# Deliverable 2.5 (D2.5)

# EU BON portal

# M51

| | |
|---|---|
| Project acronym: | EU BON |
| Project name: | EU BON: Building the European Biodiversity Observation Network |
| Call: | ENV.2012.6.2-2 |
| Grant agreement: | 308454 |
| Project duration: | 01/12/2012 – 31/05/2017 (54 months) |
| Co-ordinator: | MfN, Museum für Naturkunde - Leibniz Institute for Research on Evolution and Biodiversity, Germany |
| Delivery date from Annex I: | M51 (February 2017) |
| Actual delivery date: | M51 (February 2017) |
| Lead beneficiary: | CSIC |
| Authors: | Antonio García Camacho (CSIC, Doñana Biological Station – Spanish National Research Council, Spain)<br>Hannu Saarenmaa (UEF, University of Eastern Finland, Digitisation Centre, Finland)<br>Tim Robertson, Kyle Braak (GBIF, Global Biodiversity Information Facility, Denmark)<br>Israel Peer (GlueCAD Ltd. – Engineering IT, Israel)<br>Veljo Runnel (UTARTU, University of Tartu, Natural History Museum, Estonia)<br>Lyubomir Penev, Teodor Georgiev, Pavel Stoev (Pensoft, Bulgaria)<br>Donat Agosti (Plazi, Switzerland) |

**This project is supported by funding from the specific programme 'Cooperation', theme 'Environment (including Climate Change)' under the 7th Research Framework Programme of the European Union**

**Dissemination Level**

| | | |
|---|---|---|
| PU | Public | ✓ |
| PP | Restricted to other programme participants (including the Commission Services) | |
| RE | Restricted to a group specified by the consortium (including the Commission Services) | |
| CO | Confidential, only for members of the consortium (including the Commission Services) | |

*This project has received funding from the European Union's Seventh Programme for research, technological development and demonstration under grant agreement No 308454.*



# Executive Summary


## Introduction

The goal of the deliverable is the release of the European Biodiversity Portal, along with the set of tools and web applications developed under the umbrella of the WP2 and in particular its task 2.5.

The deliverable details each portal section and summarises the architectural and design decisions considered, evaluated and applied during the software development process.

## Progress towards objectives

The objectives of EU BON work package 2 "Data Integration and Interoperability" are the following:

1. Establish an information architecture for the EU BON project that will be compatible with the global GEO BON, the INSPIRE directive, other European projects, and the LifeWatch research infrastructure.
2. Develop data integration and interoperability between the various networks, and, with the new generation of data sharing tools, enhance linking between observational data, ecosystem monitoring data, and remote sensing data.
3. Develop new web service interfaces for data holdings using state-of-the-art standards and protocols. Register the networks on the GEOSS Common Infrastructure (GCI) using harmonised metadata.
4. Develop a new portal to enable fast access to EU BON integrated data and products by researchers, decision makers and other stakeholders.
5. Ensure global coordination of development efforts through an international data interoperability task force and adoption of the results through helpdesk and a comprehensive training programme.

The current deliverable presents a review of the portal, but covering in addition the results of the efforts to accomplish all the objectives proposed, already covered by the deliverables D2.1, D2.2, D2.3 and D2.4.

## Achievements and current status

The document describes the final status of the portal.

## Future developments

Further improvement in the portal and tools may be applied, as part of the sustainability plans of the project. In addition, given that GI-cat is part of the core of the GEO Discovery and Access Broker (GEO DAB), the EU BON Metadata Broker may be replaced by a view in the GEO DAB, depending on the incorporation of EU BON's data providers as GEO providers, in the same terms that EU BON harvests their datasets.

# Table of contents

# 1. Introduction

## 1.1. Background and description of the work

The Description of Work (DoW) depicts the European Biodiversity Portal as one of the primordial products to be generated under the umbrella of the EU BON project, setting its importance as a gateway not only to data and metadata, but also to the set of tools, results, factsheets and other outcomes generated by the EU BON consortium.

The Task 2.5 (European Biodiversity Portal) is defined as follows in the Description of Work (DoW).

> *A European Biodiversity Portal (EBP) will be developed as the main GEO BON information hub. It will link to relevant databases and information systems, policy contacts and recommendations, and structured advice for assessing relevant distributed information/datasets for different user groups, including contributions from citizen science data gathering gateways. The EBP will technically integrate the various data sources under one search facility and spatially/temporally oriented user interface. The portal will build on the tools developed by task 2.3, functions developed by task 2.4. It will provide access to full detailed data, geographic visualisation and remotely sensed data. It will be closely linked to the GCI and GEO Portal, and access layers and data from GEOSS sources. The portal would also act as showcase for the products from the analytical and modelling activities of other WPs and support workflows for building such products using the registered e-services. The portal will also serve general dissemination functions for WP8. (Lead CSIC; UEF, GBIF, UnivLeeds, Pensoft, FIN, Plazi, GlueCAD, NBIC; Months 1-54)*

Encompassing the needs of heterogeneous potential user groups is not a trivial matter, given that the portal itself should accomplish the needs of public users, decision makers and scientists as well. Due to this fact, the development process has been based on an evolutionary methodology. Throughout the Task 2.5, several prototypes were generated, finally moving to an agile cycle, which bases its functioning in releasing changes and new components as soon as they have been developed, presenting them to the stakeholders and elucidating new overall requirements.

Related milestones include **MS251** "Specification for the EU BON portal" (Month 27), **MS252** "Technical workshop, launch of registry and beta portal" (Month 39) and **MS253** "EU BON portal test release" (Month 41). Additionally, all the milestones of the Task 2.7 which document the advice of the international advisory group of EU BON for informatics work and the interim review have been instrumental in determining the requirement.

An initial version of the portal was developed by CSIC and presented during the EU BON Citizen Science Round Table, 27 November 2014. This version or first functional prototype was centred on the integration of LTER and GBIF dataset discovery, providing a single search for retrieving metadata from both biodiversity data sources. Additional providers could be added by means of configuring an OGC Catalogue Service for the Web (OGC-CSW) interface in the portal administration section.

The milestone MS251 set the grounds of the portal, describing the specifications and requirements of the final version of the portal. Nevertheless, during the EU BON General Meeting in Cambridge (1-4 June 2015), both prototype and requirements document were discussed, coming to an end with a set of amendments to apply to the portal requirements. The EU BON consortium concluded that the portal should be focused on acting as a showcase of EU BON results and not only to the datasets, given the plethora of tools and products developed and maintained under the umbrella of the project, and their potential as an enforcement of the biodiversity conservation in Europe.

Taking into consideration this approach, EU BON decided to move the European Biodiversity Portal to an Enterprise Content Management System, considering Drupal and Liferay as alternatives and

finally choosing the latter. Liferay enables the users to create not only a portal but a set of portals or sites in the same Liferay installation, thus enabling EU BON to create sites for specific testing sites or for particular groups, e.g. the Citizen Science site.

CSIC has designed the user interface of the portal, following an initial design proposed by Pensoft. During a workshop held in Seville in January 2016 portal's design and interface were discussed and approved. The domain names chosen during the workshop were http://biodiversity.eubon.eu and http://portal.eubon.eu.

The recommendations of the EU BON International Advisory Group had also a direct influence on the development on the portal, in particular in regards to the user interface. Taking in consideration the corrections suggested by the advisory group during the meeting held on 29 November 2016, the development team decided to upgrade from Liferay 6.2 to Liferay 7.0 and polish the portal interface, taking into consideration the usability of the system, and also improving the overall design.

## 1.2. Specifications traceability

The Milestone MS251 established the specifications of the portal (**Table 1**), later refined during the agile software development process. **Table 2** summarises the implementations of each specification in the final product.

**Table 1.** Summary of the specifications of the portal (MS251)

| Specification Code | Description |
|---|---|
| **SPEC-01** | Provide a filtered search user interface |
| **SPEC-02** | Integrate biodiversity networks/providers/test sites by metadata |
| **SPEC-03** | Supply heterogeneous biodiversity related data |
| **SPEC-04** | Integrate the EU BON Taxonomic backbone |
| **SPEC-05** | Provide an interface to GEOSS |
| **SPEC-06** | Provide an interface to DataONE |
| **SPEC-07** | Provide geospatial capabilities for filtering and visualisation |
| **SPEC-08** | Provide different user interfaces for different user roles |
| **SPEC-09** | Web content management |

**Table 2.** Specification - Implementation traceability

| Specification Code | Implementation |
|---|---|
| **SPEC-01** | · Dataset search interface.<br>· Analysis tool: Spatial Dataset Browser<br>· Analysis tool: Species Richness application<br>· Analysis tool: Species Population trends visualisation |
| **SPEC-02** | · GEOSS GI-cat instance deployed and configured<br>· Biodiversity networks and test sites integrated in GI-cat.<br>· Metadata internally harmonised to the ISO 19115/19139 metadata model (INSPIRE compliant). |
| **SPEC-03** | · Dataset search UI |
| **SPEC-04** | · Taxonomic backbone called from the Dataset Search interface to correct taxa filters |
| **SPEC-05** | · GI-cat output profiles provide external network to retrieve/harvest biodiversity datasets: GI-cat service connector, OpenSearch API, |

| | |
|---|---|
| | OGC CSW. |
| **SPEC-06** | · GI-cat output profiles provide external network to retrieve/harvest biodiversity datasets: OpenSearch API, OGC CSW. |
| **SPEC-07** | · Dataset search geospatial filter.<br>· Analysis tool: Spatial Dataset Browser<br>· Analysis tool: Species Richness application<br>· Analysis tool: Species Population trends visualisation<br>· Analysis tool: Business Intelligence dashboards, filter by country, georeferenced occurrence points.<br>· Geospatial layer representation portlet (test sites). |
| **SPEC-08** | · Portal visualisation configurable by means of Liferay authorization and accounting system.<br>· Several user groups and roles created, e.g., product list administrators, citizen science user group. |
| **SPEC-09** | · Liferay offers an advanced Enterprise Web Content Management System.<br>· Users can administer their own sites, content or components. |

# 2. Overview of the European Biodiversity Portal

The European Biodiversity Portal is a multisite web portal based on Liferay 7.0. Specific components have been developed using JavaScript frameworks and templating languages, such as Freemarker and Velocity. The overall design follows the Material Design guidelines, which help to provide a flat, modern and responsive design (**Figure 1**).


**Figure 1.** EU BON Biodiversity Portal - Home page

Not only should the European Biodiversity Portal act as a showcase of EU BON results, but also help data-agnostic users to discover, share and retrieve data. Taking this into account, the portal navigation was envisaged to follow a natural data workflow in Biodiversity:

Discover and retrieve data → Analyse data → Share data

→ Generate data products → Share data products

## 2.1 Data section

The data section provides the user with an interface for retrieving datasets from the data providers linked or harvested by the EU BON Metadata Broker, an instance of GEOSS GI-cat. The dataset search tool corrects the taxonomy constraints entered by the users, requesting data to the EU BON Taxonomic Backbone[1]. The user can enter taxa by vernacular or scientific name, exactly or approximately, or by the identifier assigned by the data provider. After performing a search, a list with the previously-harvested datasets will be listed, including links to the original metadata and their geographic coverage (**Figure 2**).

