.ventura@eurac.edu

Roberto Monsorno
Eurac Research
Viale Druso 1
Bolzano, Italy
roberto.monsorno@eurac.edu

Simone Tritini
Eurac Research
Viale Druso 1
Bolzano, Italy
simone.tritini@eurac.edu

Daniele Antonucci
Eurac Research
Viale Druso 1
Bolzano, Italy
daniele.antonucci@eurac.edu

Alexander Jacob
Eurac Research
Viale Druso 1
Bolzano, Italy
alexander.jacob@eurac.edu

Abstract

Our main objective is to provide a user-friendly data platform for researchers working with big environmental data available on our spatial data infrastructure and collaborating external cloud services. As a research centre, involved in many international projects, well defined and standardized access to data and processing facilities is paramount for effective collaboration with our partners. Collaboration means not only to share data and metadata but also reproducible software solutions. Consequently, partners can adopt those solutions into their own SDI and possibly become part of federated data network. To this end, building on open source solutions is key and comes with the additional benefit of being able to build solutions tailored to specific project requirements. Offering interfaces based on existing and widely adopted OGC standards guarantees interoperability with external applications and partners. Nevertheless, it remains a challenge to maintain the EDP on a state of the art level with new systems and solutions appearing frequently. To keep up with the pace the EDP is designed with modular components that can be updated or substituted when better solutions become available. Part of this is also to contribute to the development of new solutions and standards like the H2020 openEO project. It defines an API for accessing big earth observation cloud processing infrastructures and fosters open development of backend-drivers and clients libraries in R, Python and Javascript for harmonized access to a very heterogeneous service landscape.

Keywords: Environmental Data, Open Source, Spatial Data Infrastructure, OGC Standards, API

1 Research requirement

Monitoring environment (natural or urban areas) implicates acquisition of long time-series of heterogeneous data in various formats. Input data sources range from remote over proximal to in-situ sensing and after more than 10 years we collected about 1,5 Petabyte of data. To improve and speed up the research daily work, it is necessary to define a data life cycle to handle this big dataset (Data lifecycles analysis: Towards intelligent cycle, 2017) and simplify both discoverability and access to it, figure 1.

One of the biggest challenges in a long-running big data platform for smarter cities lies in the real capacity not only to drill into the data, but also, and most importantly, to leverage existing social and geographical ties between all main actors of local communities (LocalFocus: A Big Data Service Platform for Local Communities and Smarter Cities, 2018). Recently we started working with Internet of Things (IoT) applications where environmental conditions represent the trigger to set-up automatic actions ("ACT" in figure 1) for the connected devices. This function is becoming more and more

Figure 1: Data workflow to consider in research work.

Source: copyright own by the authors.

requested not only by researchers to find new applications fields but also by companies and common citizens for different purposes like fields irrigation, alarm systems for human or animal safety etc.

2 Environmental Data Platform

Currently we are setting up a spatial data infrastructure called Environmental Data Platform (EDP) in order to satisfy the aforementioned requirements of the researchers. This EDP integrates many applications in order to get the best usability in term of performance, big-data capability and data processing service that are not provided in a unique existing platform. Existing platform are mainly too specialist as smart cities, remote sensing oriented not easy to manage heterogeneous data. Our approach aim to be as more interoperable as possible, to allow for a simplified access to different sources of heterogeneous data.

The functionalities that we consider are:

- **Authentication** & **Authorization** services for the platform components
- **Discovery** of data and services
- **Visualization** tools
- **Analysis** & Processing tools
- **Access** Data services
- Data **Feeding** services (e.g. Automatic download of data, data collection from Wireless Sensor Networks)
- **Knowledge Base** (Code and Documentation repository)

The adoption of the Open Geospatial Consortium (OGC) standard implementations helps to guarantee interoperability between applications and external services. There is a global effort to first define interoperability requirements for environmental science, business and policy, and then develop and implement consensus-derived, free and open environmental IT standards that meet those requirements while co-evolving with the larger IT standards framework and advances in IT (OGC Information Technology Standards for Sustainable Development, 2015).

