



Deliverable 2.3 (D2.3)
Registry and Metadata Catalogue
M39

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: M39 (February 2016)

Actual delivery date: M39 (February 2016)

Lead beneficiary: GBIF

Authors: Tim Robertson (GBIF, Global Biodiversity Information Facility, Denmark)
 Federico Mendez (GBIF, Global Biodiversity Information Facility, Denmark)
 Kyle Braak (GBIF, Global Biodiversity Information Facility, Denmark)
 Hannu Saarenmaa (UEF, University of Eastern Finland, Digitisation Centre, Finland)
 Antonio García Camacho (CSIC, Consejo Superior de Investigaciones Científicas, Estación Biológica de Doñana, Spain)

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.

All intellectual property rights are owned by the EU BON consortium members and protected by the applicable laws. Except where otherwise specified, all document contents are: "© EU BON project". This document is published in open access and distributed under the terms of the Creative Commons Attribution License 3.0 (CC-BY), which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.



Table of contents

1	Executive summary	4
1.1	Introduction.....	4
1.2	Progress towards objectives.....	4
1.3	Achievements and current status.....	4
2	Overall architecture	6
2.1	Introduction.....	6
2.2	General registry architecture	6
2.3	Common data model.....	9
3	Enhancing the GBIF Registry	9
3.1	Architectural overview	11
3.2	Real time data indexing and visualization	12
3.3	Promoting consistent citation and tracking of use.....	13
3.4	Monitoring data trends over time	13
3.5	Improving the management of data vocabularies.....	15
3.6	Vocabulary changes to support sample-based data	16
3.7	Data mobilization	17
4	The Data Access Broker	21
5	Connecting other systems	23
5.1	Pensoft Biodiversity Data Journal	23
5.2	Plazi treatment database.....	23
5.3	Citizen science observations through PlutoF	23
5.4	EuMon.....	24
6	Future developments	24
7	References	25

1 Executive summary

1.1 Introduction

The goal of the Deliverable is to provide a functional system enabling the registration and discovery of datasets through EU BON and the GEOSS Common Infrastructure. Realising an effective open data infrastructure to achieve this requires attention to both technical and social aspects. A highly functional system is only useful if it has the trust and support of the data publishing community to share data through it. This document provides both the technical overview of the system deployed and also insight into the surrounding tools and processes that have been identified and devised to ensure the effective uptake by the data publishing community.

1.2 Progress towards objectives

The objectives of the EU BON registry and metadata catalogue are the following:

- To build upon the existing GBIF and LTER registries and metadata catalogues
- To connect these catalogues to the GEOSS Common Infrastructure
- Support the registration of entities such as networks, projects, sites and datasets
- Expose the entities through a web service interface for machine access
- Provide a unified access to heterogeneous data

1.3 Achievements and current status

This document consists of a general architectural overview describing the deployed system, followed by 4 sections detailing specific progress.

1. The first section details the enhancements to the GBIF Registry which included:
 - a. A revision of the underlying database in PostgreSQL and development of web services exposed through a RESTful API and an Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH) interface.
 - b. Development of a near real-time data indexing system providing metrics and visualisations such as maps and charts to help build compelling reasons for publishers to share data.
 - c. A revision of the data model to better support data provenance and accurately credit the various institutions involved in data publishing and technical hosting.
 - d. Collecting metrics about the access patterns of datasets and exposing this as a set of services to the data publishers.
 - e. Generating national reports (produced quarterly) about the nature of data published. These provide graphs showing changes over time for a variety of dimensions and are now used as indicators by the CBD for Aichi Target 19¹ ².

¹ <http://cbd.int/gbo4>

² <http://www.bipindicators.net/numberofgbifrecordsvertime>

- f. Promoting a consistent citation format built around Digital Object Identifiers (DOI) and assigning DOIs to all datasets, with a view to linking cited use in scientific papers to source datasets in the future.
 - g. Implementing a versioning mechanism for the data vocabularies used during the publishing and indexing process.
 - h. Developing a new data format to allow registration of sample based datasets such as those used in monitoring and survey studies. This new standard was deployed in the GBIF Registry to coincide with the GBIF Integrated Publishing Toolkit release v2.3.
2. The second section details the deployment of a Data Access Broker (DAB) to connect the GBIF and LTER registries and EU BON testing sites to the wider EU BON informatics architecture. The DAB is a deployment of GI-cat with standard accessors allowing connection to common formats of web services, such as the Open Geospatial Consortium (OGC) *W*S* services, and custom adapters for the GBIF and LTER registries. The GI-cat installation provides the brokering of registered entities into the GEOSS Common infrastructure and semantic querying. In developing this the following activities were undertaken:
 - a. Deploy GI-cat and GI-axe instances in CSIC's servers, both available through EUBON portal test subdomain (<http://test-eubon.ebd.csic.es/gi-cat/>, <http://test-eubon.ebd.csic.es/gi-axe/>).
 - b. Configure a subset of the network entities recommended for consideration in the MS241 Annex 1 section, either harvesting or linking them as an external web service. In particular, GBIF datasets were harvested through its OAI-PMH endpoint.
 - c. Join efforts with LTER developers and technical representatives in order to modify current endpoints to be consumable by Gi-cat through standardised services.
 - d. Join efforts with ESSI-Lab GI-cat developers, in order to modify GI-cat EML accessor to retrieve metadata from EML harvest lists.
3. The third section details the activities that have connected significant systems or datasets to the EU BON registry. This includes:
 - a. The connection of the citizen science platform (PlutoF) developed by Tartu Ülikool through the GBIF Registry.
 - b. Connection of the Pensoft Biodiversity Data Journal to EU BON through the GBIF Registry.
 - c. Connection of the Plazi treatment database to EU BON through the GBIF Registry.
 - d. Harvesting of LTER-Europe DEIMS datasets through GI-cat, using a new accessor for EML files and harvest lists.
 - e. Harvesting of GBIF dataset registry through GI-cat, using the OAI-PMH 2.0 endpoint provided by GBIF.
4. The fourth section lists future developments that will use the platform described in this deliverable as a basis to provide improved services for citations, licensing and data visualization (EBV Explorer).

2 Overall architecture

2.1 Introduction

Task 2.4 (Metadata registry and catalogue) is described as follows in the Description of Work (DoW).

Building on the existing GBIF and LTER registry and metadata catalogues, an enhanced and integrated metadata system will be developed for EU BON. The various entities such as networks, projects, sites, and datasets identified in the analysis and mobilization efforts of WP1 will be described in the new registry/catalogue. The entity descriptions should include web service interfaces or other access points, and will also be registered at the GCI and other indexing services. In order to overcome heterogeneity of data, accommodate multilingualism, enhance discoverability and interoperability, and facilitate querying in portals, the use of Knowledge Organisations Systems (KOS; e.g., thesauri) will be explored. (Lead GBIF; UEF, CSIC, Pensoft, MRAC, INPA, IBSAS; Months 9-51)

Central to the efficient functioning of any network is the presence of a registry. This can be thought of as a database with human and web service interfaces for the registration and curation of network entities. The registry stores inter-relationships of entities (e.g., which institution hosts which data sets), enabling discovery and access (via cached data and/or technical access points). This information is typically captured in metadata records that are stored in a metadata catalogue (a metadata database). To make clear the distinction between a metadata catalogue and a registry, the latter should provide machine interfaces and authoritative content. Therefore it must be well curated by data managers whilst being open enough to easily allow connection.