[1] https://cybertaxonomy.eu/eu-bon/utis/

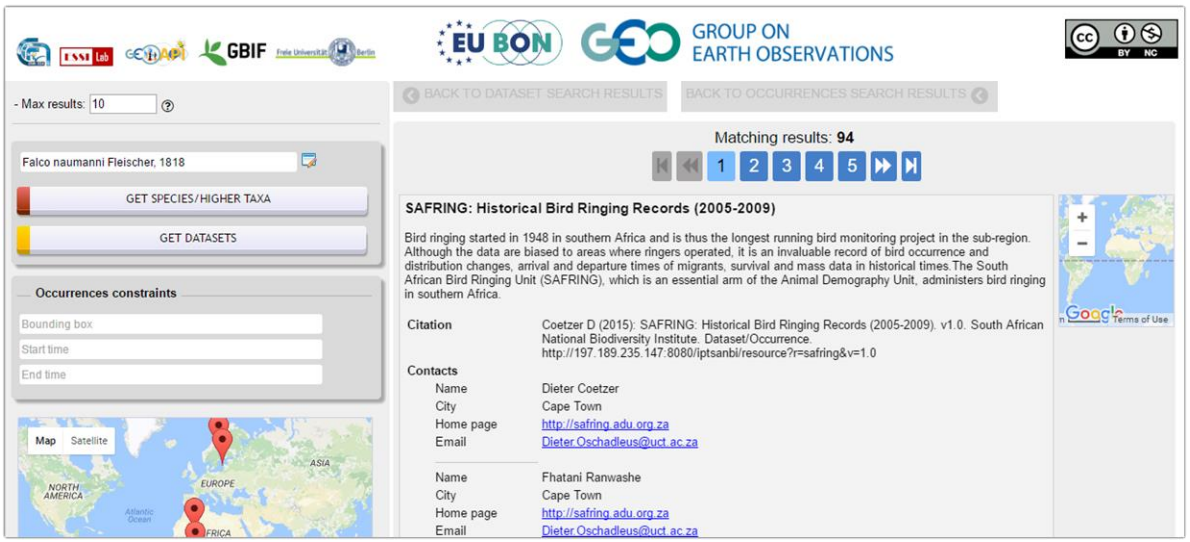


**Figure 2.** Dataset discovery

## 2.2 Analysis section

This section lists a set of tools for data analysing and knowledge discovery. Given that GBIF provides access not only to metadata but also to harmonised data, all analytical tools use GBIF as data backbone, either by accessing its APIs or by retrieving data from a periodical snapshot of the GBIF occurrences across Europe.

These are the tools put at the users' disposal:

- Spatial Dataset Brower. See section 5.1.
- Species Richness Application. See section 5.2.
- Species Population Trends Visualisation. See section 5.3.
- Business Intelligence Dashboard. See section 5.4.

## 2.3 Services section

This section offers a set of tools for helping users to share data. It also provides a collection of tools, services and data layers helpful in order to fulfil a proper data analysis.

- Tools for data sharing. An inventory and assessment of software tools available for biodiversity data sharing and data publishing.
- Environmental layers. A wiki with a set of gridded environmental data layers, which have been selected by the EU BON scientists for their usefulness for data analysis and modelling.
- AquaMaps for EU BON, a joint project of EU BON, FishBase and SealifeBase, consisting of standardised distribution maps for over 1,032 European fishes and marine mammals.
- EU BON IPT. Data hosting service provided by the EU BON partner GBIF using their Integrated Publishing Toolkit (IPT).
- EU BON Taxonomic Backbone (**Figure 3**). The EU BON taxonomic backbone or Unified Taxonomic Information Service (UTIS) allows running a federated search on multiple European checklists and returns a unified result set of the individual responses of the various

checklists. It integrates a set of taxa providers, accessing them through REST APIs or WSDL/SOAP web services.

The backbone integrates the Pan-European Species directories Infrastructure (EU-Nomen), EUNIS which fully covers Natura 2000, thus it is fully compliant with the Appendix 3 of the INSPIRE directive. Additionally, it retrieves taxa from the Catalogue of Life (CoL), the World Register of Marine Species (WoRMS), the GBIF Checklist Bank and the Plazi TreatmentBank. It offers a single REST API for retrieving taxa from all taxonomic providers connected.

– EU BON Helpdesk. Advice, assistance and training in using EU BON tools, products and services. Provided by the EU BON partner Museum of Central Africa, in Tervuren, Belgium.

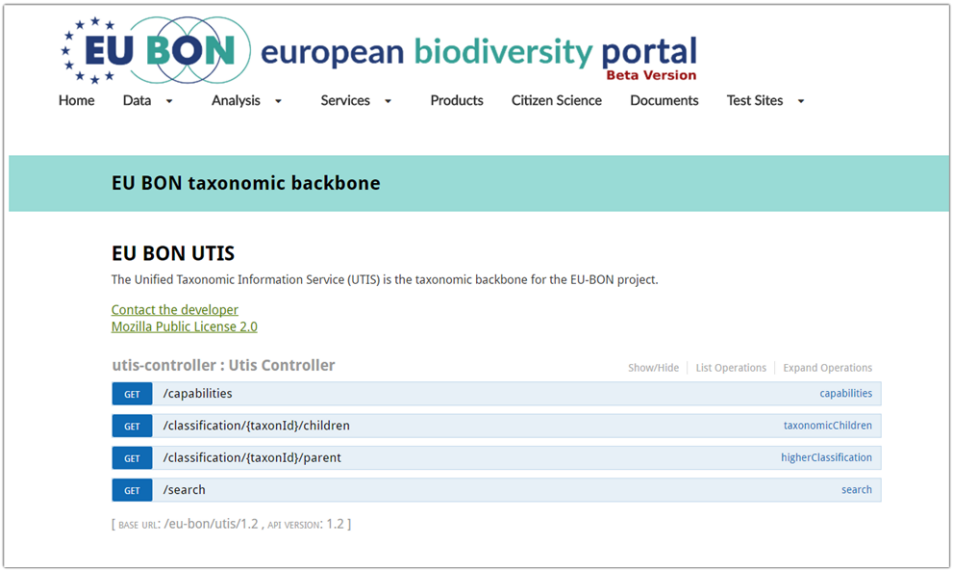


**Figure 3.** Taxonomic Backbone API

## 2.4 Products section

This section provides the user with an interface for identifying EU BON products from a set of more than forty elements, either tools, factsheets or other products (**Figure 4**). The list can be filtered by audience and by topic.

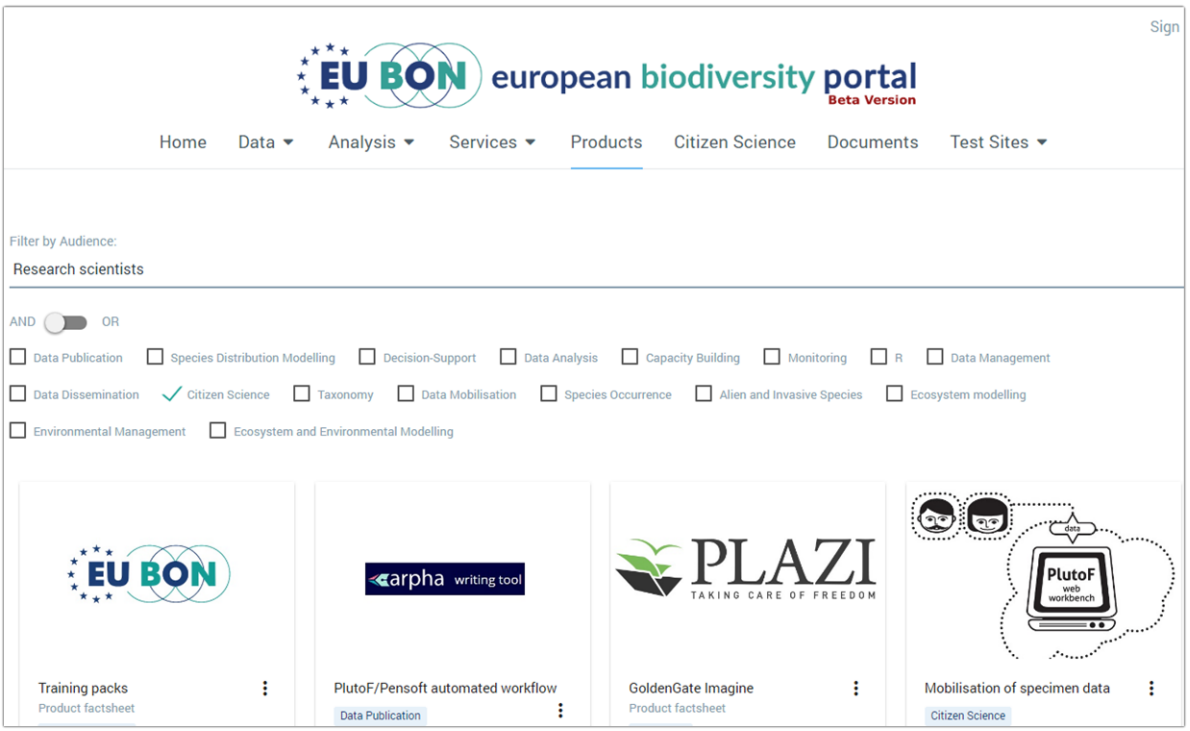


**Figure 4**. EU BON Products

# 3. Software architecture and components

The Deliverable **D2.1**, "Architectural design, review and guidelines", established an initial architecture of the portal, based on the service-oriented architecture[2] implemented through web services and a Java EE 6 web development stack. This initially laid out architectural approach was applied in the development of the first iteration of the development process, releasing the later established as EU BON European Biodiversity Portal, first prototype. This architecture was deprecated in further versions, due to the transition to an Enterprise Content Management System, the use of GI-cat as main broker/metadata registry and also to the transition to JavaScript-based tools.

Moreover, the D2.1 depicted the overall need of service orchestration within the core of the service-oriented architectural implementation, by means of deploying an Enterprise Services Bus, a core component which would help to implement such service collaboration as business processes, as well as to ease the connection of heterogeneous data sources. The final design of the portal does not base its core in an ESB, due to the following design decisions:

- The GEOSS GI-cat, as part of the GEO Discovery and Access Broker, already implements the business processes, orchestration and translation to a common business data model, based on the ISO 19115 data model.
- The Taxonomic Backbone deals internally with the connection to heterogeneous data sources, the service mediation and data translation to a common model, offering a JSON/XML REST API, thus an additional orchestration was not needed.

2 D2.1, EU BON Recommendation 9: "EU BON should adopt the Service Oriented Architecture model."

- The ECMS used for implementing the portal already offers mediation services and service orchestration, based on internal Enterprise Service Bus, mediation or workflow engines, such as Activiti BPM or ServiceMix. They could be used if complex process implementation were needed.

The ECMS chosen for building the portal and their child sites is Liferay, either 6.2 till the beta version or Liferay 7.0, in the pre-final and final version. It was chosen among other CMS or ECMS systems, such as Drupal or Wordpress, due to several facts:

- Liferay acts not only as a portal, but also as a "portal of portals", being able to manage hundreds of sites under the same Liferay instance.
- The templating system offered a good solution for integrating data sources and tools.
- The set of portlets provided with the system offered a good solution for building the sites from scratch.
- The learning curve is not as steep as in Drupal.
- The development team was familiar or had expertise either with Liferay or Java, which reduced the risk of getting familiar with a new infrastructure or programming language in the short period of time needed for releasing a proper beta version.

Additionally, Liferay 7 follows the OSGi standards, which ensures a proper modularity of each component. Moreover, its templating system allows a manageable solution for integrating existing applications, data bases and data structures, for instance the product listing tool (HTML5 + JavaScript + structure + FreeMarker template) or the component for representing WMS/WFS layers (HTML5 + JavaScript + OpenLayers 3).

Data integration and external service integration was finally implemented by means of REST APIs and JavaScript clients:

- The dataset discovery tool is implemented in JavaScript using the GEO DAB API.
- The Spatial Dataset Browser and the Species Population Trends tools are a JavaScript based, which links to GBIF services through its REST APIs.
- The Citizen Science Directory is a portlet, Java based, that links to PlutoF REST API.
- The Taxonomic Backbone is linked and requested by its REST API.

# 4. Metadata integration and dataset discovery

The Milestone **MS241** "Specifications for registry and metadata catalogue" established the recommendation of using GI-cat as metadata registry, thus aligning the registry with the GEO efforts, which additionally contributed to the release of the GEO Discovery and Access Broker as main core of the GEOSS infrastructure. The new evolution of the GEOSS Common Infrastructure, known as GCI 2.0, will be entirely based on the GEO DAB, and the release of the new version of the GEOSS Portal is part of this transition. Therefore, EU BON should be aligned with the GEO's strategy.

The GEOSS GI-cat broker offers a system for adding data providers, through a set of service connectors that could consume a wide range of standard interfaces, such as EML, OAI-PMH, OGC-CSW, THREDDS or OpenSearch. It can be deployed along an additional broker, GI-axe, specialised in georeferenced data layer formats, OGC-WMS, OGC-WFS and ArcGIS among others.

