
# Proceedings of the 2017 conference on

# Big Data from Space

## (BiDS'17)

28th-30th November 2017
Toulouse (France)

*Edited by P. Soille and P.G. Marchetti*



Joint Research Centre

EUR 28783 EN

**Contact information**
Name: Pierre Soille
Address: European Commission, Joint Research Centre
Via Enrico Fermi 2749, TP 267, I-21027 Ispra (VA), Italy
Email: Pierre.Soille@ec.europa.eu
Tel.: +39 0332 78 9111

**JRC Science Hub**
https://ec.europa.eu/jrc

JRC108361

EUR 28783 EN

PDF ISBN 978-92-79-73527-1 ISSN 1831-9424 doi:10.2760/383579

Luxembourg: Publications Office of the European Union, 2017



How to cite these proceedings: P. Soille and P.G. Marchetti (Eds.), *Proceedings of the 2017 conference on Big Data from Space. BIDS' 2017*, EUR 28783 EN, Publications Office of the European Union, Luxembourg, 2017, ISBN 978-92-79-73527-1, doi:10.2760/383579, JRC108361

# EO4WILDLIFE: A CLOUD PLATFORM TO EXPLOIT SATELLITE DATA FOR ANIMAL PROTECTION

*Fabien Castel[1], Gianluca Correndo[2], Alan F. Rees[3]*

[1]Atos Integration, Toulouse France

[2] University of Southampton, IT Innovation Centre, Southampton United Kingdom

[3] University of Exeter, Penryn, UK

## ABSTRACT

EO4wildlife brings large number of multidisciplinary scientists such as marine biologists, ecologists and ornithologists to collaborate closely together while using European Sentinel Copernicus earth observation data more efficiently on a platform available over Internet dedicated to environment study and animal protection [1]. A comprehensive set of processing services and data connectors are available on the platform providing to scientists powerful tools to build innovative animal protection applications.

**Index Terms—** Copernicus, Platform as a service, big-data, cloud computing, earth observation data, data analytics, animal protection

## 1. INTRODUCTION

All the new sets of data provided by Copernicus satellites open up the way for hundreds of innovative scenarios to combine animal tracking data with remotely-sensed earth observation data. In order to reach such important capabilities, an open service platform and interoperable toolbox supported by a scalable cloud infrastructure are being designed and implemented. It offers high level data processing services. The platform front end will offer dedicated services that will enable scientists to connect with several animal tracking databases, access large data collections from Copernicus satellites, sample relevant environmental indicators, and finally run environmental models and simulations using these big data sources in a scalable processing environment.

The research is leading to the development of web-enabled service compliant to OGC2 (Open Geospatial Consortium) enabling data interoperability in geospatial data access and processing services.

## 2. PROBLEMATIC

Scientists can use the huge Copernicus datasets for various purposes. Mainly, they aim at identifying the key environmental factors that drive the distributions of animals. By building predictive models the goal is to improve management and decision-making about animal protection. Model results are extrapolated, in line with various climate change scenarios, to determine likely future population distributions and aid understanding of how environmental conditions may alter phenology and demographic processes.

Exploiting these rich datasets is a challenge for scientists. A wide diversity of products are available through different systems, platforms and interfaces, but this profusion of options can be overwhelming and scientists do not always have the technical capabilities to access these sources and to process the downloaded data. That's why the EO4wildlife platform aims at providing a quick and easy access to a comprehensive set of EO datasets, as well as a toolbox of services for data filtering, processing and visualization.

## 3. CLOUD APPLICATION

Generally, cloud application stands for applications deployed over Internet with flexible pay-as-you-use infrastructure, "big-data" storage and scalable web services. The EO4wildlife platform perfectly fit with this description.

In particular, several layers can be distinguished when dealing with applications on the cloud. The Infrastructure as a Service layer – IaaS – provides flexibility and efficiency in order to easily scale out processing and storage capabilities. The Platform as a Service layer – PaaS – deals with applicative components deployment, resource management and security issues. The Software as a Service layer – SaaS – allows user to access the required data and to configure and run the service available on the platform.