The requirements as stated in the EU BON DoW recognised the need to draw on already existing registries and catalogues such as GBIF and LTER and to integrate these with the GEOSS Common Infrastructure (GCI). It also indicated the kinds of entities that should be documented, including not only datasets but also related networks, projects and sites. The complete package of work for the EU BON registry includes new infrastructure to connect and integrate the GBIF and LTER registries with the GCI and also enhancement of the GBIF registry itself to better allow it to act as a broker for the EU BON network.

2.2 General registry architecture

The high level architecture of EU BON as proposed by CSIC (deliverable D2.1) relies on an Enterprise Service Bus (ESB)³ that is responsible for orchestrating workflows and connecting various data and service providers (EU BON D2.1 Architectural design, 2015). The registry and metadata catalogue form one component of the ESB and is implemented by GI-cat⁴ as a broker catalogue system (**Figure 1**), first introduced to EU BON partners in the Crete meeting (First EU BON Training Even, 2014) (S. Nativi, 2009). The GEO Data Access Broker (DAB) application, based on GI-cat, connects disparate information resources from different communities that are not aware of each other's interoperability mechanisms. It can be used to harvest content from existing registries and bring them under one view. In the case illustrated in **Figure 1**, the common data model (view) is based on the ISO 19115 metadata standard.

³ http://eubon.cybertaxonomy.africamuseum.be/sites/default/files/EU%20BON_EAI_SOA-presentation.pdf

⁴ <http://essi-lab.eu/do/view/GIcat/WebHome>

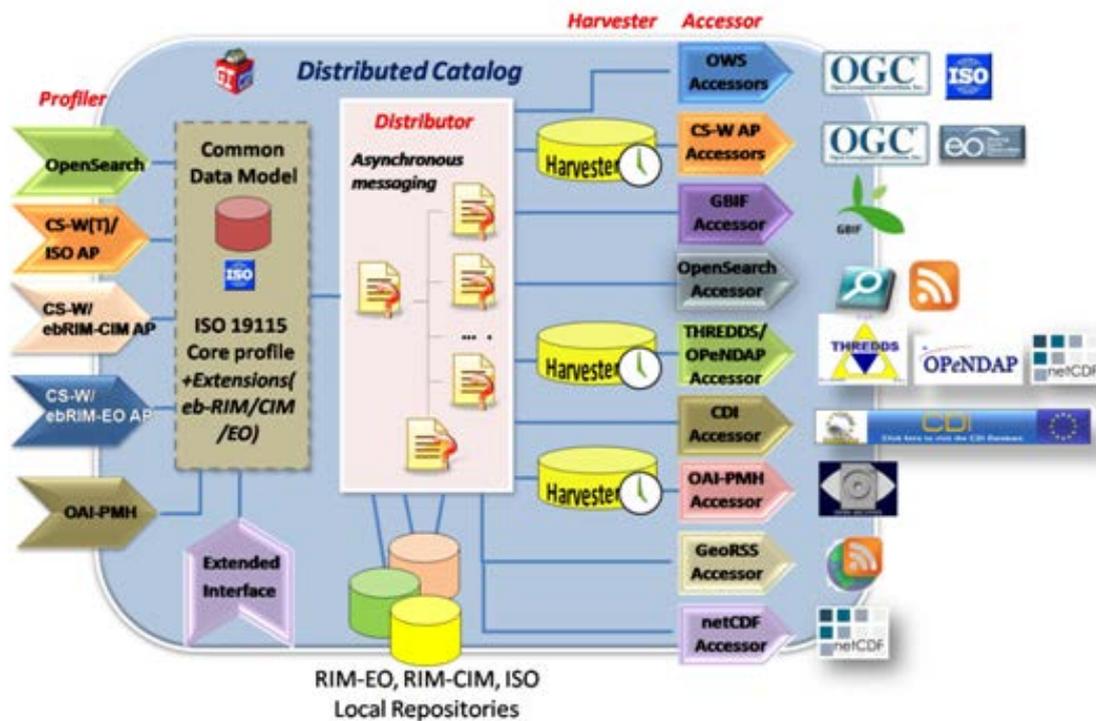


Figure 1 The GI-cat broker system featuring some catalogue query interfaces (right) and several backend mediation components (S. Nativi, 2009).

After reviewing the recommendation that GI-cat be evaluated as the brokering system for EU BON, CSIC determined that, as a specialized broker, GI-cat is a powerful solution for integrating metadata sources under a common data model and for providing an interface from EU BON to the GEOSS registry/catalogue. However, as it was lacking in connectors for common WSDL services and specific input sources needed for EU BON purposes (e.g., GenBank, EU-Nomen, WoRMS, etc.), a revision of the architecture was therefore proposed, consisting of a hybrid solution that integrates GI-cat inside a larger ESB based SOA architecture (**Figure 2**).

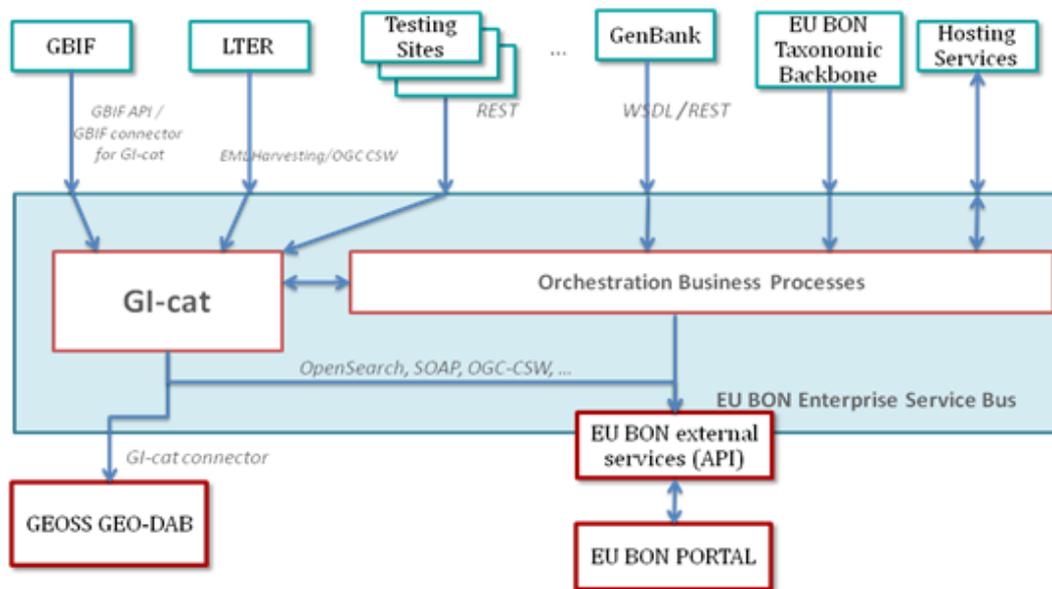


Figure 2 Proposed architecture for EU BON consisting of a hybrid solution in which GI-cat is integrated in the enterprise service bus.

The GI-cat broker system as described thus serves as an essential component of the EU BON's European Biodiversity Portal for registering and connecting the partner systems of EU BON. As described in the EU BON DoW, the European Biodiversity Portal (task 2.5)

“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 in task 2.3, functions developed by task 2.4. It will provide access to full detailed data, geographic visualisations and remotely sensed data. It will be closely linked to the GCI and GEO Portal, and access layers and data from GEOSS sources”.

The broker design does not impose common APIs on each participating partner system. Rather, while communities are encouraged to adopt, where appropriate, well known and widely deployed standards such as OGC web services, in their absence, accessors for each system have been developed and centrally managed within GI-cat.

It was recognised that there is significant burden in developing custom adapters for each network and thus partners were encouraged to make use of existing registries (e.g., GBIF, DataONE) to ensure that only a few accessors were actually needed (**Figure 3**). Not only did this reduce the technical complexity involved, but also helps ensure the necessary helpdesk activities are in place, and that the system would be maintained beyond the life of the EU BON project.