The dataset search interface, developed by ESSI-lab along with CSIC, consists a JavaScript client for retrieving datasets previously harvested by GI-cat, discovered by means of their metadata along with GBIF dataset information. The taxonomy filter request taxa information to the EU BON Taxonomic Backbone through its REST API.

On the other hand, the development strategy for the EU BON registry involved enhancement of the GBIF registry (**Figure 5**), which is closely connected to all the primary biodiversity dataset discovery services of GBIF.org. The registry is administrated by GBIF with full helpdesk support to publishers and a commitment from GBIF to sustain this beyond the life of the project.

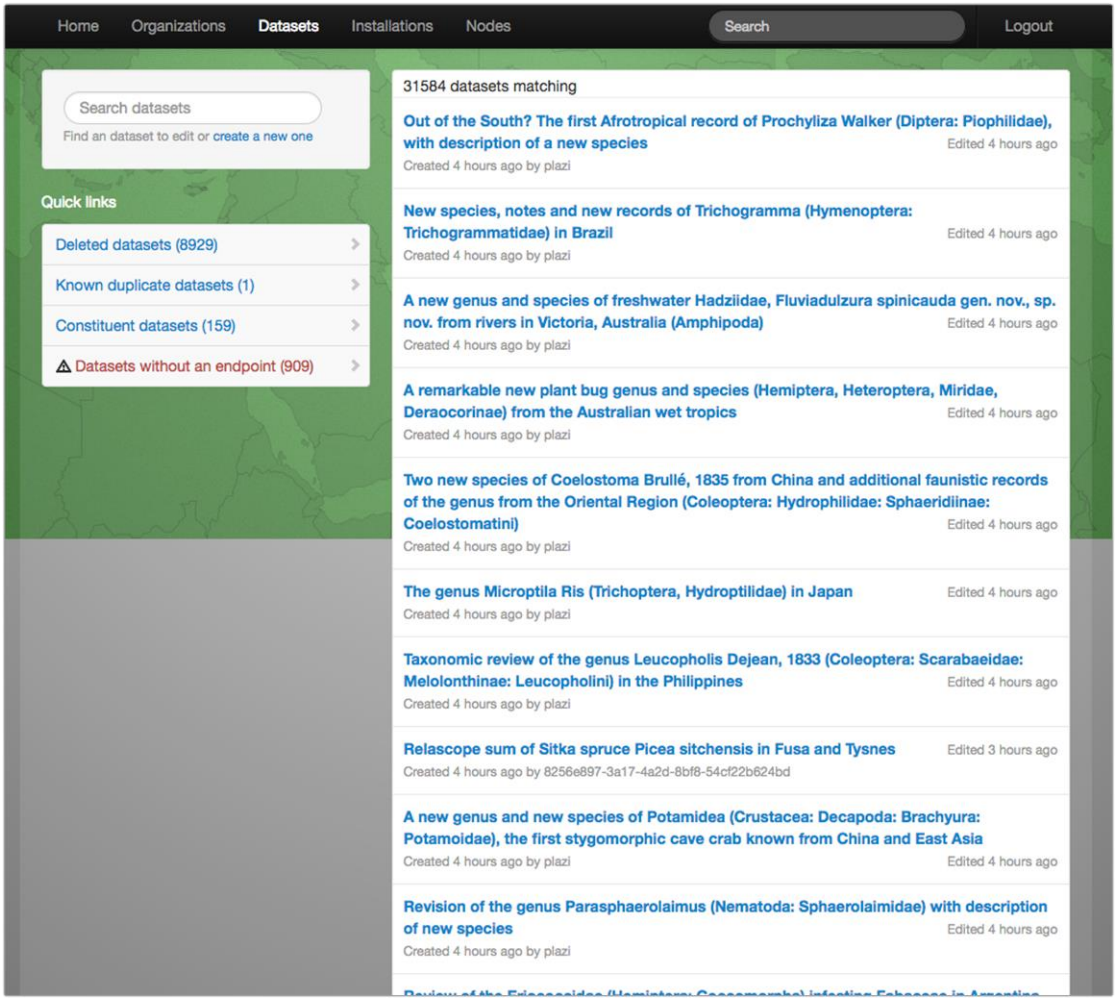


**Figure 5.** GBIF Dataset Registry

### Integration of GBIF in DataONE as a Member Node

Integration with DataONE has not been completed. GBIF have signed a Memorandum of Cooperation with DataONE to establish a GBIF DataONE MemberNode which will allow all datasets registered in the EU BON network to be discoverable and connected to DataONE. The development[3] for this has been started and is expected to be completed during 2017.

[3] https://github.com/gbif/dataone

# 5. Analysis tools

## 5.1. Spatial Dataset Browser

The spatial dataset browser (**Figure 6**) provides the ability to discover datasets containing evidence-based biodiversity data for a region and timespan. Unlike a metadata discovery system, which typically uses descriptive data about a dataset, the browser makes use of all record level content to index the geographic bounds of the content accurately at multiple scales and to year resolution. Such an approach provides a more accurate search result for the user than possible with metadata only approaches. The application was developed to explore a more graphical approach to dataset discovery for those datasets registered in the EU BON registry.

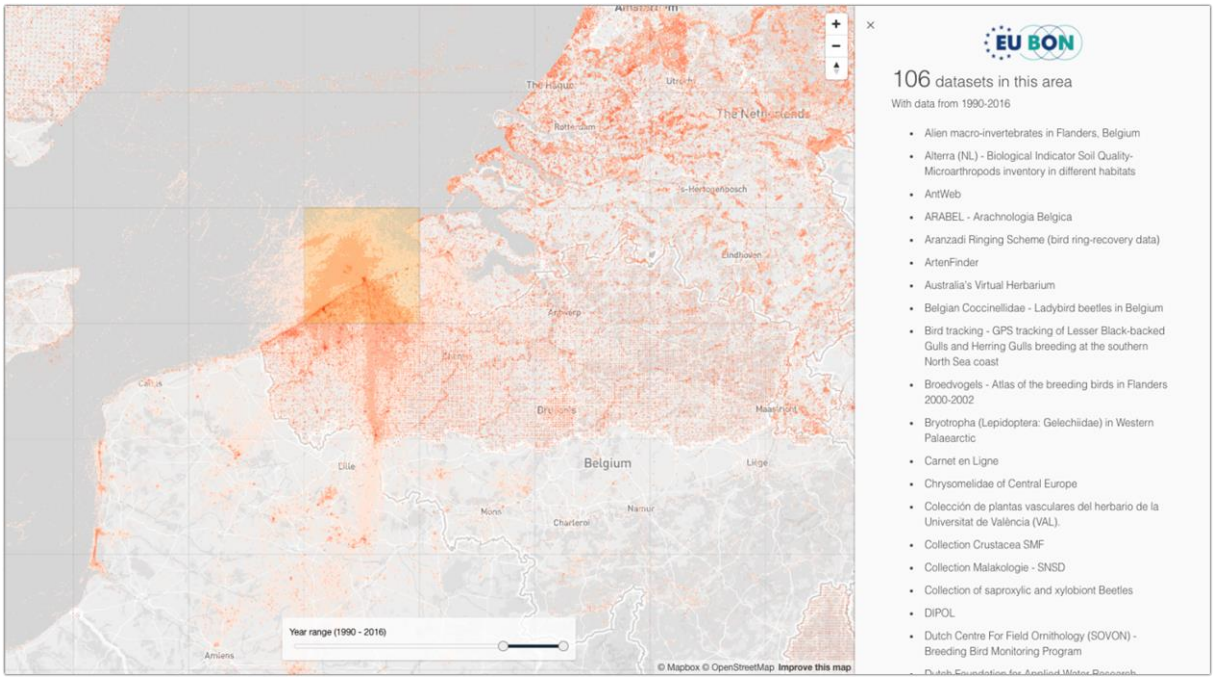


**Figure 6.** Spatial Dataset Browser

In developing this application, the developers also chose the opportunity to evaluate the emerging web mapping technologies built on vector tiles as opposed to simple image based mapping solutions[4]. The user feedback received confirmed that there is good support for mobile devices and the usability experience was relatively positive. This experience was carried through to the occurrence trend tools, and other mapping solutions in forthcoming revisions to the GBIF API.

The future roadmap for this pilot will be to bring the functionality into the GBIF API and dataset discovery services and surfaced on GBIF.org. The functionality will be enhanced with other dimensions such as publishing institution, keyword searches, etc., and also to include the index of all species contained. A combination of taxonomic, geospatial and temporal dimensions in the dataset index will provide a complete discovery service for this format of dataset. GBIF have indicated commitment to sustain, operate and advance this work as part of the daily routines of GBIF data management.

[4] https://www.mapbox.com/vector-tiles/specification/

## 5.2. Species Richness Application

The application is an analytical tool that utilises GBIF occurrence data to compute a spatial and temporal distribution of higher taxa.

The occurrence data are obtained from the GBIF as a snapshot and stored in the local database. Each record contains information on taxon, geographical coordinates of the locality, year of the observation, and number of records. Currently, there are 93 million records ranging from year 1600 to 2015 containing more than 176.5 million occurrences within the area of Europe (72° N - north boundary, 29° N - south boundary, 10° W - west boundary, 40° E - east boundary).

The list of higher taxa, as well and their hierarchy, is stored in the local database, as well. Currently, the application supports birds (class Aves), mammals (class Mammalia), frogs (order Anura), bony fish (class Actinopterygii), beetles (order Coleoptera), butterflies and moths (order Lepidoptera), mushrooms (phylum Basidiomycota), vascular plants (phylum Magnoliophyta), and all higher taxa included in them. The species data are obtained via GBIF REST API.

The input requires higher taxon, spatial extent, spatial resolution, temporal extent, and temporal resolution. The higher taxon is the subject for which the occurrences will be computed. An autocomplete input field is used which directs the user to selecting the supported taxon. Spatial extent is a bounding box of the total area that will be searched for occurrences and divided into grid. It is specified by north, east, south, and west boundaries. Spatial resolution is the size of a single cell in the result grid, specifically cell's width and height, hence the cell is a square. It is given as integer number representing geographical degrees. Temporal extent is an interval of years into which the searched occurrences must fall. The default year is 1758 - publication of the Systema Naturae by Linnaeus. Temporal resolution specifies the step between the years. Value of ‘1’ means that the application will display results for each year in the temporal extent. Value of ‘2’ will display every second year with results cumulated for every two years.

An optional input for species is used to compute the relative abundance of species in the selected higher taxon. Naturally, the species must belong to the selected higher taxon.

The application uses GBIF REST API's species function “/species/search” to retrieve accepted species of the selected higher taxon (**Figure 7**). These are then searched in the local occurrences data, while additional parameters are applied (spatial extent and temporal extent). Retrieved records are processed in order to place them into corresponding cells with specified size (spatial resolution), and to arrange them into years considering the temporal resolution. The cell is specified by the year it contains records for, and by the area it covers. Cells are created from the bottom-left geographical coordinates of the spatial bounding box. Considering the cell size can be any integer number, as well as the bounding area can be of any size, the bounding area is "filled" with the desired-sized cells from left to right, and from bottom up. Remaining cells in the last column on the right and the last row on the top side of the bounding area can be of smaller width and height, accordingly.