There are many benefits with such an approach. Besides the economical and practical aspects, there is a strong incentive for sharing. On the platform, everyone can be both a producer and a consumer, and discover new opportunities within the community. EO4wildlife offers a catalogue of resources and added value services that is continuously enriched as new members join and contribute to the ecosystem.



doi: 10.2760/383579

## 4. DATA ACCESS

### 4.1. Tracking Data

EO4wildlife aims at being connected to existing platforms where scientists host their data. Initially the Seabird (http://seabirdtracking.org/) and the Seaturtle (http://seaturtle.org/) tracking databases are targeted.
The Seabird Tracking Database aggregates data from more than 150 contributors to provide the largest collection of seabird tracking data in existence. In total, the Seabird Tracking Database holds information for 114 species in more than 11 million locations, corresponding to more than 20 thousand tracks. It serves as a central store for seabird tracking data from around the world and aims to help further seabird conservation work and support the tracking community.
The seaturtle.org platform hosts Argos tracking data for all seven species of sea turtle and around 70 other animal species that include cetaceans, pinnipeds, elasmobranchs and birds located around the globe. The platform hosts data for over 1000 tracking projects and has amassed over 14 million data points.

### 4.2. Environmental Data

The EO4wildlife platform provides connectors to access data from various data sources, such as the CMEMS [2] catalogue, indexing hundreds of ocean-related EO products, or the AVISO catalogue for the altimetry domain. An internal EO4wildlife data catalogue is maintained to aggregate products from all these external sources.
Environmental datasets generally are voluminous. The Figure 1 is an example of a dataset that is used on the platform. This dataset provides data on sea surface temperature and sea ice concentration during a 10 year period from 2007 to 2017 for a global geographic coverage. The overall size for the dataset is near 0.5 Terabytes. The platform aims at targeting tens of datasets similar to this one.

## 5. PROCESSING SERVICES

The platform hosts a series of basic data analytics services to enable scientists to implement analytic workflows [3]. These services can be divided in three main categories.

### 5.1. Pre-processing and Aggregation

The first one is **data pre-processing and aggregation.** It includes the pre-processing, cleaning and aggregation of the data prior to the analytical step. The pre-processing of geospatial data sets is an important step when dealing with potentially imprecise information such as animal positions. These services allow to recognize and to eliminate all data elements which are clearly unrealistic considering the knowledge of the domain. (E.g. the animal is not capable of travelling at such velocities). Moreover, the pre-processing services allow filling missing data values and accommodate different data grids by interpolating values which were not directly collected and represented in the data sets. Aggregation services reconcile data represented with different spatial or temporal resolutions providing functionalities to sample environmental observations and aggregate them in the right granularity to fuel niche modelling algorithms. In this category are also included services to process animal tracks, to provide grouping of tracks in trips or gridding a number of tracks to study the population distribution.

| Global Ocean OSTIA Sea Surface Temperature and Sea Ice Analysis | | | |
|---|---|---|---|
| Variables | size/day (MB) | size/year (GB) | total size (GB) |
| analyzed_sst | 50 | 18,25 | 196,55 |
| sea_ice_fraction | 25 | 9,125 | 98,275 |
| analysis_error | 50 | 18,25 | 196,55 |
| Whole product | 125 | 45,625 | **491,375** |

**Figure 1: EO Dataset volume**

### 5.2. Data Mining

The second category is data mining and contains services processing animal tracks and satellite marine observations in order to model animals' use of space and correlate this information with available environmental observations. This category is further subdivided in two sub-categories of services: animal tracks based services and statistical environmental services.
Animal tracks based services analyse the tracks alone in order to estimate the animals' home range and the foraging grounds.
Statistical environmental services assess the statistical relevance of environmental observations in modelling animals' presence and implement environmental niche models (ENM) to understand the marine species' habitats.
Data mining techniques have been implemented in different scenarios to model the preferred animals' habitat following a literature review for each marine species. At the moment these modelling techniques include:

- Environmental Envelope Model (EEM)
- Generalised Additive Model (GAM)
- Generalised Linear Model (GLM)
- Boosted Regression Trees (BRT)
- Random Forests (RF)

These services support scientists in modeling marine animal niches by means of environmental observations which can then be used in creating projections of animal presences in the future (see Figure 2 for an example of niche model produced using RF).