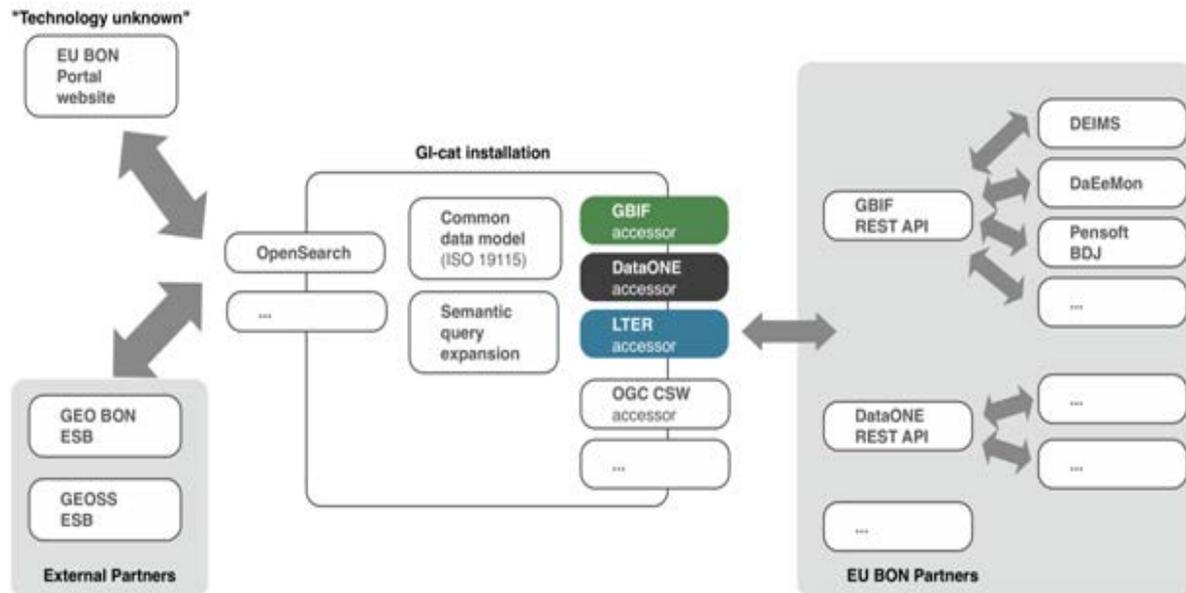


Figure 3 The architecture for the EU BON registry as deployed using GI-cat.

2.3 Common data model

The common data model supported by GI-cat is based on the OGC 19115 geospatial metadata standard. Widely promoted and adopted, e.g., within the EU INSPIRE framework⁵, this specification is well suited for describing geospatial resources in general (e.g., temporal and geospatial coverages) but is somewhat lacking in expressivity for certain aspects of biodiversity related resources, e.g., for describing taxonomic coverage only in a flat keyword list and detailed field sampling methodologies.

Last versions of GI-cat incorporate a new EML accessor that facilitates the consumption of either single EML files or EML harvest lists, as described in LTER DEIMS⁶. This accessor translates EML metadata into ISO-19115/ISO 19139 metadata model; in particular, the taxonomic coverage is translated into metadata keywords, which is not optimal but sufficient to discover datasets, thus the extension of the core metadata model of GI-cat may result not necessary for EU BON purposes. The GI-portal, on the other hand includes a semantic browser that utilises the GEMET thesaurus, which approach might work for taxonomies and needs to be explored. Nevertheless, this extension and the semantic functions of the GI-portal will be assessed in collaboration with ESSI-Lab, main developers of the GI-cat and GEO DAB components.

3 Enhancing the GBIF Registry

The requirements as stated in the EU BON DoW recognised the need to draw on and build upon the work previously achieved with the GBIF Registry. At the outset of the EU BON project, the GBIF Registry served the need of the GBIF network, but lacked an open API to allow other systems to connect and had several issues relating to the MySQL data model and data validations performed. Today the GBIF Registry is a robust database, connected to many data publishing systems through a RESTful API and to a real time data indexing service and has a fully functional helpdesk support

⁵ http://inspire.ec.europa.eu/documents/Metadata/MD_IR_and_ISO_20131029.pdf

⁶ <https://data.lter-europe.net/deims/eml/harvest-list-all.xml>

through the GBIF Secretariat. This section details the enhancements that were implemented to the GBIF Registry as part of the EU BON project.

The role of the GBIF Registry can be summarised as a component offering:

- An authoritative source of information (metadata) on institutions, datasets, networks, technical services and other key entities as required by registry partners. Due to the nature of the network and tools in use, multiple versions of this information are often available. Where this occurs, the registry aims to provide the most complete representation by merging sources and harmonizing conflicting views where possible. This simplifies consumption to clients by providing a unified view of metadata in a consistent format. Links to external representations available through other formats (e.g. the Ecological Metadata Language⁷) are available to clients.
- A source of information on inter-relationships between datasets, institutions and other entities. A common need relates to the hosting arrangement where one party hosts a Dataset on behalf of another, which might itself be a superset of other datasets. Modelling of dataset relationships provides an indication of where duplicate content might exist and how to correctly determine the attribution chain for all parties involved in the data management lifecycle.
- A trustworthy identifier assignment (minting) service for institutions and datasets. Identifiers are allocated a Universally Unique Identifier (UUID) on first registration, and a DOI is issued if none exists already for the dataset.
 - An identifier resolution service allowing external clients to submit a known identifier and resolve this to the registry assigned identifier. Thus clients already using (e.g.) a Global Registry of Biodiversity Repositories (GRBio) identifier⁸ can interact with the registry using those identifiers if registered. The number of identifier systems recognised is expected to grow continuously as more systems are connected.
- A mechanism to help coordinate distributed system activities by
 - Providing preferred technical access points where multiple routes exist;
 - Offering stable identifiers for registered entities and
 - Providing notification services of significant events such as a dataset being registered.
- A discovery mechanism for users and machines for
 - Registered network entities;
 - Technical endpoints and
 - Data definitions (e.g. Standards) such as the extensions and vocabularies used in the Darwin Core Archive⁹ format.
- Discovery is provided through indexing of metadata in the EML standard, and through flexible tagging of entities using simple key value pairs of tags, optionally in a restricted namespace. Tagging may be done publicly (no namespace), allowing anyone to make use of the tag, or by maintaining a private collection of tags (private namespace). Private tagging

⁷ <https://knb.ecoinformatics.org/#external/emlparser/docs/index.html>

⁸ <http://grbio.org/>

⁹ <http://www.gbif.org/resource/80636>

ensures a registry client can define their own terms (vocabulary) for tagging and be assured others cannot make changes to their tags.

- Metrics on the nature of data within a dataset, and metrics about how data is accessed through the GBIF.org data system are provided on a Dataset level.
- National analytical reports that summarising changes to data available over time are published periodically.

The software for the GBIF Registry is available on GitHub¹⁰ and the API is documented publically¹¹.

3.1 Architectural overview

At the heart of the registry is a PostgreSQL database¹² accessed exclusively through a web services API. To enable faceted search, an embedded Apache SOLR index is maintained and exposed in the API. The registry emits messages to a messaging bus (RabbitMQ) to enable components to subscribe to significant events, such as a newly registered dataset to be crawled. This is depicted in **Figure 4**.