The challenging goal consists in creating a working environment for the analysis of time-series, made of heterogeneous data (raster, vector, punctual observation, models etc.). This requires multiple applications that provide web services for data access.

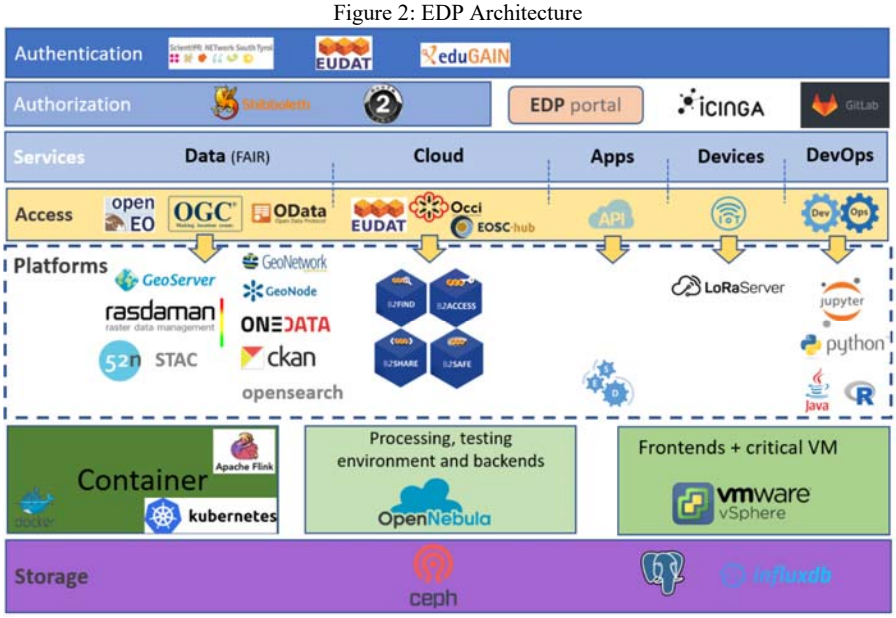
2.1 Architecture

The system is based on a cloud computing platform, considering the capability to access any information, at any time, regardless of the resources required and the location of the infrastructure, data, application or user (ISSI Scientific Report 15, 2018).

We have a virtualization environment and various hypervisors (green level in figure 2) to host all the Virtual Machines (VM) according to researcher/application requirements such as scalability, simplicity, etc. We organized the system in front-end VMs, available from internet, and back-end VMs, available only from intranet to protect applications and data from non-authorized accesses. In addition, we separated application servers from the big data-server and databases nodes to share datasets among applications and to set up easily backups and other maintenance tasks.

Datasets are stored in file system (mainly the archive of raster files), databases (PostgreSQL, InfluxDB) or datacubes (Rasdaman) and are available for the VMs to read/write

Figure 2: EDP Architecture



Source: copyright own by the authors.

operations accordingly with Access Control Lists (Organizing Access to Complex Multi-Dimensional Data: An Example from the ESA SEOM SInCohMap Project, 2017).

The general idea is to have many components (cloud back-ends) that provide functionalities to different clients using a unique API. We are collaborating in the openEO project (H2020 grant 776242 http://openeo.org/) to develop an open API (yellow box in figure 2) to connect R, Python, Javascript and other clients to big earth observation cloud back-ends in a simple and unified way. This step is fundamental to simplify user work by permitting a unique authentication and giving a unique access to web services according to user permissions.

We will benefit by openEO results using its API and developing missing drivers for our back-end API like SOS and InfluxDB. This API, compared to existing API, is the only one that aim to have a unique data access and that easily permit to develop missing drivers to improve the project result and usability.

The web portal is the main place where, after authentication, a user can discover and preview dataset, set up processing and store new outputs. Here the user will discover services available for the dataset of interest (visualization, processing, download, etc.). This improvement of the EDP is one of the main goals of the FESR1094 project DPS4ESLAB.