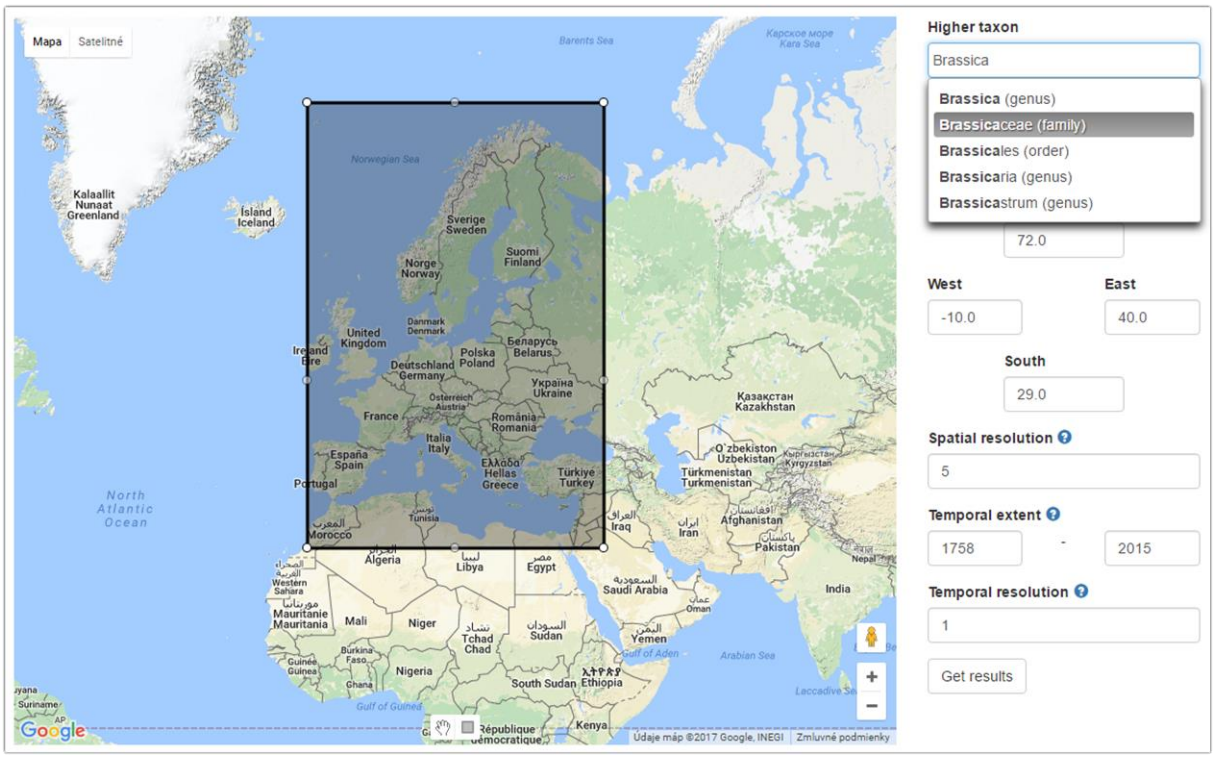


**Figure 7.** Selecting the higher taxon from the list of suggestions by auto complete field

Each query's results are presented as occurrences results (Occurrences tab), species results (Species tab), and optionally relative abundance of species in the higher taxon (Ratio tab). Google Maps API is used to display the appropriate layers of results. Each layer contains only cells containing common year. One can slide through or switch the layers using the "player widget". Each cell contains information on its location, number of occurrence records, species it contains, and number of occurrence records of the specific species if it is given.

The occurrence tab (**Figure 8**) shows distribution of the occurrences in the area, the species tab (**Figure 9**) displays species distribution the occurrences exist for in the area, and the relative abundance is the ratio of selected species' occurrences to the occurrences of the higher taxon. The described functionality is accessible via REST API, as well, which produces results in JSON or XML format.

The usage of the cached data instead of live ones using GBIF API is a cost to pay for high data access speed. The drawback of this approach is that the data must be periodically updated.

The future expectations in the development include support of all higher taxa in GBIF, optimization, and any additional functionality that might prove useful in the context of the application.

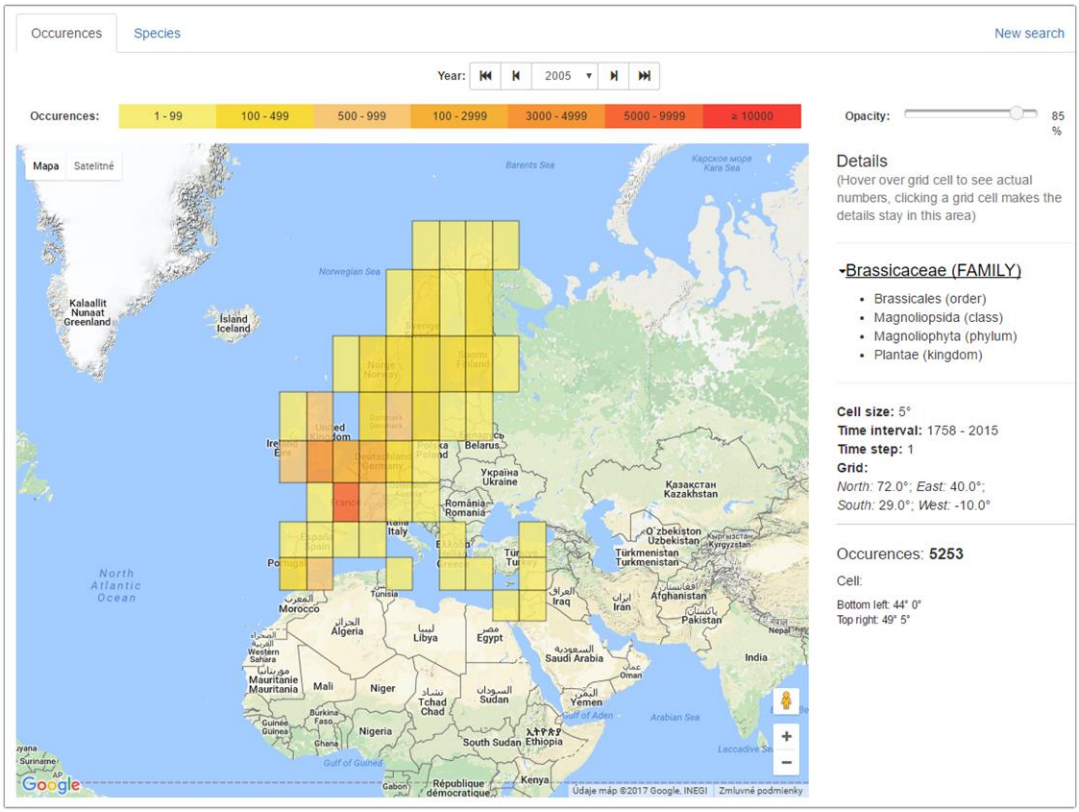


**Figure 8.** Occurrences results tab for 2005, with input parameters on the right side. Number of occurrences and the cell info are shown when mouse cursor is moved over the cell

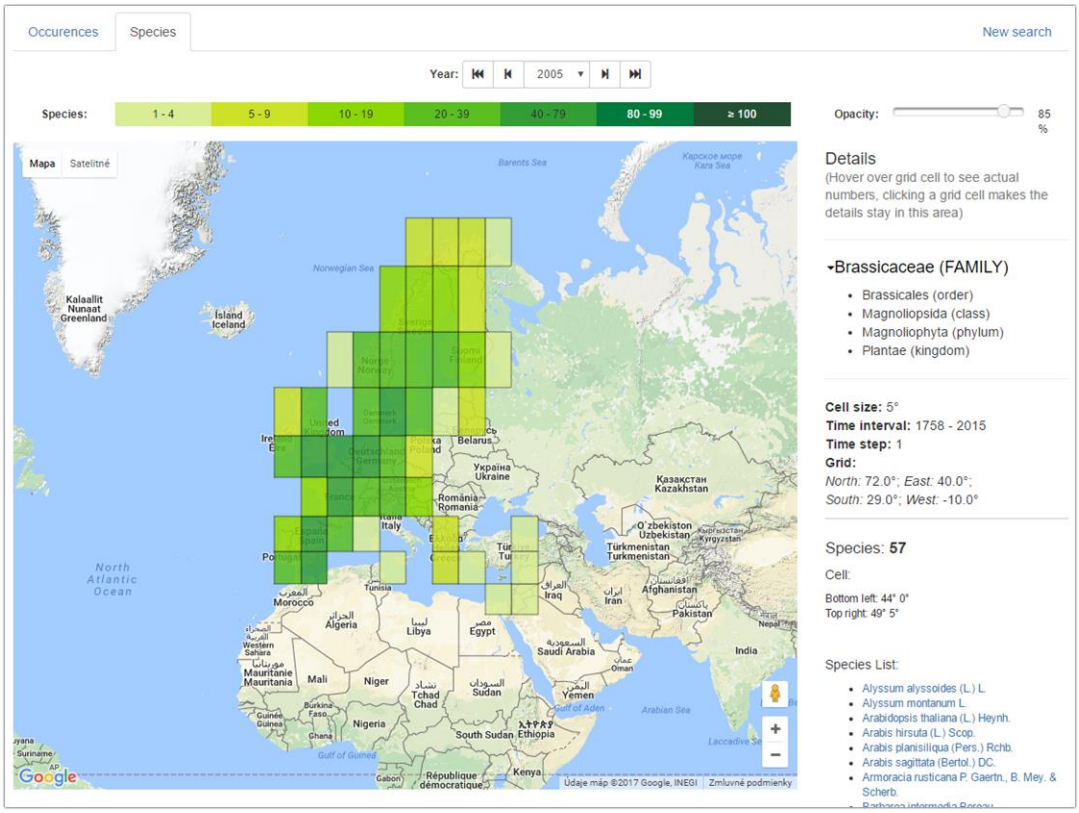


**Figure 9.** Species results tab for 2005. Moving a mouse cursor over specific cell shows the list and total amount of species in the cell

## 5.3. Species Population Trends Visualisation

The 'occurrence trends' is a pilot tool intended to help users determine if GBIF mediated data is likely to be suitable for deriving species population trends. The tool allows a user to navigate to a location and timespan of interest, and then it provides an analysis showing if the population for the species is in decline or increase (**Figure 10**). This analysis involves normalizing the data for the species against all data collected for the higher taxonomic group. By normalizing against the group, the analysis attempts to compensate for biases in sampling effort. In many cases this will work well, but in other cases this may provide misleading insights – however, by providing a transparent view of the process, the user can judge if there appears to be sufficient data to investigate further.

To explain the algorithm, please consider the following example of the early thorn moth (*Selenia dentaria*) normalized against its order (*Lepidoptera*):

- In 1968, there were four occurrences (of a total of 12 lepidopterans) = 33.33%
- In 2002, there were three occurrences (of a total of 6 lepidopterans) = 50%
- The normalized population is said to have grown by 50% and the trend is thus positive

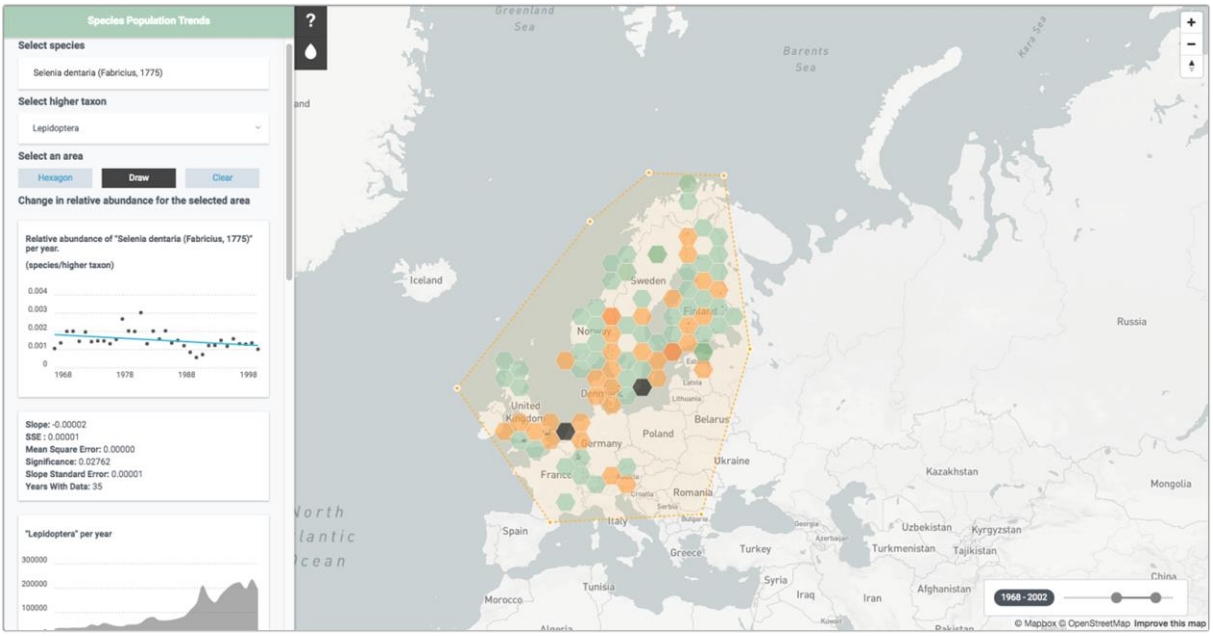


**Figure 10.** Species Trends Visualisation, filter by species and area described by a polygon. Results as hexagons

In order to create the tool, the mapping services of GBIF.org have been fully revised to support vector tiles, the ability to aggregate data into hexagons and the ability to perform simple linear regressions. By incorporating this into the GBIF API and data management, the data is always up-to-date and the ongoing sustainability of the tool will be assured within the GBIF product suite.