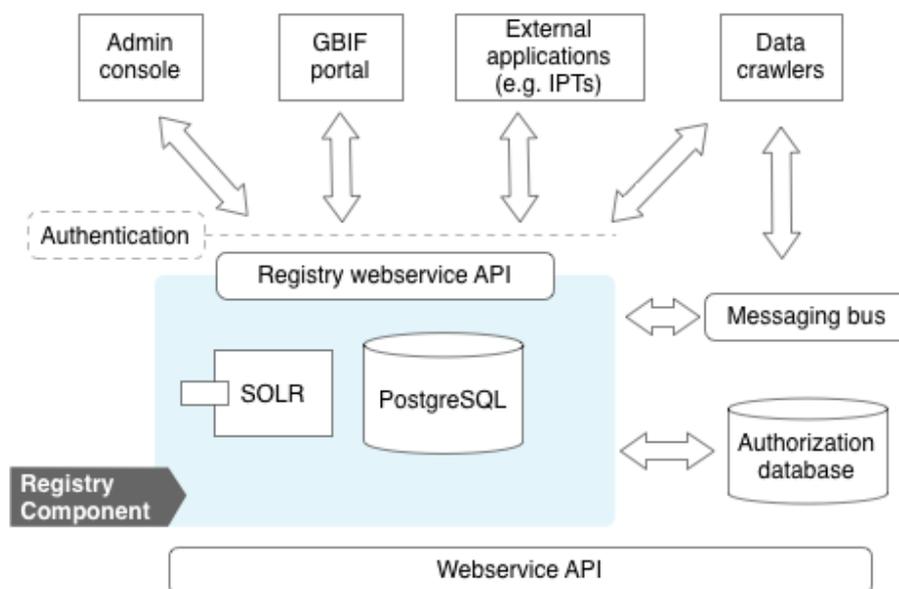


Figure 4: High level overview of GBIF Registry Architecture

All access with the GBIF Registry is through the web services API enabling external systems to interact. Authentication is provided through HTTP basic Authentication for individuals and a token-based approach for external systems. Authorization is scoped by network entity whereby systems are granted permission to create and edit sub entities under their domain. For example, an institution is able to create and edit datasets and their metadata that are explicitly owned by themselves, but cannot edit the Dataset entities connected to another institution. In general, the Registry API contains an extensive set of functions that allow to search, retrieve and update information about: datasets, dataset contacts, institutions, organizations, nodes, technical installations and GBIF downloads.

¹⁰ <https://github.com/gbif/registry>

¹¹ <http://www.gbif.org/developer/registry>

¹² The database schema and change history is shown on <https://github.com/gbif/registry/tree/master/registry-ws/src/main/resources/liquibase>

3.2 Real time data indexing and visualization

The key standards used to describe datasets are the Dublin Core, ISO 19115, FGDC geospatial standards and the Ecological Metadata Language (EML). Of these only the EML standard allows the author to document biodiversity domain specific content, such as the taxonomic coverage. All of these standards share a common limitation that they enable only very basic discovery means, and in practice it is common that metadata is sparsely populated. Recognising this limitation the GBIF Registry is connected to a near real-time data indexing system. As datasets are registered and when data is made available in a recognised standard, the indexing system downloads, parses, performs quality control mechanisms and indexes the content. This enables several things. Firstly, it enables richer discovery of relevant datasets. The metadata might accurately list the taxonomic families and geographic extent of the data, but cannot indicate if the dataset has information about a species at a specific location. The real time indexing system however does index content fully, and is able to provide answers to this question. Secondly, the indexing process is able to provide metrics about the nature of the data contained to help allow users determine if the dataset is of interest (see **Table 1**). These are surfaced through GBIF.org¹³ and through the API and include:

Table 1 Datasets statistics

Type of dataset	Metric	Description
Occurrence	Kingdoms covered	An indication of the biases across kingdoms in the dataset
	Basis of record covered	An indication of nature and amount of data shared – e.g. Observation versus fossil collections
	Issues in interpretation of data	A summary of the amount of data that has been flagged with quality issues
	Kingdom, basis of record and geo-referencing cube	A 3 dimensional summary indicating the biases between kingdom, the nature of data (basis of record) and whether the data has coordinate information
Taxonomic checklist	Kingdoms covered	An indication of the biases across kingdoms in the dataset
	Taxonomic ranks	The number of taxa per rank within the dataset
	Issues in interpretation of data	A summary of the amount of data that has been flagged with quality issues
	Overlap with GBIF and the Catalogue of Life	A summary of how much the species described overlap with the GBIF backbone and Catalogue of Life taxonomies
	Common names, distributions and descriptions	A summary of how many species have supplementary data

¹³ Example: <http://www.gbif.org/dataset/271c444f-f8d8-4986-b748-e7367755c0c1/stats>

Thirdly, by indexing data, the system is able to generate map visualizations showing more detailed views of the geographic coverage, adjustable by time, than is possible through only the metadata standards. These help provide a better picture about the nature of the dataset to assist users in determining if it is of interest.

3.3 Promoting consistent citation and tracking of use

In February 2015, GBIF began issuing DOIs for all newly published datasets and began to recognize and display publisher-assigned DOIs for existing datasets. GBIF then retroactively issued DOIs to all existing datasets that were missing a publisher-assigned DOI. In that way, all datasets in GBIF.org are now assigned a DOI.

User downloads on GBIF.org also now receive GBIF-issued DOIs. The citation file bundled with each download explicitly lists the DOI of each dataset referenced. This approach significantly simplifies references to any and all datasets represented in user-defined search results, even complex ones comprising occurrences from many different datasets.

DOIs for datasets and downloads gives users a stable, easy-to-use model for citing data sources. Updates to GBIF.org support this new citation model by displaying DOIs both for datasets and for user downloads. This approach also improves publishers' ability to track how and where users apply their data, in both research and web applications.

In March 2015, version 2.2 of the GBIF IPT was released capable of automatically connecting with either DataCite or EZID to assign DOIs to datasets. This version also enabled the citation to be auto-generated for a dataset in a format based on DataCite's preferred citation format [<https://github.com/gbif/ipt/wiki/IPT2Citation.wiki>], which includes the dataset version and link to the dataset homepage (e.g. DOI). Human readers of the citation can thus locate the exact version of the dataset cited, facilitating scientific reproducibility.

3.4 Monitoring data trends over time

Building an authoritative registry of datasets requires consideration to both technical and social aspects. It is not enough to provide a technical infrastructure but rather a lot of engagement activities need to succeed in order to populate the registry with the necessary content. To help support those championing open data and provide means to quantify their mobilization efforts, a data monitoring system was developed and processes put in place to recalculate the metrics quarterly. The metrics are presented as a series of charts illustrating the amount of data mobilized, and are available for all data¹⁴ (e.g. globally) and also for each country¹⁵. **Figure 5** is one such chart illustrating the total growth in species occurrence records over time globally.