Finally, we are using a GitLab repository to store the implemented code. This tool is useful to collaborate in development of processing algorithms or web applications. Thanks to this, we have the possibility to change versions, to assign issues and review the history of changes.

[This work is part of the research activities of the project DPS4ESLAB, funded by the European Union Investments in favor of growth and employment programme under grant agreement No 8141/2016 del 04.05.2018]

2.2 Applications

The applications are the core of the EDP and have to provide functionalities described above and to satisfy user's requirements of usability. Possibly, EDP have to simplify researcher's work considering they are interested in the output of their analysis and not in the technology behind.

Most of the applications are web oriented and implement OGC standards web services. They are mainly open source solutions that implement OGC standards, in order to:

- customize application's code;
- guarantee to partners the possibility to easily reproduce adopted solutions;
- reduce costs of the infrastructure and its maintenance.

The decision of which software to choose, between all the existing solutions, is driven by aspects like the robustness, the fulfillment of the requirements, the possibility to customize the code (open source), the community and developers support and for sure the state of the art condition.

The figure 2, in the white box, represents the most important applications we use that we can group as follow:

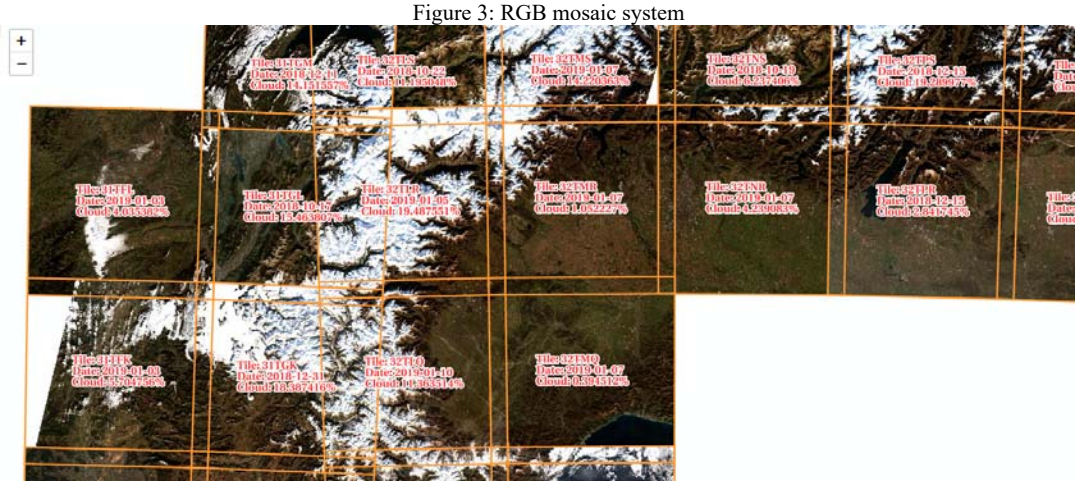
2.2.1 Catalogues and data management systems

- Geonetwork: provides a metadata service using the Catalogue Service for the Web standard (CSW);
- GeoNode: is a web-based application and platform for developing geospatial information systems (GIS) and for deploying spatial data infrastructures (SDI);
- CKAN: is a fully-featured, mature, open source data management solution, to make your data discoverable and presentable.

2.2.2 Spatial platforms and services

- GeoServer: is an implementation of a group of open standards such as Web Feature Service (WFS), Web Map Service (WMS), and Web Coverage Service (WCS) that we use to share spatial data and to create web maps;

Figure 3: RGB mosaic system



Source: copyright own by the authors.

- Rasdaman: RASter DAta MANager allows us to store and to query massive multi-dimensional arrays, such as sensors, image, simulations, and statistics data appearing in domains like earth, space, and life science (WCS, WCPS & WMS);
- LoRa Server for IoT applications: provides open-source components for building LoRaWAN networks; together they form a ready-to-use solution for IoT applications;
- SOS: the 52°North Sensor Observation Service provides an interoperable web-based interface for inserting and querying sensor data and sensor descriptions.

2.2.3 Databases

- PostgreSQL: is a powerful, open source object-relational database system;
- Influxdb: is a powerful database to store and manage time-series.

