



Deliverable 2.2 (D2.2)

Data sharing tools

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: MRAC, Royal Museum for Central Africa

Authors: Larissa Smirnova, Patricia Mergen, Quentin Groom, Aaike De Wever (MRAC, Royal Museum for Central Africa)
 Pavel Stoev, Lyubomir Penev (Pensoft, Pensoft Publishers Ltd, Bulgaria)
 Israel Peer (GlueCAD, GlueCAD Ltd. – Engineering IT, Israel)
 Veljo Runnel (UTARTU, University of Tartu - Natural History Museum, Estonia)
 Antonio García Camacho (CSIC, Spanish Council for Scientific Research - Doñana Biological Station, Spain)
 Timoty Vincent (INPA, Brazil)
 Hannu Saarenmaa (UEF, University of Eastern Finland, SIB Labs - Digitalium, Finland)
 Donat Agosti (PLAZI, Switzerland)
 Christos Arvantidis (HCMR, Hellenic Centre for Marine Research, Greece)

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 content

1. Executive Summary	4
Introduction	4
Progress towards objectives	4
Achievements and current status	5
Future developments.....	6
2. Background and prerequisites – general introduction	7
3. Inventory of tools.....	16
Types of tools	16
Tools surveyed by EU BON.....	17
4. Requirements.....	18
User communities to be supported.....	18
Selected tools	19
Enhancements required	23
Testing.....	23
5. Implementation	26
GBIF Integrated Publishing Toolkit (IPT).....	26
DEIMS: Drupal Ecological Information Management System	30
Spreadsheet tools.....	31
Biodiversity Data Journal and ARPHA publishing platform	34
Plazi TreatmentBank and DwC	38
Metacat and Morpho.....	42
PlutoF	45
6. Future developments and conclusions.....	48
Challenges.....	48
7. Bibliographic references	54
Annex 1: Non-exhaustive list of tools	56

1. Executive Summary

Introduction

Biodiversity information is growing exponentially due to the expanding number of biodiversity related projects, initiatives, and networks collecting data around the world. A substantial portion of these data come from citizen science initiatives, and often differ from more "traditional" data collected by trained scientists. Mobilization and integration of data from such diverse origins is thus of major importance and is one of the key objectives of the EU BON project.

Data mobilization is a broad term that includes data sharing, data publishing, and involvement of scientific and citizen communities in data generation.

Cooperation of many tasks across the entire project has been required, including T1.5, T2.2, T2.4, T2.5, T2.7, T2.8, T3.4, T5.2, and T8.5.

Progress towards objectives

This report provides conceptual and practical advice for implementation of the available data sharing and data publishing tools enhanced or adopted by EU BON. The report begins with an introduction to the complex world of data, metadata, and data integration. The concepts of data sharing and data publishing are clarified. A comprehensive review of the existing tools for metadata, occurrence data, and ecological data is compiled. A detailed description of the tools, their pros and cons, is followed by recommendations on their deployment and enhancement.

This is done from the perspective of the needs of the biodiversity observation community with an eye on the development of a unified user interface to this data – the European Biodiversity Portal (EBP). We described the steps taken to develop, adapt, deploy and test these tools. This document also gives an overview of the objectives and challenges that still need to be achieved in the remaining part of the project.

After a detailed analysis of tool requirements, recommendations are given on what tools best satisfy the needs of different user groups within the biodiversity observation community. A small number of tools, name the GBIF Integrated Publishing Toolkit (IPT), spreadsheet tools, DEIMS, Metacat, the ARPHA Publishing Platform, TreatmentBank, and PlutoF were selected for deployment. Additional tools, which may be used for data sharing, such as those used by organizations to comply with the requirements of the INSPIRE¹ directive, have been included, as have other spatial analysis and crowd-sourcing tools. These tools also contribute significantly to the resources of the community which is why they have been included in this report.

¹ Infrastructure for Spatial Information in the European Community (<http://inspire.ec.europa.eu/>)

The main challenges identified are:

- there is a variety of tools but none can, in and of itself, satisfy all the requirements of the wide variety of data providers!
- gaps in data coverage and quality demand more effort from data mobilization.

To fully meet the user requirements a combination of tools have been selected, which, in the form of a work-flow, will mobilize data. Some of the tools are also used to further process the data, including paper publication. Outreach campaigns and training sessions have been organized and are planned in the future to target effort on data mobilization where gaps have been identified.