¹⁴ <http://www.gbif.org/analytics/global>

¹⁵ <http://www.gbif.org/analytics/country/published>

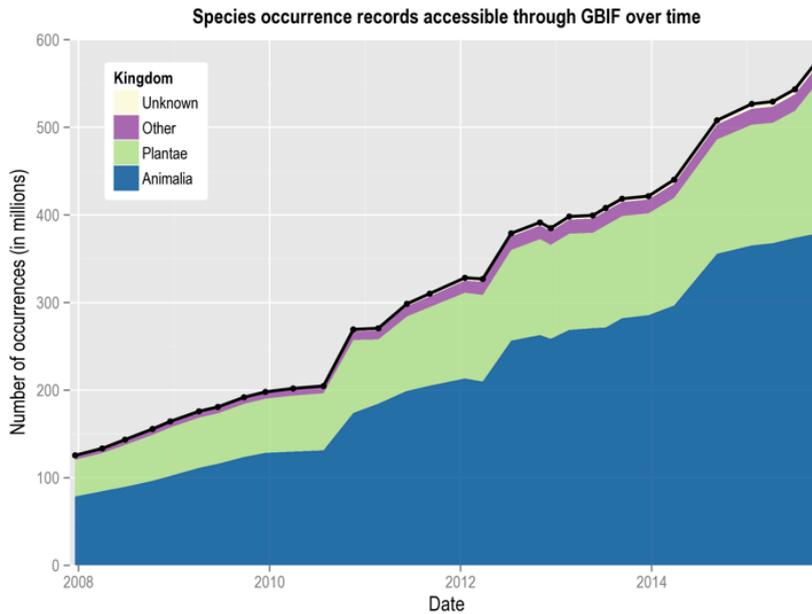


Figure 5 Growth in data registered in GBIF over time

Each quarter, the metrics are recalculated using the latest data quality interpretation routines and taxonomic backbone. Reprocessing these data each time is required to ensure that the charts reflect change over time in data availability without influence of improving data interpretation technique. The following summarises the metrics that are tracked over time for each country and globally (**Table 2**).

Table 2 Dataset metrics

Metric	Description
Records by kingdom	The number of available records categorized by kingdom.
Records for Animalia	The number of animal records categorized by the basis of record.
Records for Plantae	The number of plant records categorized by the basis of record.
Species count by kingdom	The number of species with available occurrence records, categorized by kingdom.
Species count for specimen records	The number of species associated with specimen records.
Species count for observation records	The number of species associated with observation records.
Records by year of occurrence	The number of occurrence records available for each year since 1950.

Species by year of occurrence	The number of species ¹⁶ for which records are available for each year since 1950.
Records by day of year	The number of occurrence records available for each day of the year.
Species by day of year	The number of species for which records are available for each day of the year.
Completeness	Indicate changes in the completeness of records when completeness is defined to include an identification at least to species rank, valid coordinates, a full date of occurrence and a given basis of record (e.g. Observation, specimen etc.).
Taxonomic precision	Indicate changes in the number of available records, which include an identification at least to the species rank. The numbers of records identified to an infra-specific rank or to a genus are also shown.
Geographic precision	Indicate changes in the number of available records, which include coordinates for which no known issues have been detected.
Temporal precision	Indicate changes in the number of available records, which include a complete date including year, month and day.
Geographic coverage for recorded species	Indicate changes in the number of species for which records are available from a range of localities. The earth's surface is divided into a series of increasingly fine grids (one degree, half degree and tenth of a degree). Species are categorised according to the number of cells in each of these grids for which GBIF has available data for the species since 1970. The charts show changes in the number of species recorded in only one such grid cell, in between two and twenty such grid cells, etc. Greater distribution of records typically increases the value of the data for various modelling activities.
Data sharing	Indicates trends in whether the data about biodiversity within a country is published by institutions within the country, or is available from abroad.

3.5 *Improving the management of data vocabularies*

During the publishing process data vocabularies are used to “map” data into a consistent format. The vocabularies provide the documented data and enable machines to read data tables and integrate data. These vocabularies fall under 2 categories:

1. Application schemas which documents the fields / columns that can be used in a data table and the type of information recorded in each row. A commonly used schema is the species

¹⁶ Species counts are based on the number of binomial scientific names for which GBIF has received data records, organised as far as possible using synonyms recorded in key databases such as the Catalogue of Life. Since many names are not yet included in these databases, some proportion of these names will be unrecognised synonyms and do not represent valid species. Therefore these counts can be used as an indication of richness only, and do not represent true species counts. All data have been processed using the same, most recent, version of the common GBIF backbone taxonomy, and comparisons over time are therefore realistic.

occurrence schema¹⁷, which indicates each row represents a record documenting evidence of a species at a place and time.

2. Vocabularies indicating the range of values a single field should have. Commonly used examples are the basis of record vocabulary¹⁸ or the ISO 3166 standard listing the countries and their 2 digit codes.

During the initial design of the application schemas no mechanism for versioning was put in place, while the field level vocabularies were always declared as transient. The lack of versioning in the application schemas posed a significant issue as data standards have evolved over time. This was addressed by moving to strongly versioned schemas, whereby each schema is registered with the date it is issued which is included in the URI for the resource itself – an example being http://rs.gbif.org/core/dwc_occurrence_2015-07-02.xml. This change in registration scheme was coordinated with the release of the version 2.3 of the Integrated Publishing Toolkit.

3.6 Vocabulary changes to support sample-based data

Since the significant revision in 2009, the Darwin Core vocabulary provided a rich set of terms, organised into several classes (e.g., Occurrence, Event, Location, Taxon, Identification). Many of these terms were relevant for describing sample-based data. Synthesising several sources of input (GBIF organised workshop on sample data, May 2013; discussions on the TDWG mailing list; discussions on the EU BON mailing list), a small set of terms relating to sample data were identified as essential, some of which are already present in the DwC vocabulary. These terms were:

1. *eventID*: an identifier for the set of information associated with an *Event*; may be a global unique identifier or an identifier specific to the data set.
2. *parentEventID**: an identifier for the broader Event that groups this and potentially other Events. May be a globally unique identifier or an identifier specific to the dataset.
3. *samplingProtocol*: the name of, reference to, or description of the method or protocol used during a sampling event.
4. *sampleSizeValue**: a numeric value for a measurement of the size (time duration, length, area, or volume) of a sample in a sampling event. A *sampleSizeValue* must have a corresponding *sampleSizeUnit*.
5. *sampleSizeUnit**: the unit of measurement of the size (time duration, length, area, or volume) of a sample in a sampling event. A *sampleSizeUnit* must have a corresponding *sampleSizeValue*.
6. *organismQuantity**: a number or enumeration value for the quantity of organisms. An *organismQuantity* must have a corresponding *organismQuantityType*.
7. *organismQuantityType**: the type of quantification system used for the quantity of organisms. An *organismQuantityType* must have a corresponding *organismQuantity*.

Five of the seven terms are new. Four of them are required to be used in pairs: *sampleSizeValue* with *sampleSizeUnit*, *organismQuantity* with *organismQuantityType*.

¹⁷ http://rs.gbif.org/core/dwc_occurrence_2015-07-02.xml

¹⁸ http://rs.gbif.org/vocabulary/dwc/basis_of_record.xml

Upon GBIF and EU BON proposal, these five new terms were ratified by the Biodiversity Informatics Standards organisation TDWG on 19 March 2015.

In order to encode sample-based data, a new core called the *Event* (i.e. sampling event) core¹⁹ was deployed for use with the Darwin Core Archive standard. This was accompanied by a revised Occurrence schema²⁰, which included two additional terms, *organismQuantity* and *organismQuantityType*. **Figure 6** shows a typical arrangement of data files for a Darwin Core Archive when using an *Event* core.