All data that is used for this application is available for download through GBIF.org.

## 5.4. Business Intelligence Dashboards

Business intelligence is defined as the set of strategies, processes, applications, data, products, technologies and technical architectures that are used to support the collection, analysis, presentation and dissemination of business information. Dashboards represent a powerful tool for transforming data into knowledge, by means of representing aggregated data tables into charts, maps and data

summaries that can provide an at-a-glance view of include key-performance indicators (KPI), relative to a particular objective or business process.

Considering the convenience or representing aggregated data for the policy makers, scientists and public users, the development team decided to build a data warehouse with a snapshot of GBIF data, corresponding to all European occurrence records. A data warehouse is a repository of integrated data from one or more disparate resources, often implemented as denormalised data bases that follow a star, snowflake, or hybrid schema. The star schema is the simpler and faster, which consists of several dimension tables connected to a central table, the facts table, which references the dimensions' tables by their primary keys.

The EU BON Portal data warehouse follows the star schema design, and it consists of three dimensions: time, location and taxonomy, along with a central facts table. The initial data was a snapshot of GBIF data, exported as a CSV file, whose columns were a of the Darwin-Core available terms. This file consisted of more than 217 million records and about 100 GB of data.

For building and populating the data warehouse, the developers used Pentaho 6.x/7.0 Data Integration (PDI), with which Extract, Transform and Load processes (ETL) could be defined to transform the initial data into the denormalised tables and also to update them periodically. The facts table includes each dimension's primary key and a column for representing the number of occurrences per taxa, location and time.

For storing the database, PostgreSQL was chosen, although an initial MySQL based data warehouse was design and tested for the beta portal release. PostgreSQL 9.6 introduces parallel queries, which can reduce the time needed for performing complex data queries, as the ones needed for building a BI dashboard.

For creating OLAP cubes and the dashboards, the development team used Pentaho Business Intelligence Server, and in particular this set of tools:

- Mondrian for creating the OLAP cubes (dimensions organised as hierarchies, facts table definition).
- Saiku for allowing OLAP analyses in a web environment.
- Pentaho CTools to build the dashbard, complemented with Ivy IS CDE components.

As depicted in the **Figure 11**, with the EU BON Biodiversity Portal dashboard, users can filter by any taxa level, time range or country, and after a few seconds the charts will be updated with aggregated data from the data warehouse.

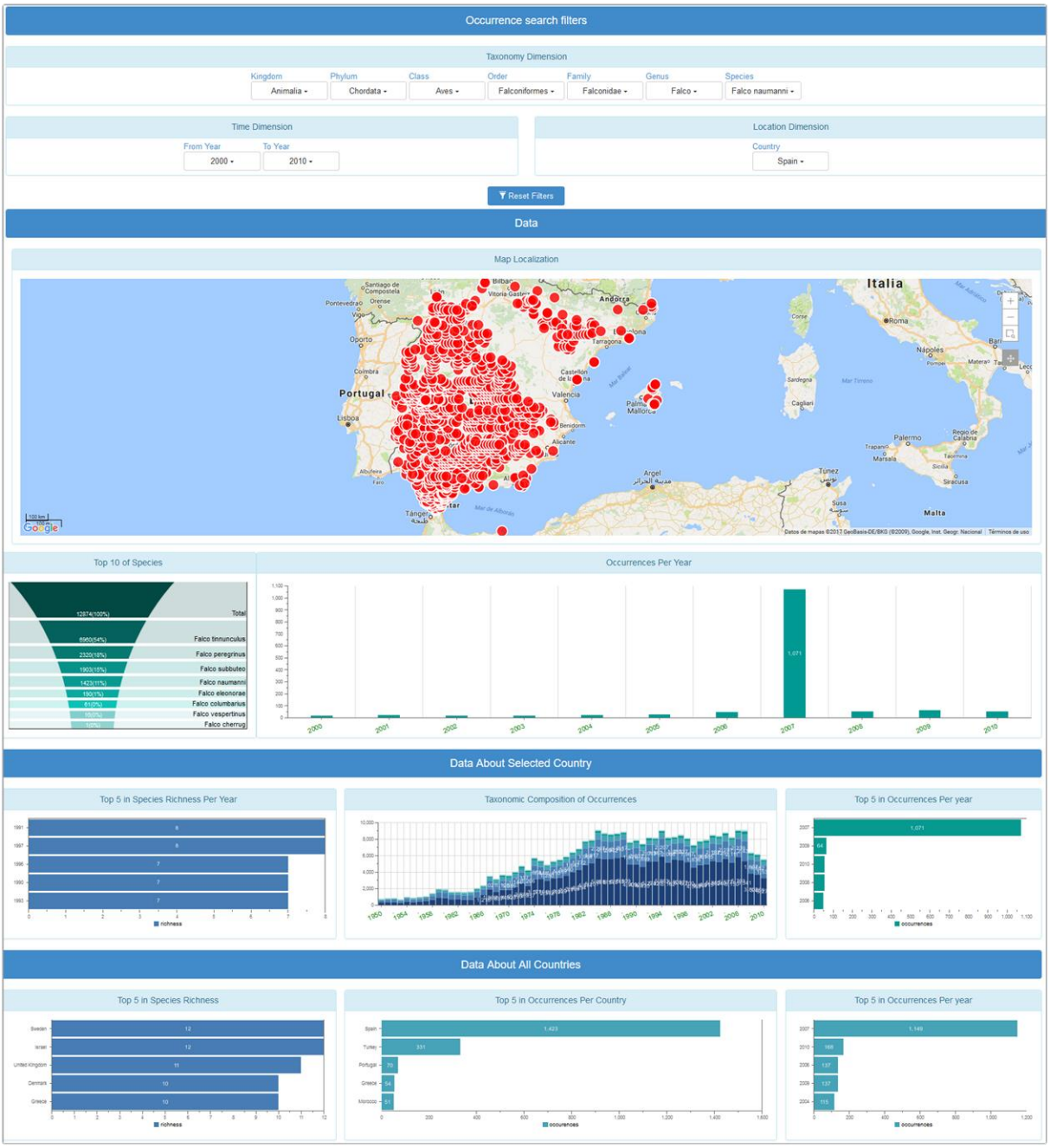


**Figure 11.** Dashboard, filtering by species, country and time range

# 6. EU BON services and tools for data publishing and sharing

## 6.1. Directory of tools for data publishing and sharing

EU BON offers a wide range of tools for helping data providers to share their own data collections (**Figure 12**). The list was implemented as a dynamic data list in Liferay, providing access to each tool and their guidelines.

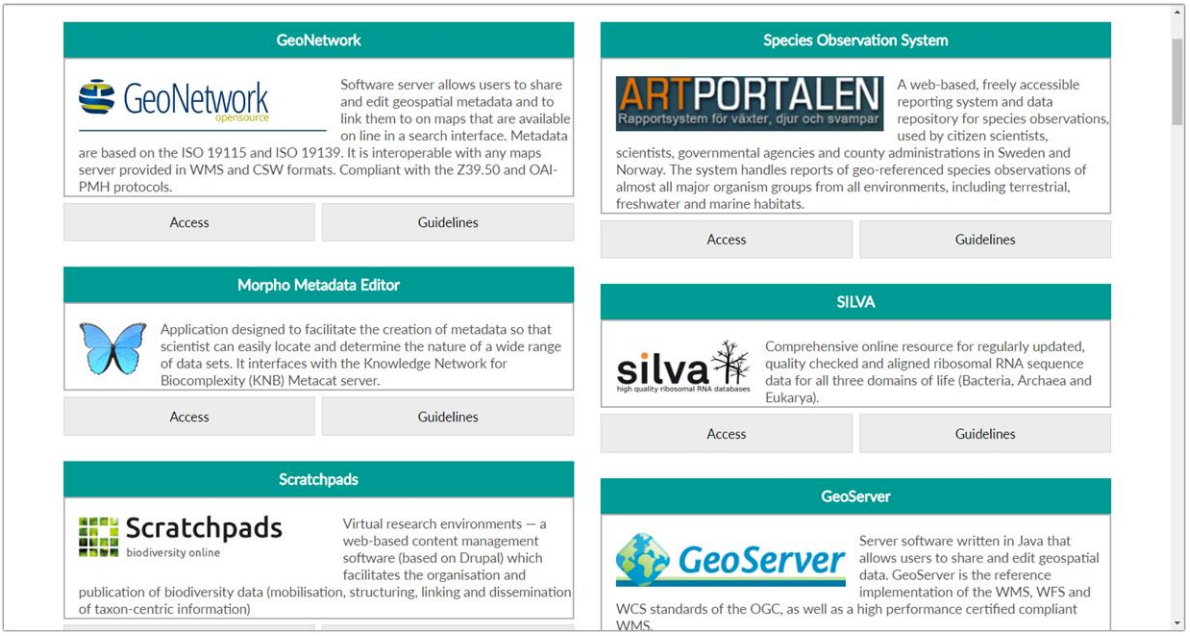


**Figure 12.** Data sharing tools

## 6.2. EU BON GBIF IPT instance

The GBIF IPT is the most established data repository tool in use within the GBIF network. It is deployed in more than 500 institutions across 52 countries and is responsible for the publication of more than 3000 datasets. The EU BON project focused on research into a sample-based data standard, and in enhancing the IPT to accommodate that standard. This was communicated and tested through both the EU BON project and the GBIF Nodes network and was put into production in IPT release version 2.3.

The development process included the deployment of a prototype repository – the Test EU BON IPT instance[5], which is embedded in the EU BON biodiversity portal as a sandbox for research and training.

Early adopters began publishing the first round of real sampling-event datasets in late 2015. The Production EU BON IPT instance[6] (**Figure 13**) was setup to accommodate publishers sharing real datasets but unable to find another IPT host for their data.