Achievements and current status

The conclusion was that the choice of tools should be defined by the needs of those observing biodiversity – the end user community in the broadest sense – from volunteer scientists (citizen scientists), exploring and recording life around them *via* their mobile devices, to decision makers looking for processable and reliable data to build reports and forecasts upon it.

Short description of selected tools:

GBIF IPT²: Tool to publish and share biodiversity data sets and metadata through the GBIF network. Allows publication of three types of biodiversity data: i) primary occurrence data (specimens, observations); ii) species checklists and taxonomies; iii) sample-based data from monitoring programs.

Spreadsheet tools: 1) GBIF Spreadsheet processor is a web application that supports publication of biodiversity data to the GBIF network using pre-configured Microsoft Excel spreadsheet templates; 2) DataUp tool is the tool developed by DataOne to help environmental scientists to upload files to a repository for data management.

The ARPHA Publishing Platform³: Narrative (text) and data integrated publishing workflow to mobilize, review, publish, store, disseminate, make interoperable, collate and re-use data through the act of scholarly publishing. Three types of biodiversity data supported: i) primary occurrence data (specimens, observations), ii) species checklists and taxonomies, iii) sample-based data from monitoring programs.

TreatmentBank⁴: A platform to store, annotate, access and distribute taxonomic treatments and the data objects within. It offers with GoldenGate⁵ and respective XML schemas

² <http://www.gbif.org/ipt>

³ <http://www.pensoft.net/>

⁴ <http://plazi.xuul.org/resources/treatmentbank/>

⁵ <http://plazi.org/?q=GoldenGATE>

(TaxonX⁶, TaxPub⁷) tools to convert unstructured text into semantically enhanced documents with an emphasis on taxonomic data like treatments, scientific names, materials observation, traits or bibliographic references.

Metacat⁸ and Morpho⁹: Metacat is a repository that helps scientists store metadata and data, search, understand and effectively use the data sets they manage or those created by others. A data provider using Metacat can become DataONE member node with a relatively simple configuration. Morpho is an application designed to facilitate the creation of metadata.

Implementing mobile app tools with the PlutoF API¹⁰: Online service to create, record, manage, share, analyze and mobilize biodiversity data. Data types cover ecology, taxonomy, metagenomics, nature conservation, natural history collections, etc.

Future developments

The data providing tools occupy a strategic interface between the data mobilization and making the data accessible and usable on the portals (both in terms of data, metadata or even already processed data). Future developments will thus go in the direction of minimizing the identified barriers to data mobilization on one hand and enhancing the workflow towards the stakeholders by filling the known gaps.

One such major gap, as reported from the gap analysis performed by WP1, is the time lag between published datasets, compared to the apparently huge number of those still hidden within the repositories of institutions. The number and the diversity of data and metadata standards in circulation may also be an obstacle to potential providers of biodiversity data. Likewise, the same is true for the diversity of software tools. Hence EU BON has focused on the empowerment of existing data sharing tools and standards by broadening their interoperability, connectivity and sharing capabilities, rather than adding new tools.

These further enhancements of the tools selected for their adequacy with the objectives of EU BON will be achieved in the next steps, by involving massively the different stakeholders and outreach to additional data providers. The work done at the testing sites will now be extended further to real life implementation of the identified tools in larger networks of GBIF, LTER, and LifeWatch, but also by encouraging smaller structures and individual researchers such as those identified by the EuMon project to use them. In this regard the helpdesk and the associated training activities will play a major role. The whole EU BON consortium is however also committed to contribute to the overall outreach efforts and be active in the implementation and enhancement of the selected data providing tools.

⁶ <http://plazi.org/?q=taxonx>

⁷ <https://github.com/tcatapano/TaxPub/releases>

⁸ <https://www.dataone.org/software-tools/metacat>

⁹ <https://www.dataone.org/software-tools/morpho>

¹⁰ <https://plutof.ut.ee/>

2. Background and prerequisites – general introduction

Biologists are joining the Big-Data club (Marx, 2013)¹¹. This comes about due to the efforts of genomics (molecular sequence data), but also as a result of biodiversity monitoring programs. Big Data are determined not only by the volume, but also by the variability and complexity of data (Fig.1). Life science disciplines are producing such variable and complex datasets that they can easily compete with other disciplines for the title of Big Data.