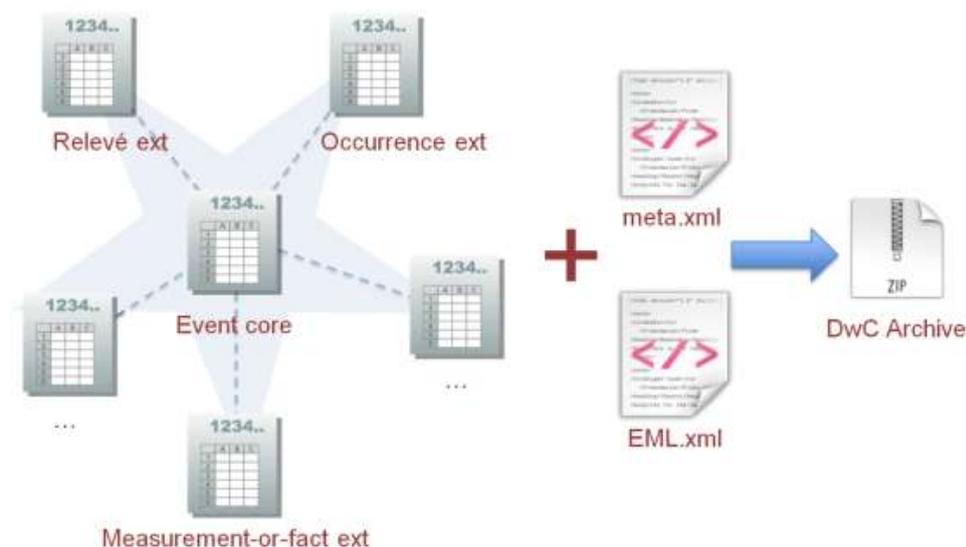


Figure 6 Example structure of Darwin Core Archive using an Event core

3.7 Data mobilization

Since 2013 around 163,631,617 occurrence records, coming from European institutions, have been mobilised through the GBIF network. Those records are distributed in 12,585 datasets, the distributions of datasets and publishers per country are listed in the following table (see **Table 3**):

Table 3 Distribution of dataset and publishers per country

Country	# Of Datasets	# Of Publishers
Germany	9869	32
United Kingdom	1338	13
Bulgaria	206	3
Spain	199	76
France	158	41
Nederland	145	28
Ireland	125	1
Norway	105	6

¹⁹ http://rs.gbif.org/core/dwc_event_2015_05_29.xml

²⁰ http://rs.gbif.org/core/dwc_occurrence_2015-07-02.xml

Poland	97	28
Belgium	87	11
Denmark	64	21
Finland	56	11
Sweden	43	8
Switzerland	21	11
Portugal	16	9
Austria	13	12
Estonia	10	3
Russia	7	4
Andorra	7	1
Slovenia	5	3
Czech Republic	5	1
Iceland	4	1
Slovakia	1	1
Croatia	1	1
Hungary	1	1
Italy	1	1
Luxembourg	1	1
Total	12,585	329

The European datasets constitute one of the main sources of occurrence data indexed in the GBIF Portal, those datasets contains valuable information of biological collections not only about European data but also about data mobilized from other regions, the table below (**Table 4**) lists the largest (in terms of occurrence records) datasets published since 2013 in the GBIF network.

Table 4 Sample of European datasets

Dataset/Publisher/DOI	Country	Records
Artdata <i>Published by: ArtDatabanken</i> http://doi.org/10.15468/kllkyl	Sweden	40,806,330
INPN – Donnes flore des CBN agrges par la FCBN <i>Published by: SPN – Service du Patrimoine naturel, Musum national d’Histoire naturelle, Paris</i> http://doi.org/10.15468/omae84	France	20,976,931
Dutch Vegetation Database (LVD) <i>Published by: Alterra, Wageningen UR</i>	Nederland	9,767,671

http://doi.org/10.15468/ksqjep		
Tiira information service <i>Published by: Birdlife Finland</i> http://doi.org/10.15468/p09mpw	Finland	5,198,246
Norwegian Species Observation Service <i>Published by: The Norwegian Biodiversity Information Centre (NBIC)</i> http://doi.org/10.15468/zjbzel	Norway	4,880,875
Botanical Society of the British Isles - Vascular Plants Database additions since 2000 <i>Published by: UK National Biodiversity Network</i> http://doi.org/10.15468/hogxp7	United Kingdom	4,801,101
Geographically tagged INSDC sequences <i>Published by: European Molecular Biology Laboratory (EMBL)</i> http://doi.org/10.15468/cndomy	United Kingdom	4,648,711
Naturalis Biodiversity Center (NL) – Botany <i>Published by: Naturalis Biodiversity Center</i> http://doi.org/10.15468/ib5ypt	Nederland	4,519,597
Naturgucker <i>Published by: naturgucker.de</i> http://doi.org/10.15468/uc1apo	Germany	4,285,817
Florabank1 - A grid-based database on vascular plant distribution in the northern part of Belgium (Flanders and the Brussels Capital region) <i>Published by: Research Institute for Nature and Forest (INBO)</i> http://doi.org/10.3897/phytokeys.12.2849	Belgium	3,660,135
Bird tracking - GPS tracking of Lesser Black-backed Gulls and Herring Gulls breeding at the southern North Sea coast <i>Published by: Research Institute for Nature and Forest (INBO)</i> http://doi.org/10.15468/02omly	Belgium	2,483,064
British Bryological Society - Bryophyte data for Great Britain from the British Bryological Society held by BRC <i>Published by: UK National Biodiversity Network</i> http://doi.org/10.15468/gvqhjb	United Kingdom	2,419,291
Suffolk Biological Records Centre - Suffolk Biological Records Centre (SBRC) dataset <i>Published by: UK National Biodiversity Network</i> http://doi.org/10.15468/ab4vwo	United Kingdom	2,120,907
Vascular plants in Denmark recorded under the The Nationwide Monitoring and Assessment Programme for the Aquatic and Terrestrial Environments (NOVANA) <i>Published by: Danish Nature Agency</i> http://doi.org/10.15468/m40vfk	Denmark	1,854,895
Bristol Regional Environmental Records Centre - BRERC October 2009	United Kingdom	1,581,060

<i>Published by: UK National Biodiversity</i> http://doi.org/10.15468/vntgox		
Network Vegetation data from protected areas in Denmark (3 in the Danish Nature Protection Act) <i>Published by: Department of Bioscience, Aarhus University</i> http://doi.org/10.15468/ar7pbr	Denmark	1,116,766
Banco de Datos de la Biodiversidad de la Comunitat Valenciana <i>Published by: Biodiversity data bank of Generalitat Valenciana</i> http://doi.org/10.15468/b4yqdy	Spain	1,035,068
Phanerogamic Botanical Collections (S) <i>Published by: GBIF-Sweden</i> http://doi.org/10.15468/yo3mmu	Sweden	1,005,058
FLORIVON <i>Published by: Dutch Foundation for Botanical Research (FLORON)</i> http://doi.org/10.15468/ke2ody	Nederland	900,519
Vlinderdatabank - Butterflies in Flanders and the Brussels Capital Region, Belgium <i>Published by: Research Institute for Nature and Forest (INBO)</i> http://doi.org/10.15468/njgbmh	Belgium	761,660
PlutoF platform observations <i>Published by: Natural History Museum, University of Tartu</i> http://doi.org/10.15156/bio/587440	Estonia	741,963
Fundación Biodiversidad, Real Jardín Botánico (CSIC): Anthos. Sistema de Información de las plantas de España <i>Published by: Anthos: Spanish Plants Information System, Biodiversity Foundation-Royal Botanical Garden, CSIC</i> http://doi.org/10.15468/4wnutv	Spain	720,399
Finnish Entomological Database: Lepidoptera <i>Published by: Finnish Museum of Natural History</i> http://doi.org/10.15468/nojfbd	Finland	708,284
Shropshire Ecological Data Network - Shropshire Ecological Data Network Database <i>Published by: UK National Biodiversity Network</i> http://doi.org/10.15468/bmwtox	United Kingdom	701,889

4 The Data Access Broker

GI-cat is a specialised brokering system developed for the GEOSS portal within the context of the EuroGEOSS project, subsequently becoming the core of the GEO Discovery and Access Broker. It is comprised of a JavaEE application that acts as a broker among standardised sources and catalogues, and a set of libraries and connectors to provide standardised input and output capabilities.