2.2.4 Other applications

- Shiny server: is a server program that makes Shiny (an R package) applications available over the web;
- Jupyter Notebook: is an open-source web application that allows you to create and share documents that contain live code, equations, visualizations and narrative text;
- Data Exchange Server (DES): performs different tasks in the domain of data handling and pre-processing in order to automatize easily processes.

3 Pilot cases

In this section, we describe three applications developed in specific projects that use EDP's components. We describe them to give an idea about what is possible to achieve with the EDP.

3.1 RGB system

RGB automatic system is an application set-up in the contest of the **Sentinel Alpine Observatory** (SAO), which aim to develop and provide **satellite products and services**.

The output of the RGB application is an RGB image mosaic, that cover all the Alps, updated every day with the new tiles downloaded where cloud coverage less than 20% and no-data less than 30% in the single tile.

The result is accessible by a WMS layer, created using Geoserver application, or can be visualized in a simple interactive web map that we prepared.

Every tile in the map (figure 3) has information about acquisition time and cloud coverage both in the metadata, and in a separated vector layer, which represent bounding box of the tiles.

The tool uses Python codes and the DES application in order to automatize download, pre-processing, processing of the tiles and to update the group-layer's store and metadata in Geoserver. The vector layer, the tiles grid visible in figure 3, is stored using the PostGis extension of PostreSQL DB and the tile information are easily updated with a SQL query, using the HTTP post method.

3.2 ExcEED

According to the International Energy Agency, buildings currently account for 40% of primary energy consumption in most countries and are a significant source of carbon dioxide emission.

The European Union is undertaking consistent action to enforce and promote energy efficiency in building sector. The new energy efficient constructions and the deeply renovated buildings are the expression of the European efforts to decrease CO_2 emissions. Despite that, several issues are still uncovered.

The actual policy that regulate the design of new buildings and retrofit are not truly effective, the energy consumption of the building is generally not in line with what has been foreseen in the design phase as well as the indoor comfort are rarely considered in the evaluation of building quality. Moreover, there are common mistakes about how the buildings are retrofitted or built not evaluating in continuous their performance.

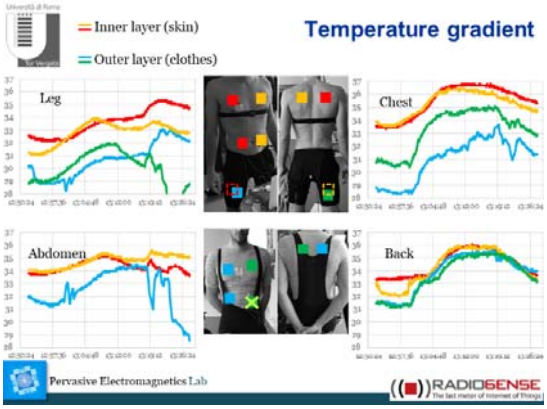
The solution to these problems lies in the analysis of the real performance of new and retrofitted buildings and the comparison to design and rating criteria. Indeed, it is necessary to know in which point of the building construction methodology we are, to improve the quality and the performances of the future building sector.

The European Energy Efficient building & district Database (ExcEED) project aims to create a solid and well-structured database with measured and qualitative information from beyond the state of the art buildings. ExcEED is the first implementation of a comprehensive database of design and performance data for state of the art buildings in Europe. To reach this aim several aspects have been defined. As first step, a collection of tailored Key Performance Indicators that will "transform data into information" have been created and saved in the database. Detailed metadata has been collected to gather the most important information of the building. Thereafter, a set of tools that allows geo-clustered, statistical, and knowledge analysis of the data has been set up. Thanks to all of them, it was possible to benchmark different buildings both on energy and comfort points of view. The tools have been included in a suitable dashboard that is able to get the data manually via both ftp connection or uploading *csv* file and connecting with the Building Management System (BMS) of the building.

Our work focalized on the development of Geo-cluster tool. The latter is a web application of the EDP that allows to visualize data and metadata collected as well as to have a first analysis of buildings performances. The web application is developed using Shiny-Server that provides several libraries to analyse and plots different results. Moreover, the shiny application can be easily customized integrating custom R code.