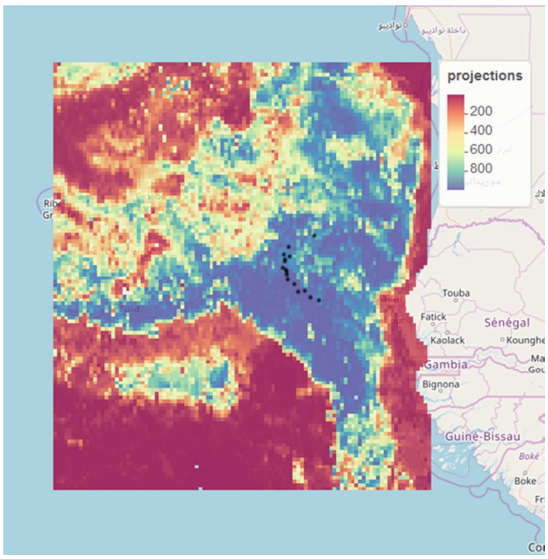


**Figure 2: Marine turtle habitat modelled using RF near Cape Verde (August 2009)**

### 5.3. High Level Fusion Services

The last category contains **high level fusion services.** These services make use of multiple data sources to better estimate animals' position, behavior and modelling animals' habitats. This category includes the Track & Loc service which enables the estimation of submarine trajectories for animals equipped with pop-up or archival tags [6].

## 6. IMPLEMENTATION

### 6.1. Architecture Overview

The platform is composed of several functional components. An internal data catalogue aggregates georeferenced products metadata from various external sources. An ingestion component allows retrieving this data on-demand for exploitation by the platform services. The service manager component allows developers to manage the lifecycle and the execution of their services. At the end of the chain, EO4wildlife makes available built-in visualization features for standard geographic data (OGC WMS/WFS standards) produced by the services.

The service management mechanism is built on the containerization concept (i.e. Docker [4]). By encapsulating each service into an independent and self-sufficient container, the platform ensures total freedom for the service developers (preventing language, framework or libraries constraints) and an easy portability on the cloud. An orchestration technology (i.e. Kubernetes [5]) is used to manage container life cycle so that the underlying infrastructure becomes totally transparent. This technical architecture based on standards aims at creating a collaboration space for scientists in term of data and service sharing

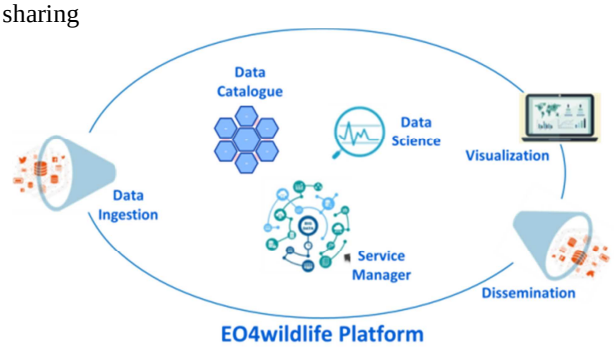


**Figure 3: Functional view**

### 6.2. Data Management

The ambition of EO4wildlife is to grant scientists easy access to tens of earth observation datasets, thus dealing with terabytes of data (see 4.2). Instead of hosting permanently all the data to the platform, a repository is set up and acts as a local cache from the remote data warehouse. Data is day by day downloaded following platform user's requests and temporarily stored on the local cache. This cache is common to all users. When providing data to a service, the platform deals with the merging of these daily files in order to have one global file corresponding to what was requested for the service execution. Finally, a periodical purge mechanism ensures that the cache size does not overreach the local disk size.

### 6.3. Interoperability

Integrating smoothly the EO4wildlife platform into the existing ecosystems of animal monitoring applications is a key element for scientists. EO4wildlife was designed to complete existing systems. For this purpose, it provides convenient interface for external application to upload data into the platform. Indeed the internal module in charge of the file management on the server side exposes a REST web service enabling any external application (Seabird and Seaturtle for instance) to upload data to the platform.