GI-cat common data model is based on ISO 19115 plus extensions (see **Figure 7**). The distributor translates provided data and metadata to this data model, and exposes the information translating again to several catalogue formats. Consuming EML interfaces directly may require the development of a new GI-cat accessor.

As a specialised broker, GI-Cat can be configured as a direct-access mediation service or as a metadata harvester. In both approaches it can be connected to external data providers, acting as input sources modelled on a set of standardised formats, e.g. OGC (CSW, WMS, WFS), THREDDS, OAI-PMH, etc. It can provide a subset of standardised interfaces to act as an output provider, e.g., CSW output, OpenSearch interface, OAI-PMH harvesting services, etc.

As far as GI-Cat can be configured to provide OGC-CSW services to distribute the functionality and return search results, and given that OGC-CSW is supported by many information facilities, that single interface could be the output of EU BON to other biodiversity networks.

As mentioned before, translating EML metadata into ISO 19115 metadata is not totally straightforward, and in the case of the taxonomic coverage of each dataset, GI-cat transforms each taxon listed into ISO metadata keywords.

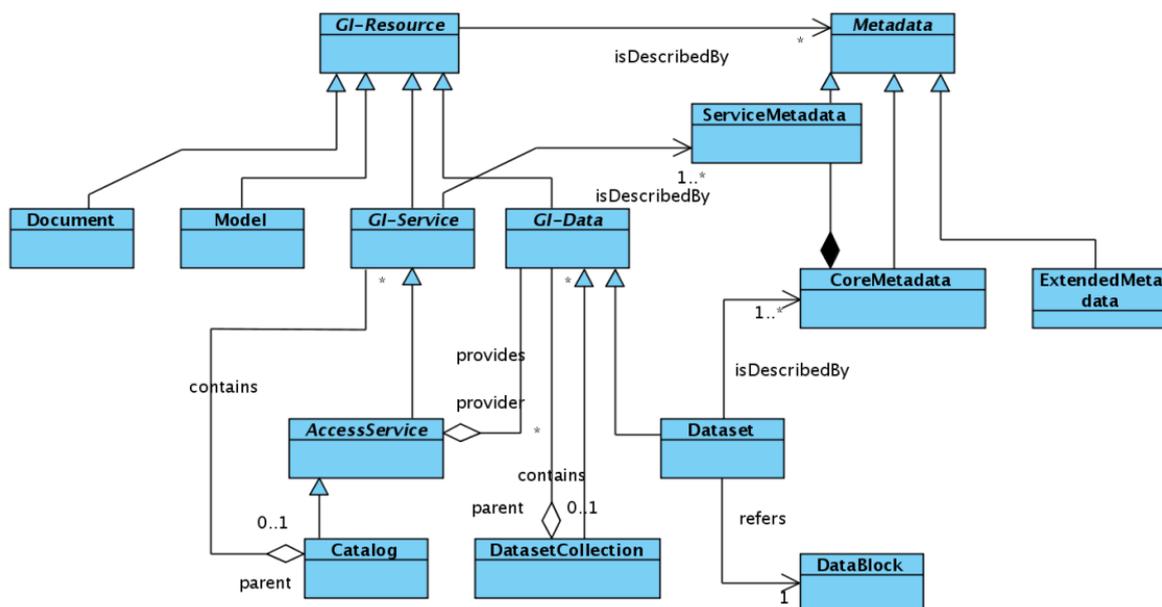
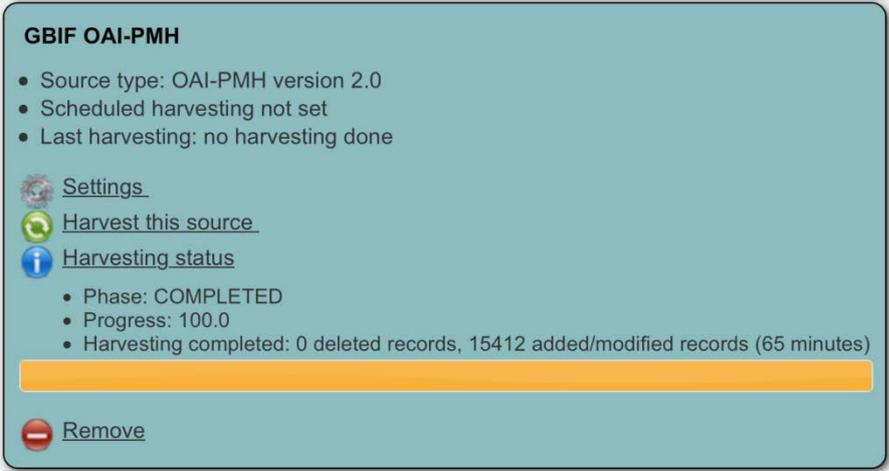


Figure 7 GI-Cat Data Model

Most EU BON test sites provide standardised interfaces that GI-cat could consume, directly or uploading metadata to broader biodiversity networks, in particular LTER and GBIF. Thus, consuming those network entities or harvesting them was stated as a main requirement of the EU BON metadata registry, and in particular the metadata broker.

In the case of GBIF, as far as they provided an OAI-PMH harvesting endpoint (see **Figure 8**), it could be configured in GI-cat, being able to harvest 15412 datasets, caching their metadata in the GI-cat local database



GBIF OAI-PMH

- Source type: OAI-PMH version 2.0
- Scheduled harvesting not set
- Last harvesting: no harvesting done

[Settings](#)

[Harvest this source](#)

[Harvesting status](#)

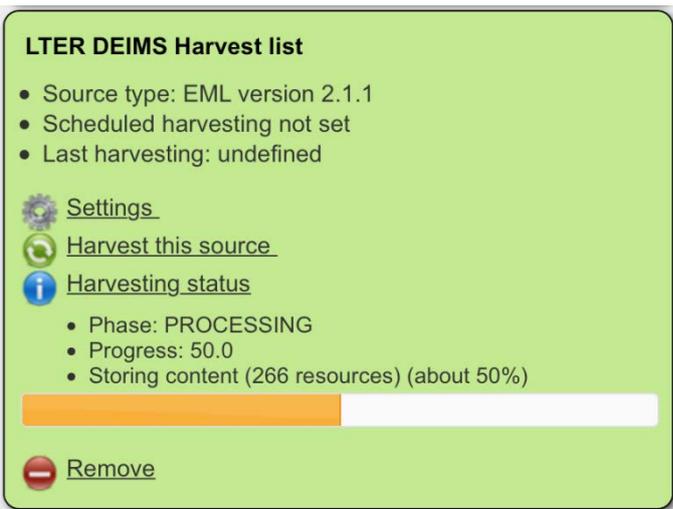
- Phase: COMPLETED
- Progress: 100.0
- Harvesting completed: 0 deleted records, 15412 added/modified records (65 minutes)

[Remove](#)

Figure 8 OAI-PMH Harvesting status

The integration of LTER DEIMS metadata was not straightforward in the first stages of the development. The first approach to the integration used a GeoNetwork instance, which stored DEIMS metadata and transformed it into the ISO metadata model, using XSLT stylesheets, and providing a OGC CSW repository as well. This CSW endpoint could be properly configured in GI-cat, but the taxonomic coverage sections were not included into the translation into the ISO model.

The second approach to the integration used a new EML accessor developed by ESSI-Lab and available in the last stable release of GI-cat (version 11.x). However, this approach required to incorporate EML files one to one, a particularity that could require to build external processes to automate the discovery of new EML files and their configuration in GI-cat. After envisaging feasible solutions and managing this issue with ESSI-Lab, they provided us with a new EML accessor, that could consume harvest lists directly. After configuring it properly, EU BON GI-cat instance could retrieve 425 datasets from DEIMS (see **Figure 9**), scheduled after the initial harvesting process.