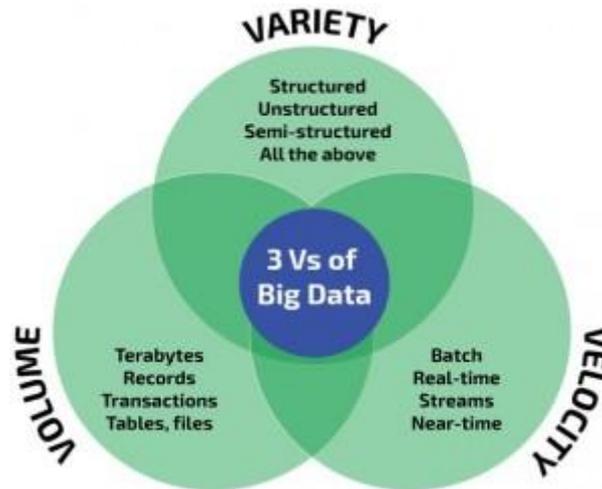


Figure 1. The Three V's of Big Data (borrowed from the “Big Data: Algorithms for Data Preprocessing, Computational Intelligence, and Imbalanced Classes”¹²).

Additional data sources come from citizen sciences initiatives, remote sensing, satellite imagery and the vast corpus of digital literature, which open new perspectives for data mining. This huge amount of data is of high scientific value and potential. It should be mobilized to become more accessible *via* data portals, such as the Global Biodiversity Information Facility¹³ (GBIF), Long-term Ecological Research Network¹⁴ (LTER), and DataONE¹⁵. Programs such as GEO BON¹⁶ and projects like EU BON, which belong to the Global Earth Observation System of Systems¹⁷ (GEOSS) use these primary data sources to detect change in biodiversity. These initiatives have identified data mobilization and integration as important goals.

¹¹ <http://www.nature.com/nature/journal/v498/n7453/full/498255a.html>

¹² <http://sci2s.ugr.es/BigData>

¹³ <http://www.gbif.org/>

¹⁴ <http://lternet.edu/>

¹⁵ <https://www.dataone.org/>

¹⁶ <http://geobon.org/>

¹⁷ <http://earthobservations.org/geoss.php>

The growing importance for the need to enhance and facilitate tools for access, sharing and publishing of biodiversity data is closely related with several factors:

- data explosion caused by mass digitization, computerization and public involvement, coined “crowd sourcing” or “citizen science” (which raises issues of data quality and standardization),
- climate change and the increasing loss of biodiversity raises the pressing need for more and accurate data to enable assessments, analyses of trends and traits in order to provide decision-makers with solid scientific-based recommendations and solutions,
- growing number of the intermediate agents (international initiatives, projects and infrastructures) designed to make the link between the data and policy easier, faster and more efficient by trying to fill in the gaps in data, mobilize data through boundaries and disciplines, provide the services to data providers (tools, standards, best practices, training).

Gathering, managing and analyzing of biodiversity data is demanding because: i) they include many different types of data; ii) the amount of data is large; iii) relevant data sources are fragmented and widely distributed, and iv) their coverage is often incomplete (Hoffmann, 2014).

Therefore, the EU BON Description of Work defines the task T2.3 as follows:

“This task will work with international partners (task 2.7) to scope the requirements and build new releases of data sharing tools for relevant data providers. These open source tools implement the selected interoperability mechanisms (task 2.2) and data publishing mechanisms (task 8.5) for use by the relevant networks, and provide registration and query functions towards the GCI. As the basis of development, existing tools for metadata, occurrence data and ecological data from GBIF and LTER will be used. New tools for sharing habitat data will be investigated. A model for distributed development will be adopted. (Lead MRAC; UTARTU, UEF, GBIF, Pensoft, Plazi, GlueCAD, INPA, IBSAS; Months 9-51)”

The initial project outputs were dedicated to the evaluation and gap analysis of different data sources and data types (deliverable D1.1), which allowed the production of further recommendations, best practices and solutions for the storage and management of selected biodiversity data types such as taxonomic backbone data, data stored in bio-repositories, species profile data, and citizen-science based data. Furthermore, a comprehensive analysis of the required information architecture and review of the available data standards was made (deliverable D2.1).

This report was preceded by a milestone (MS231) document in the spring of 2015 where an inventory of tools was made and a specification for the data sharing tools of interest to the EU BON project were laid out. In this report we extend the work by presenting a detailed assessment of the selected tools for data sharing and data publishing, the development and enhancement of the selected tools, and results of their testing in the real environments.