[5] http://eubon-ipt.gbif.org/
[6] http://cloud.gbif.org/eubon

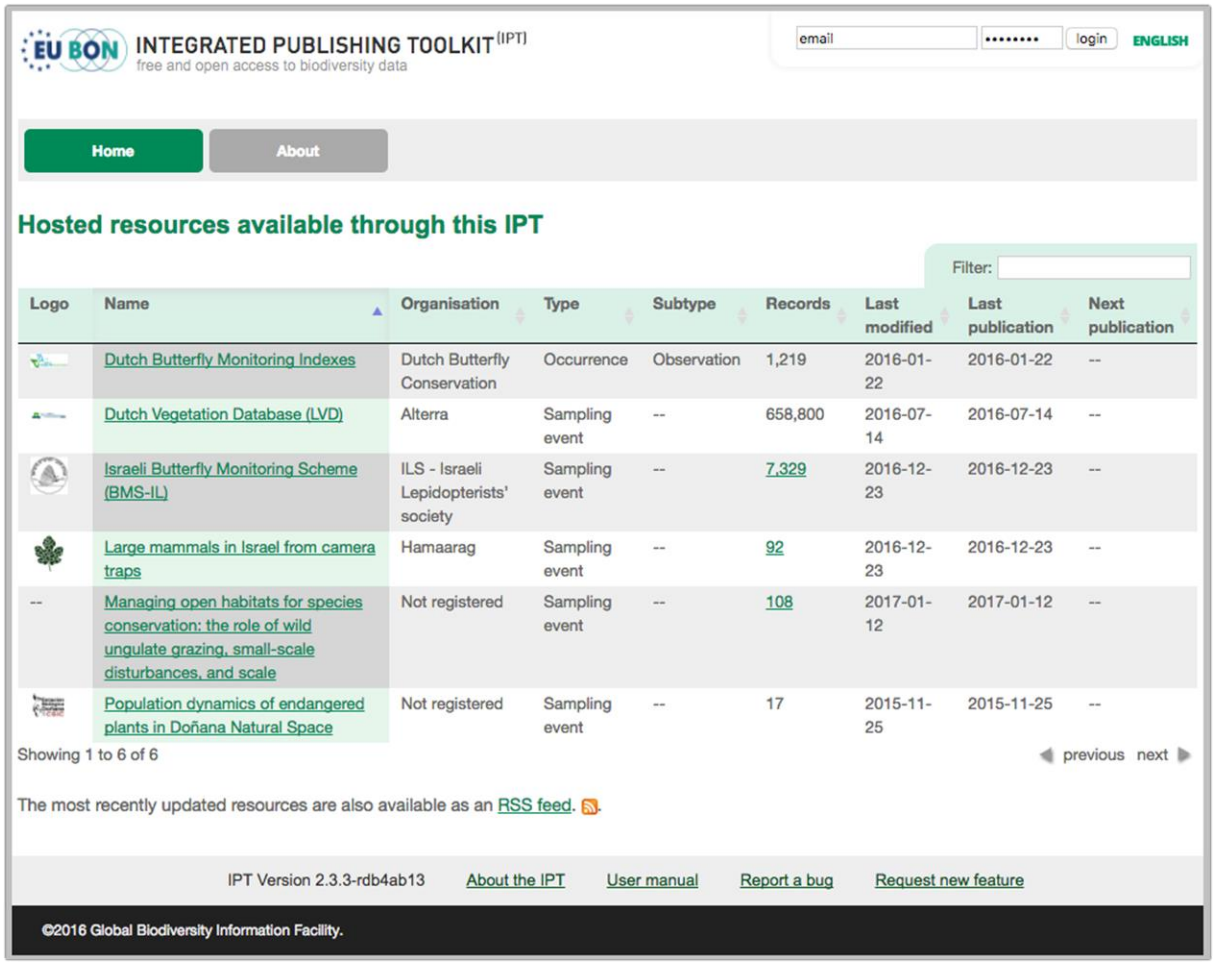


**Figure 13.** EU BON IPT instance

In several cases the EU BON project used the test installation for exploratory work, but collaborated with the National GBIF activities when moving to production. This is a positive step to retain national ownership and local expertise.

Currently the Production EU BON IPT hosts 5 sampling-event datasets. This includes two datasets that are worth highlighting:

- Israeli Butterfly Monitoring Scheme (BMS-IL)[7]: Work on this dataset lead to the development of a data model for systematic monitoring scheme data that can be adapted to many other types of systematic monitoring, for many taxonomic groups. For more information, see this blog post[8].
- Dutch Vegetation Database (LVD)[9]: Work on this dataset lead to the development of a data model for vegetation plot data, which Turboveg v3 will support as an export format. For more information, see this blog post[10].

[7] http://cloud.gbif.org/eubon/resource?r=butterflies-monitoring-scheme-il
[8] http://gbif.blogspot.dk/2017/01/sampling-event-standard-takes-flight-on.html
[9] http://cloud.gbif.org/eubon/resource?r=lvd
[10] http://gbif.blogspot.dk/2016/07/probably-turbovegs-best-kept-secret.html

Note that in the future, all datasets hosted on the Production EU BON IPT will be migrated to another trusted IPT host, ideally located in the dataset's country of origin or a GBIF offered repository as part of the sustainability plan.

IPT version 2.3.3 is a more robust and secure version that was released in December 2016. It includes a couple of enhancements that make it easier to publish and visualise the contents of sampling-event datasets[11]. A more complete description of what is new in version 2.3.3 can be found in this blog post[12].

## 6.3. Toolbox for Scholarly Publishing and Dissemination of Biodiversity Data (ARPHPA-BioDiv)

The EU BON's Toolbox for Scholarly Publishing and Dissemination of Biodiversity Data (ARPHA-BioDiv) (**Figure 14**) is a set of standards, guidelines, recommendations, tools, workflows, services, and journals, based on the Pensoft's ARPHA Journal Publishing Platform of Pensoft. It is designed to ease scholarly publishing of biodiversity and biodiversity-related data that are of primary interest to the EU BON and GEO BON networks.

The core element of ARPHA-BioDiv Toolbox is the ARPHA Journal Publishing Platform developed by Pensoft, along with the associated journals: Biodiversity Data Journal (BDJ), Research Ideas and Outcomes (RIO), and One Ecosystem. ARPHA is an innovative publishing solution that supports the full life cycle of a manuscript, from authoring and reviewing to publishing and dissemination. The data publishing strategy of ARPHA aims at increasing the proportion of structured text and data within the article content, so as to allow for both human use and machine readability to the maximum possible extent.

### Case studies

#### a) Fauna Europaea Special Issue

One of the major data mobilisation initiatives realised by ARPHA and the Biodiversity Data Journal[13] (BDJ) is the publication of data papers on the largest European animal database 'Fauna Europaea'. A new series 'Contributions on Fauna Europaea'[14] was launched in the beginning of 2014. This novel publication model aimed to assemble within a single collection 57 data-papers on different taxonomic groups covered by the Fauna Europaea project, as well as a range of accompanying papers highlighting various aspects of this project (gap-analysis, design, taxonomic assessments, etc.). The first two papers were published on 17 September 2014 and until the end of 2016, 12 articles altogether have been added in BDJ (de Jong et al., 2014).

#### b) EU BON Open Science Publication Pilot

The journal Research Ideas and Outcomes (RIO) was designed to publish all outputs of the research cycle, from research ideas and grant proposals to data, software, research articles and research collateral, such as workshop and project reports, guidelines, policy briefs, and Wikipedia articles, to name just a few. Within RIO Journal, EU BON realised one of the first ever open science collections

[11] https://github.com/gbif/ipt/issues?q=is%3Aissue+milestone%3A2.3.3+is%3Aclosed+label%3Asampling-event
[12] http://gbif.blogspot.dk/2017/01/ipt-v233-your-repository-for.html
[13] http://bdj.pensoft.net
[14] http://bdj.pensoft.net/browse_user_collection_documents.php?collection_id=9&journal_id=1

of project publications entitled: Building the European Biodiversity Observation Network (EU BON) Project Outcomes[15]. Up to date, the collection contains 11 publications.

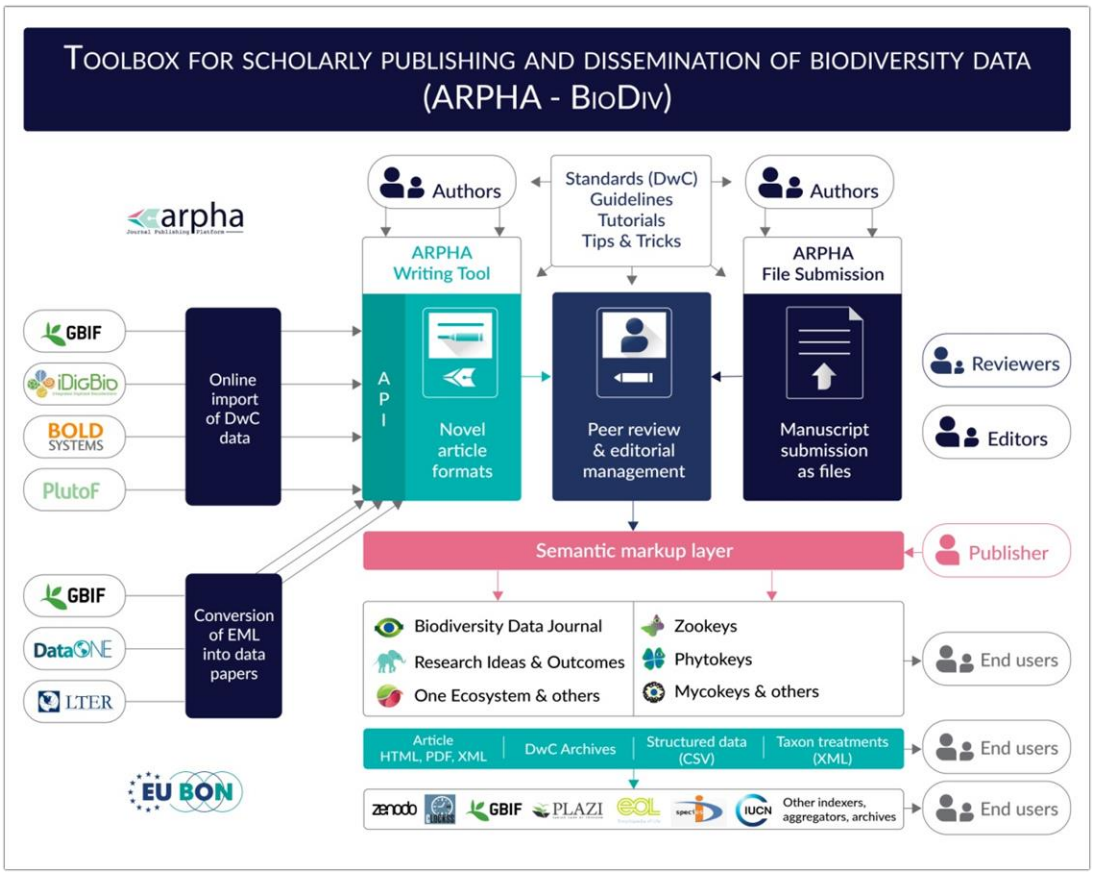


**Figure 14.** EU BON's Toolbox for Scholarly Publishing and Dissemination of Biodiversity Data (ARPHA-BioDiv).

## 6.4. Plazi TreatmentBank

Plazi's TreatmentBank (**Figure 15**) is a novel approach to free data relevant for biodiversity monitoring from scientific publications. It is making use of the highly standardized way descriptions of the world's species published, started in 1753. These descriptions (taxonomic treatments) are the scientists' diagnosis of species they deal with. In the last years, approximately 17,000 new species are being described with a multiple of re-descriptions. Typical data elements are the scientific names, observation records, morphological descriptions, illustrations or citations. The legal aspects of these elements has been discussed by Patterson et al., 2014 (doi: http://www.biomedcentral.com/1756-0500/7/79) and Egloff et al., in Press (doi http://dx.doi.org/10.1101/087015).

[15] http://riojournal.com/browse_user_collection_documents.php?collection_id=2&journal_id=17

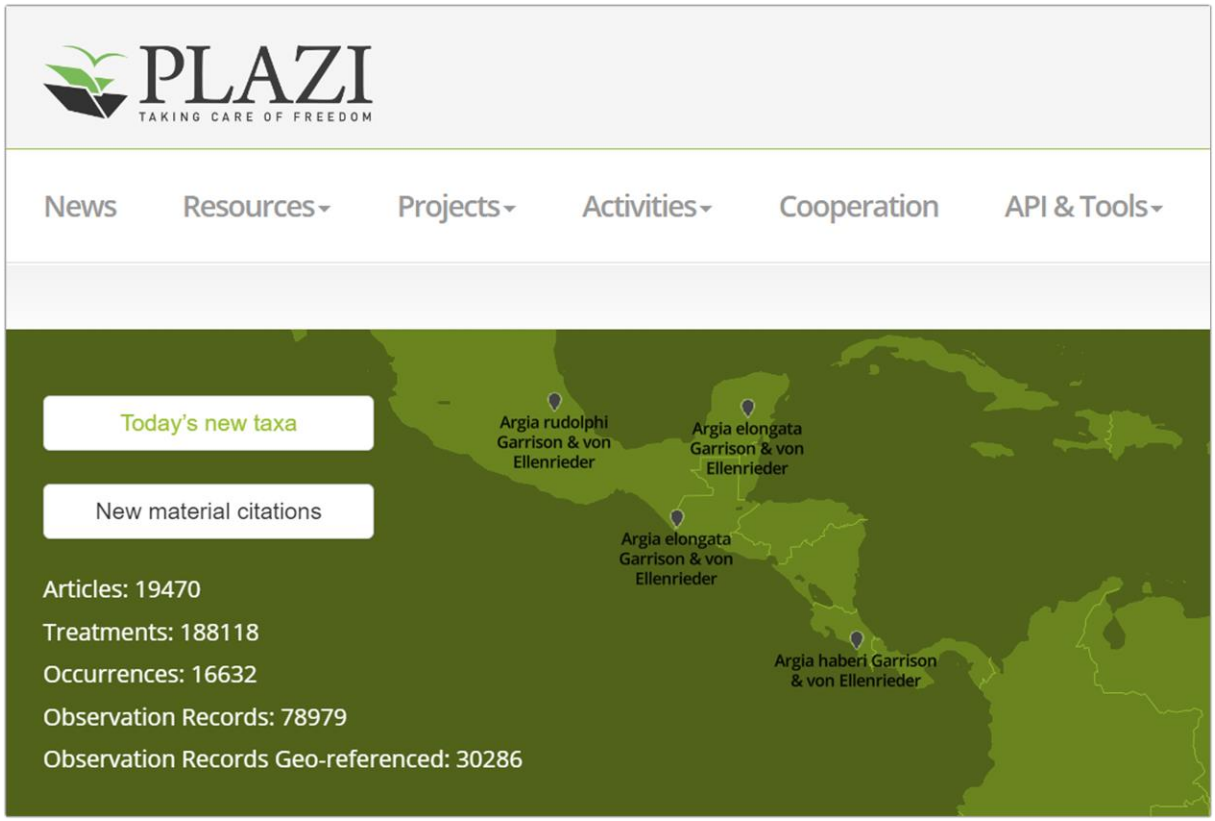


**Figure 15**. Plazi main webpage

TreatmentBank is a service that includes data mining, on search and visualization, and data dissemination.