LTER DEIMS Harvest list

- Source type: EML version 2.1.1
- Scheduled harvesting not set
- Last harvesting: undefined

[Settings](#)

[Harvest this source](#)

[Harvesting status](#)

- Phase: PROCESSING
- Progress: 50.0
- Storing content (266 resources) (about 50%)

[Remove](#)

Figure 9 Harvest list status

5 Connecting other systems

5.1 *Pensoft Biodiversity Data Journal*

Biodiversity Data Journal (BDJ) is a community peer-reviewed, open-access, online platform, designed to accelerate publishing, dissemination and sharing of biodiversity-related data of any kind. All structural elements of the articles – text, morphological descriptions, occurrences, data tables, etc. – are treated and stored as data, in accordance with the Data Publishing Policies and Guidelines of Pensoft Publishers.

All data published through the BDJ are made available openly and registered in the EU BON Registry. This is achieved by using the GBIF Registry as a broker to EU BON. The Biodiversity Data Journal connects to the GBIF Registry and automatically registers the metadata about datasets successfully completing the peer review process through the authenticated API. The data are exported, displayed and available for download to anyone in Darwin Core Archive format. This format is also used by GBIF to update the metadata and index the data itself in GBIF.org. To date, more than 160 dataset have been registered in EU BON through this mechanism²¹. The GBIF Registry recognises the existing DOIs for the datasets, and promotes those in the citation guidelines.

5.2 *Plazi treatment database*

Plazi is an association supporting and promoting the development and service of persistent and openly accessible digital taxonomic literature. Plazi tools allow a user to markup taxonomic treatments to become machine-readable. Different sections like type material, circumscriptions, references, distributions documented in literature become semantically accessible. These digitised data are made openly available by registering them in the EU BON Registry. This is achieved by using the GBIF Registry as a broker to EU BON. Plazi automatically registers and updates the metadata about datasets in GBIF. To date, more than 1188 treatments have been registered in GBIF through this mechanism²².

5.3 *Citizen science observations through PlutoF*

The PlutoF cloud based platform provides tools to create, manage, share, analyse and publish biology-related databases and projects. PlutoF mediated datasets have been connected to EU BON through GBIF. This has been achieved by using an installation²³ of the GBIF IPT to act as a repository and registration gateway. By using the GBIF IPT, the GBIF Registry and authorization mechanisms are handled automatically. Several datasets have been registered including the PlutoF citizen science data²⁴, which holds more than 700k records of species occurrence.

²¹ <http://www.gbif.org/publisher/750a8724-fa66-4c27-b645-bd58ac5ee010>

²² <http://www.gbif.org/publisher/7ce8aef0-9e92-11dc-8738-b8a03c50a862>

²³ <https://plutof.ut.ee/ipt/>

²⁴ <http://www.gbif.org/dataset/169fa761-2fb9-4022-93bd-e22b7a062efd>

5.4 *EuMon*

EuMon²⁵ was a FP6 project that ran from 2004 to 2008, and coordinated by the EU BON partner UFZ. EuMon developed a metadatabase of biodiversity monitoring schemes in Europe. This database, which is still maintained by the CKFF in Ljubljana, Slovenia, currently contains 649 entries, is the largest source of information of biodiversity monitoring activities in Europe. The EuMon metadatabase does not follow any metadata standard, nor did it have machine query facilities. In cooperation with the CKFF EU BON partners UFZ, MfN, and UEF developed these enhancements in 2015-2016. The new features include the following:

1. The EuMon database schema was extended by a number of new elements describing database and access rights.
2. Data entry forms were updated to cover the additional elements.
3. Updating of the database was enhanced so that the contact persons themselves are enabled and encouraged to update their data regularly.
4. Summary information of the new parameters was presented in graphs and statistics, as with already existing data.
5. As the EuMon database has its own data model, this was mapped to the Ecological Metadata Language (EML). Several EuMon fields have their direct counterpart in EML, but not all. Mapping to the various EML fields in an “EuMon-EML profile”.
6. The EML documents were made available through EML Harvest Lists, which is a de facto standard of the LTER network for sharing EML data.

This work is still partially underway in February 2016, but will be completed in near future. After that, all EuMon metadata will be available through the EU BON registry system at GBIF. A campaign is being planned to increase the number of entries in the metadatabase, and to approach them with a challenge to share their full data with EU BON.

6 Future developments

Taking as basis the registry and data publishing platforms described in this document, the following areas of future development have been identified:

- Provide a feed to data publishers of any cited use in scientific literature. As data are aggregated and then filtered by end users, it becomes unclear to data publishers which studies have made use in whole, or in part of their data. By promoting consistent DOI based citation in final use we aim to explore connections between aggregate data.
- Converge on a set of machine-readable open data licenses. GBIF have started a consultation with all data publishing parties and during 2016 aim to adopt a small set of licenses to remove ambiguity in scientific use.
- Development of visualisations to support Essential Biodiversity Variables. The first iteration will include a visualization targeting the Species Population class of EBV. A geographic and temporal explorer will be added to the Registry to allow discovery of datasets that can contribute data towards the EBV candidate for species distribution. In the first phase, this

²⁵ <http://eumon.ckff.si/>

will be built from an analysis of data shared through GBIF.org and then future editions will combine metadata from other datasets registered through EU BON. In subsequent developments, EBV keywords will be added to data publishing tools for data publishers to flag up. Currently, GBIF is working on extending the functionality of geospatial queries at the API level to support spatial user interfaces; it's estimated that by the end of March 2016 an initial version of the EBV Browser will be available to be evaluated by the EU BON community.

- EU BON will connect with the DataONE network. This will be achieved through GBIF developing and deploying a software stack and a repository to upload data, and possibly other EU BON partners connecting Metacat installations.
- Expansion of the Registry data model to support projects, sites and biological collections.
- Incorporation of more network entities to the GI-cat registry, in particular those entities publishing EML harvest lists (LTER Deims, Metacat).
- Transformation of existing non-standardised endpoints, or not compatible with GI-cat, to standardised services through XSL transformations and service orchestration.
- Assess and apply a likely taxonomic coverage expansion through taxonomic hierarchy injection and EU BON taxonomic backbone web service calls.
- Knowledge Organisation Systems offer a new possibility to organize biodiversity metadata. In particular, the Biological Collections Ontology (Walls et al, 2014) suggests new ways of linking information, also in quantitative sampling. During the remaining project time, these possibilities will be explored.

7 References

1. **EU BON D2.1 Architectural design.** (2015, 2 20). From D2.1 *Architectural design, review and guidelines for using standards*: http://eubon.eu/getatt.php?filename=EU_BON_D2.1_Architectural_design_review_and_guidelines_for_using_standards_10716.pdf
2. **First EU BON Training Event.** (2014, April 3). *First EU BON Training Event*, Heraklion, Crete, April 3rd 2014. *Information architecture – GEOSS perspective - Lorenzo Bigagli.*: http://eubon.cybertaxonomy.africamuseum.be/sites/default/files/uploaded_files/Bigagli_slides.pdf
3. **Nativi, S., Bigagli, L. (2009).** Discovery, Mediation, and Access Services for Earth Observation Data. In *Selected Topics in Applied Earth Observations and Remote Sensing, IEEE Journal of 2 (4)*, 233-240. doi: 10.1109/JSTARS.2009.2028584
4. **Walls, R.L., Deck, J., Guralnick, R. et al. (2014).** Semantics in Support of Biodiversity Knowledge Discovery: An Introduction to the Biological Collections Ontology and Related Ontologies. *PLOS One*. doi: 10.1371/journal.pone.0089606