[*This work is part of the research activities of the project ExcEED, funded by the European Union's Horizon 2020 research and innovation programme under grant agreement No 723858*]

Figure 4: sensor configuration in different layers



Source: C. Miozzi, S.Amendola, G.Marrocco. Tor Vergata University of Rome.

3.3 BICI

The performance of athletes doing outdoor sport activities, such as cycling, is highly affected by the thermo-hygrometric comfort, which depends on environmental and individual factors, such as geographical characteristics, season of the year, actual weather condition, physical and physiological state, and very important, the characteristics of personal garment that should protect the individual athlete. Within the BIcycle Clothing and thermal comfort estImation project (BICI), in collaboration with the company Q36.5, we used a multi-sensor approach to estimate thermo-hygrometric comfort of new textiles and designs in cycling clothes. We developed a web application to analyze the collected data that infers individual comfort zones. We integrate this datasets with some other sources of information such as weather conditions, thermal imagery to detect hot spots and distribution of water/sweat saturation and computer bike data, where we track the route and power conditions of the route (Multi-sensor approach to estimate the thermo-hygrometric comfort using new textiles and designs in cycling clothes, 2018).

The main objective is to find the balance between the level of temperature and humidity over the clothing that the user wears, in order to tolerate both environmental and physiological conditions with the feeling of being comfortable.

The web application, developed using Shiny, provides a user interface to upload and analyze the time-series using plots. The database InfluxDB stores the time-series and the metadata (date, hour, athlete, location, dresses, etc.) of every test.Conclusions

4 Conclusion

The spatial data infrastructure, called Environmental Data Platform (EDP), represents a reproducible example for research organizations or private companies that need a working environment for big data. It allows organization, discovery, processing, analysis and sharing of data. In particular, during the analysis phase the platform provides data access through web services to save time by querying only the subsets of interest and solving the problem of data format conversion. The solution not only suggests open source software to use in a private infrastructure but also suggests a workflow for big data management to increase simplicity and performances of researcher's daily work.

We constantly promote open source solution and foster adoption of OGC standards, to increase interoperability and simplify collaboration creating larger research networks. Customization of applications' code give us the possibility to satisfy specific project requirements.

Participation in the H2020 OpenEO project is helping us to improve data access service by integrating the OpenEO API in the EDP with the vision that this will improve efficiency in the daily work of researchers.

References

Bergamini, C., Bosi, F., Corradi, A., De Rolt, C., Foschini, L., Monti, S., Seralessandri M. (2018) *LocalFocus: A Big Data Service Platform for Local Communities and Smarter Cities.* IEEE Communications Magazine. DOI: 10.1109/MCOM.2018.1700597

De Lathouwer, B., Jackson, M., McKee, L. (2015) OGC Information Technology Standards for Sustainable Development. OGC White Papers available in: https://www.opengeospatial.org/docs/whitepapers

El Arass, M., Tikito, I., Souissi, N. (2017) Data lifecycles analysis: Towards intelligent cycle. In Proceedings of Intelligent Systems and Computer Vision conference, Fez, 2017. DOI: 10.1109/ISACV.2017.8054938.

Jacob, A., Vicente-Guijalba, F., Kristen, H., Notarnicola, C., Costa, A., Ventura, B., Monsorno, R. (2017) *Organizing Access to Complex Multi-Dimensional Data: An Example from the ESA SEOM SInCohMap Project*. In: Proceedings of the Big Data From Space conference, Toulouse, 2017.

Mathieu, P., Aubrecht, C. (2018) *Earth Observation Open Science and Innovation.* ISSI Scientific Report Series, volume 15 – Springer Nature. DOI: https://doi.org/10.1007/978-3-319-65633-5.

Mejia-Aguilar A., Monsorno R., Cazzaro and Bergamo L. *Journal of Physics:* Conference Proceedings Series (2018) 1065 122022. DOI: 10.1088/1742-6596/1065/12/122022.