Data exchange between different platforms can be challenging because every tracking database has its own format, generally csv-like text files. To address this issue, the EO4wildlife project defined an XML format specific to the animal tracking domain. Any tracking data can be converted to this format using a configurable CSV to XML converter. Configurations for the Seabird and Seaturtle file format are already deployed on the platform. All the services offered by the platform should use this standard format. Besides an improvement on services coherence and

versatility, the standard self-described format has other advantages. Date information becomes much more exploitable, for instance for automatic animation over time of the tracking points and automatic extraction of spatial coordinates is possible when fusing tracking and earth observation data.

### 6.4. Data Visualization

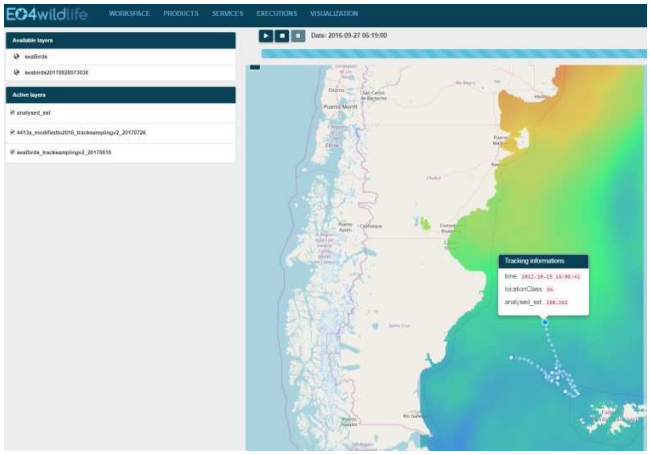


**Figure 4: Visualization panel screenshot**

The platform provides a visualization panel (see Figure 3) for tracking and earth observation data. This feature uses the OGC web standards for geographic data representation and a frontend panel based on multi layers visualization. Tracking data are handled as WFS [7] layers and earth observation product as WMS [8] layers. All the data used as inputs and/or outputs of processing services can be automatically transformed into displayable layers that can be overlaid. All the layers holding a time dimension can also be animated over time, which allows the highlighting of the key environmental factors on animal migration.

## 7. CONCLUSION & NEXT STEPS

This paper presents the services that the EO4wildlife platform can offer to the scientific community in order to develop innovative use cases based on the new generation of earth observation data. The developments of the platform will continue until 2018, developing new features following the needs of the project scientific partners and securing what exists to prepare the opening to a larger public.

## 8. ACKNOWLEDGEMENT

This work is partly funded by the European Commission under H2020 Grant Agreement number: 687275.

## 9. REFERENCES

[1] Zoheir Sabeur, Gianluca Correndo, Galina Veres, Banafshe Arbab-Zavar, Jose Lorenzo, Tarek Habib, Anne Haugommard, Fanny Martin, Jean-Michel Zigna, and Garance Weller. 2017. EO Big Data Connectors and Analytics for Understanding the Effects of Climate Change on Migratory Trends of Marine Wildlife. Springer International Publishing.

[2] CMEMS (Copernicus Marine Environment monitoring service) public website : http://marine.copernicus.eu/

[3] Z. Sabeur, G. Correndo, G. Veres, B. Arbab-Zavar, G. Ivall T. Neumann, F. Castel, J-M. Zigna, J. Lorenzo. (2017) EO Big Data Analytics for the Discovery of New Trends of Marine Species Habitats in a Changing Global Climate. 2017 Conference on Big Data from Space (BiDS'17)

[4] R. Peinl, F. Holzschuher, F. Pfitzer, Docker cluster management for the cloud –survey results and own solution, J. Grid Comput. (2016) 1–18.

[5] D. Bernstein, Containers and cloud: from LXC to Docker to Kubernetes, IEEE Cloud Comput. 1 (3) (2014) 81–84.

[6] Royer F, Lutcavage M (2009) Positioning pelagic fish from sunrise and sunset times: error assessment and improvement through constrained, robust modeling. In: Neilson JD, Smith S, Royer F, Paul SD, Porter JM, Lutcavage M (eds) Tagging and tracking of marine animals with electronic devices. Springer, Amsterdam, p 323−341

[7] OGC Web Feature Service standard format description http://www.opengeospatial.org/standards/wfs

[8] OGC Web Map Service standard format description http://www.opengeospatial.org/standards/wms


doi: 10.2760/383579