Data mining is based on two tools. For traditionally published articles in PDF or HTML format, GoldenGate Imagine, a tool considerably developed during the EU BON project, is deployed. This generic tool can be highly customized to a degree, that journals can be fully automatically mined down to individual observation records, entities can be linked to or enhanced from external resources such as the GBIF taxonomic backbone. For semantically enhanced journals, and XSLT transformation is in place. Once a document is mined and converted into an IMF file – an open source file format for scientific articles – illustrations are submitted to the Biodiversity Literature Repository[16] at Zenodo/CERN where a Digital Object Identifier is minted and included in the source document. An event is triggered to alert GBIF of new data that is transferred as a Darwin Core Archive. The DCA is based on the treatments included in the source publication. Observation records submitted to GBIF can be traced back using he persistent identifier minted by Plazi for each taxonomic treatment, and from this via the metadata to the source publication.

TreatmentBank provides a daily update of new taxonomic treatment via a continually updated Web page[17], a search interface[18] or for advanced users statistics and search tools[19]; for machine access, a description is provided[20].

[16] BLR: http://biolitrepo
[17] http://tb.plazi.org/GgServer/static/newToday.html
[18] http://tb.plazi.org/GgServer/search
[19] http://plazi.org/api-tools/statistics/
[20] http://plazi.org/api-tools/api/.

Treatments can be individually viewed or more inclusively as higher taxa or article or journals. For each result page for taxa, depending on the availability of semantically enhanced data, further options are provided (**Figure 16**). This includes on the one hand metadata and data that is part of the treatment (treatment), or referenced from within the treatment (figures, references, tables, abbreviations). Figures are shown as thumbnails with a link to the base record on BLR. On the other hand, tools provide visualizations of the taxonomic hierarchy and citation of previous treatments (Taxonomy), point distribution maps (Distribution Map), dashboards for specimen data (**Figure 17**). Finally, the content of the treatments can be downloaded as generic XML, TaxonX, RDF of DWC formats.

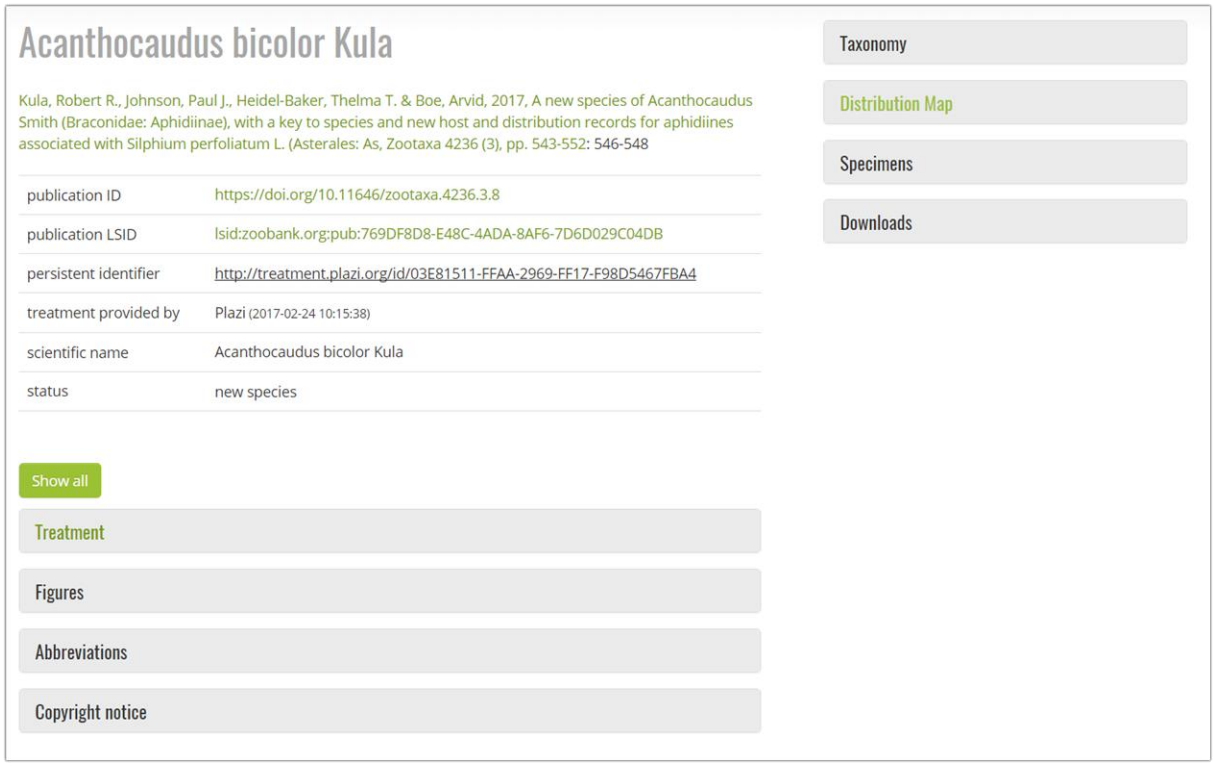


**Figure 16.** TreatmentBank taxon pages


**Figure 17.** TreatmentBank dashboard visualizing data mined from Dijkstra et al. 2015[21]

Currently, 19,471 articles have been processed resulting in 188,124 treatments, 79,023 observation records, 172,838 unique taxon names, 104,107 illustrations and tens of millions of RDF triples. For 2016, 5,055 new species to science have been discovery in the publications. With this, Plazi is one of the main name providers for GBIF. Including all treatments 43,785 new records have been added.

The data source includes 70 different journals totalling 3,220 articles in 2016, with an estimated 28% of all the new species described annually, a major complementary contribution to enhance the long tail of rare species in GBIF.

The crowd can participate in adding more data to TreatmentBank, from adding new articles to BLR as source for conversion, to converting individual articles to developing templates to automatically convert journals.

[21] http://tb.plazi.org/GgServer/summary/FF9B2A1CCA19FFEAEE22FFED4105FFB2

# 7. EU BON Product List

An important objective of the EU BON European Biodiversity Portal is to act as a showcase of EU BON products and outcomes. The products section (**Figure 18**). lists the wide range of available products, either tools, factsheets or publications. The product list was redesigned prior to launching the final version of the portal, applying the following major changes:

- The overall design follows the Material Design guidelines.
- User interface implemented in HTML5 + Freemarker + Materialize framework.
- Each product is a reversible card, with a detailed description on the back.
- Filter by audience and topics.

The factsheets and guidelines are referenced as documents and stored in the Liferay site's document repository. This repository is compatible with CMIS, thus replaceable by any advanced Document Management System, for instance Alfresco or Documentum. Nonetheless, it was not required for the purpose of the portal, given that Liferay is capable of dealing complex workflows and controlling each document's versioning.

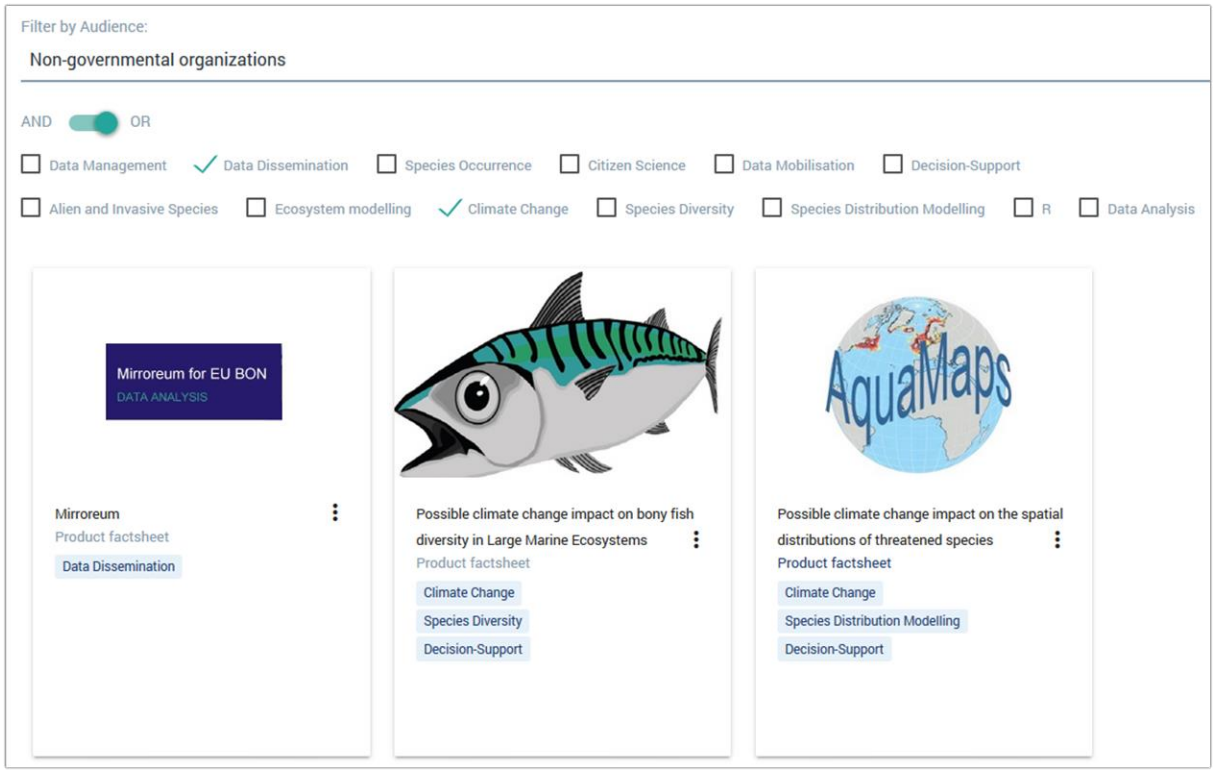


**Figure 18.** Product list, filter by audience (Non-Governmental organizations) and by topic (data dissemination or climate change)

# 8. Subsites

## 8.1. Citizen Science subsite

The Citizen Science (CS) sub-site within the EU BON portal provides outputs of the Task 1.5, which contributes to the overall EU BON strategy to directly support data mobilization efforts and generating biodiversity data. Elements of these task's objectives include plan to develop citizen science networking and gateway for biodiversity information, which aim at "*supporting an education network that will provide links to international and global initiatives and programs"*. The site provides the user with best practice examples based on evaluation of CS status and environmental education and to support linking and integrating of CS activities and data via the EU BON portal in order to achieve more comprehensive data coverage.

The CS subsite consists of several main components:

- '**Directory of CS Tools**' page is collection of references to online applications, guidelines and tools. These tools will help citizen science project managers or individual citizen scientists to collect and manage biodiversity data, set up project communication or build and participate in citizen science network. Serving as a library of tools, each tool in the directory is tagged by its category (e.g. 'Data-Sharing', 'Publishing', 'Monitoring', etc.). Each of these tools can be viewed and access using the 'Category Navigation' search engine of the CS entry page 'Citizen Science'.
- '**Start your project**' page: helps beginning citizen scientists or CS project manager to initiate their biodiversity-related activities. CS project initiators can find best practice examples that EU BON has picked. Page also provides guidelines from European Citizen Science Association (ECSA) for starting CS projects and practical introduction to the citizen science module of biodiversity data workbench PlutoF (http://www.plutof.com) developed by University of Tartu (Estonia).

- '**Directory of data providers**' offers a valuable detailed list of biodiversity-related CS data providers that can help researcher as well as CS project managers. Page is linked with PlutoF reference management module where provider data can be added and edited. The Directory of CS data providers is accumulated by Norwegian Biodiversity Information Center (Norway) and deployed on Liferay page using a portlet package developed by GlueCAD (Israel).

**Components include in CS sub-site:**

For setting up a CS project the PlutoF team (of partner UTARTU) offers a workbench and database which allows managing observations collecting, reporting and publishing to GBIF. Butterfly observers can use the 'I saw a Butterfly' app. (by GlueCAD partner) to report sporadic sites to PlutoF.

## 8.2. EU BON Test Sites: subsites

The Enterprise Content Management System that hosts the portal provides the ability of configuring a hierarchy of sites, that could act as individual web portals or in a parent-child site configuration. The development team took advantage of this functionality to build specific websites for the EU BON Test Sites which may demand to publish their information and EU BON related products in a web environment.

A set of components or portlets were configured and disposed to site administrators, in order to help them to configure their own sites:

- Web content.

- RSS Feeds (provide and consume).
- Iframes.
- Wiki pages.
- CMIS Document repository.
- A component for representing WMS/WFS layers, developed using HTML5, JavaScript and OpenLayers 3.
- Live video streaming.
- CartoDB maps.

### Doñana Biological Station Test Site

The Doñana Biological Station' site[22] was configured as a pilot of the test sites at the EU BON Portal (**Figure 19**), thus helping other site administrators to set up their specific sites' websites. It provides a detailed overview of the set of functionalities offered, including these pages and components:

- Description and news.
- RSS feeds: Doñana and EU BON events.
- Live streaming from cameras at the Doñana National Park
- Document repository
- Dashboard: several charts calculated with data collected in the Doñana Protected Area.
- Real-time sensor data, which lists a set of charts with data retrieved from the sensors deployed on the field, aggregated by steps of 5 minutes each.

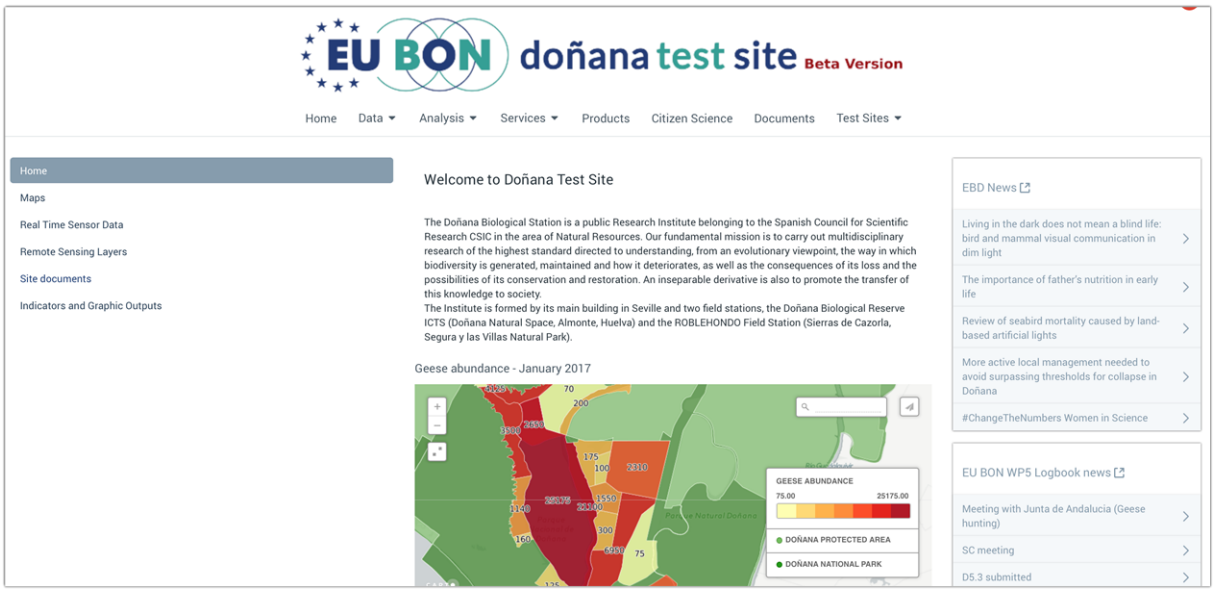


**Figure 19.** Doñana Test Site: home page

[22] http://biodiversity.eubon.eu/web/donana/home

# 9. EU BON Biodiversity Portal: elab

A virtual laboratory, "elab", can be defined as a collaborative website, in which user groups are provided with means to share data, publications, results and efforts in a shareable manner, as well as use with specific tools deployed and hosted in the elab.

As part of the adaptation to LifeWatch undertaken by the Doñana Biological Station (CSIC), a eLab hosting service was deployed, with the ability to provide specific tools running in a Software-as-a-Service environment. Technically, the elab hosting service runs onto OpenStack and OpenShift V3, providing respectively a Platform-as-a-Service and Software-as-a-Service environment. Each tool is built previously as a Docker image, then deployed, replicated and orchestrated by OpenShift and Kubernetes. Each tool is uploaded and configured once, and instantiated as many times as needed, usually one per eLab. Each tool can be also replicated on demand, either with a manual load balancing or configuring an autobalancer. Additionally, the elabs portal[23] is based on Liferay 6.2, in which each eLabs consists of a public website and an optional intranet site.

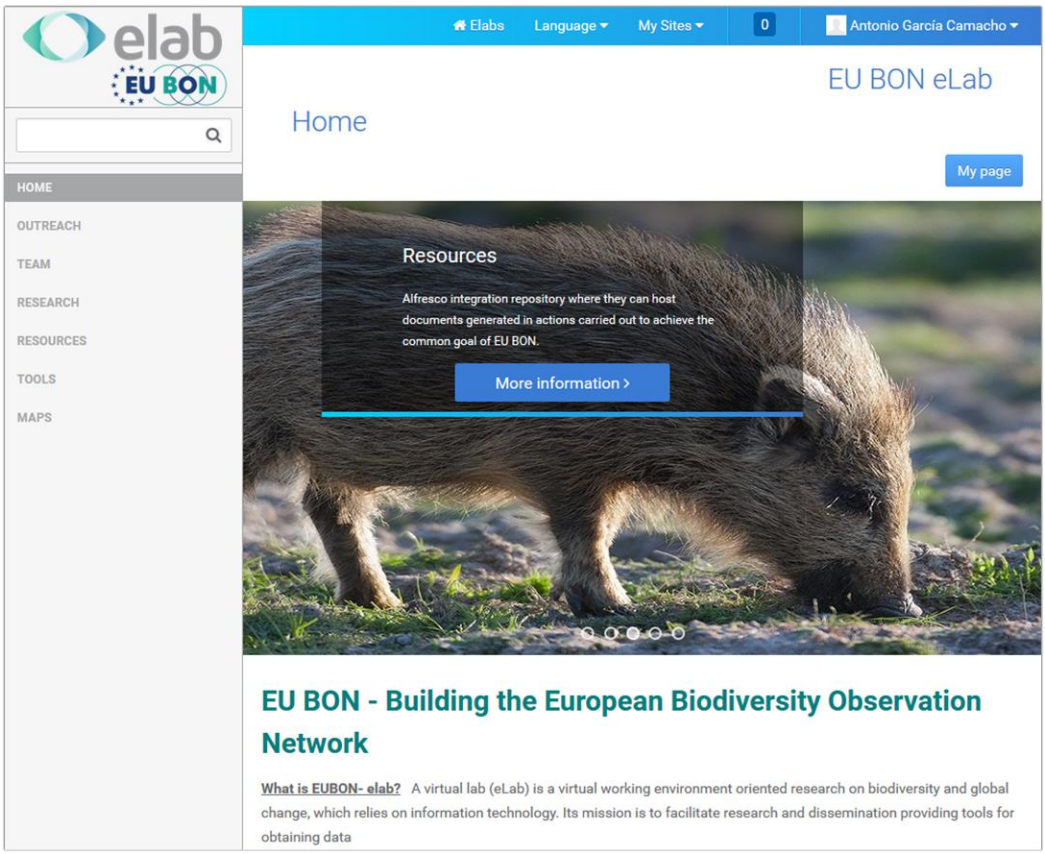


**Figure 20.** EU BON elab

The EU BON elab[24] (**Figure 20**) was configured with the following tools (**Figure 21**), although more tools are added whenever demanded by the users.

[23] http://elabs.ebd.csic.es/en
[24] http://elabs.ebd.csic.es/en/web/eu-bon-elab/tools

- Mirroreum RStudio: RStudio server with specific packages needed to run specific biodiversity-related R scripts and developed under the scope of EU BON.
- Jupyter-*: Jupyter is a web application that allows running scripts in a wide range of compilers. Jupyter-R, Jupyter-Matlab and Jupyter-Python are provided.
- Galaxy, an open, web-based platform for bioinformatics.
- GeoServer, an open source server for sharing geospatial data, for instance, as WMS and WFS layers.
- CartoDB: a web-mapping engine that enables the user to create its own interactive maps, based on uploaded data.

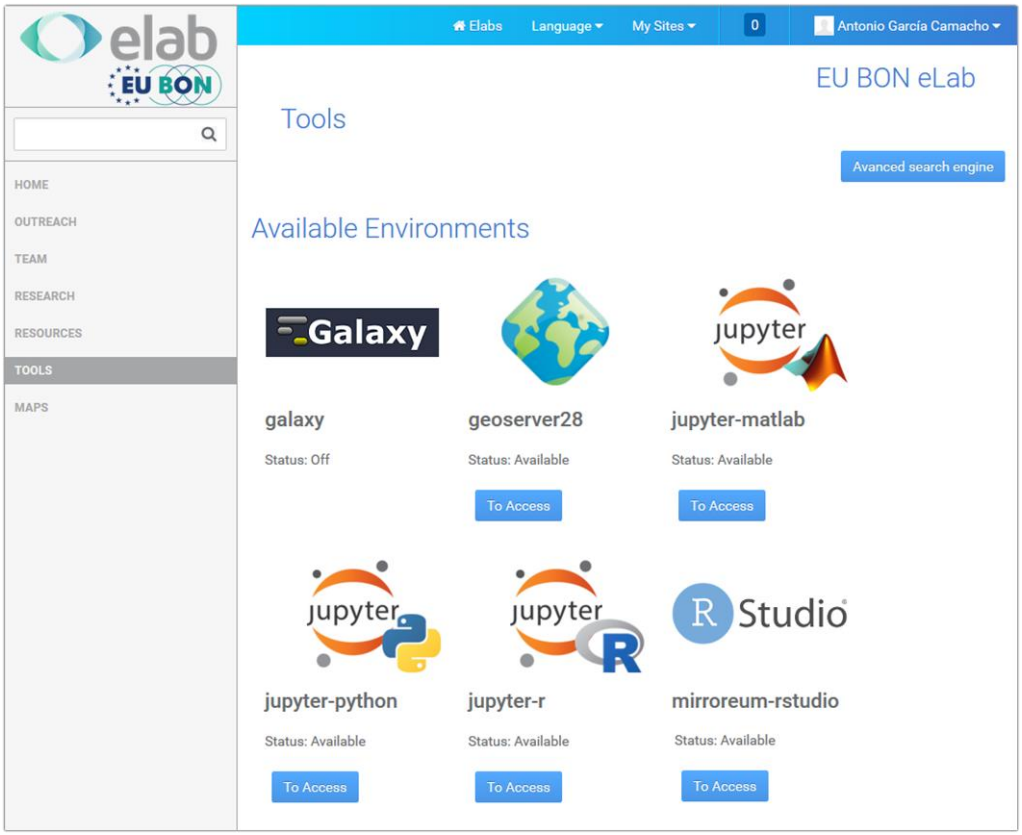


**Figure 21.** EU BON elab: tool set