Definitions and concepts

Before assessing and selecting an appropriate tool for sharing data or metadata or any other data handling, first we need a good understanding of what these terms mean and how they are used in life sciences. Within the context of biodiversity informatics one operates with terms like “data”, “standard”, “sharing”, etc., but do we all give the same meaning to them? To eliminate misunderstanding and misuse of the terms, we first introduce the fundamental concepts and definitions.

Data

The many definitions and terms which include "Data" as part of their name, coined and documented in depth through numerous biodiversity infrastructures/interoperability projects, reflects the growing complexity in handling data flows and the increased need to formalize and categorize the multiple aspects of the notion of “data”. Furthermore, the integration of biodiversity data, which may include at least formats of genetic sequences, species occurrence (distribution/abundance/biomass/production) values and habitat maps, requires clear and unambiguous identifications of the terms for data.

Data are a set of values of quantitative measurement of, or a qualitative fact on some entity in a structure of known format (e.g. spatial and tabular), typically the results of measurements. It is people and computers who collect data and impose formats on it. From these formats, information patterns and interrelations can be derived and subsequently interpreted, a process which provides evidence, which can, in turn, be used to create or enhance knowledge.

Data are often assembled in discrete units of digital content, such as files or records in a database, often expected to represent information obtained from a particular observation, sample, location, or period of time during a scientific study. These discrete units of data may be further organized into a dataset, which is an organizational tool to present a coherent and complete collection of data relevant to a particular topic. A dataset may be a single file or database, or it may be composed of thousands of files, and it is possible for a single database to contain many datasets. The organization of data into files and datasets is generally not standardized and depends on the particular needs of the individuals collecting the data and the anticipated uses of that data.

In the context of biodiversity observation network the term data should be associated with the purpose and the context in which these data are used whenever an ambiguous interpretation might arise.

Data standards

"Standards are documented agreements containing technical specifications or other precise criteria to be used consistently as rules, guidelines, or definitions of characteristics to ensure that materials, products, processes, and services are fit for their purpose"¹⁸ (ISO 2015).

Data Standards are documented agreements aim to provide consistent meaning to data shared among different information systems, programs, entities of data-consumers/users on representation, format, definition, structuring, tagging, transmission, manipulation, exchange, use, and management of data. Data standards in biodiversity science are being managed by the Biodiversity Informatics Standards organization¹⁹ TDWG.

Metadata

Metadata is “data about other data”, based on standard specific to a particular discipline. Metadata are a description of content and context of content, using predefined attributes, aim at providing a brief data about the characteristics of a resource (e.g. ‘who, what, where, when, how and on what purpose’).

In the GEOSS and GBIF contexts, from the point of view of the data provider, metadata contain information about their resources (datasets), while for the data consumer the metadata are used both to evaluate the resources and services needed to handle the data (e.g. discover, access) and to "assess appropriateness of the resource for particular needs – their so-called ‘fitness for purpose’".²⁰

Within the biodiversity domain the metadata description (file or data) should automatically be assigned to all processed and published data or object. Another requirement is that a tool for data sharing should guarantee a persistent link between the metadata and data/object. This is very important for the integrity of the information, to keep track of the origin of the data and respect IPR statements for example.

Depending on the context or usage, the same piece of information can be considered as metadata or data. The tools for data sharing can have embedded metadata templates, while in other cases the data standard is in part or entirely considered as metadata. Known standards that may fall under that case are for example Ecological Metadata Language (EML²¹), Darwin Core (DwC²²), ISO 19115 (Geographic information – Metadata²³) and Access to Biological Collection Data (ABCD²⁴), to name a few. These and other data standards have

¹⁸ <http://www.iso.org/iso/home/standards.htm>

¹⁹ <http://tdwg.org/>

²⁰ <https://code.google.com/archive/p/gbif-metadata/wikis/Introduction.wiki>

²¹ http://en.wikipedia.org/wiki/Ecological_Metadata_Language

²² http://en.wikipedia.org/wiki/Darwin_Core_Archive

²³ http://www.iso.org/iso/catalogue_detail.htm?csnumber=26020

²⁴ <http://www.tdwg.org/activities/abcd/>

been extensively reported in the EU BON deliverable D2.1 Architectural design, review and guidelines for using standards²⁵.

Data vs. information

Data or “raw data” (also known as “primary data”) is a term for information collected from a source. From the perspective of the infrastructure service provider an important distinction between raw data and information is that data entities are provided, defined and described by an external source, which is outside of the scope of the infrastructure. Raw data is multi-purpose and can be reused. Raw data doesn't yield much information until it is processed (hence interpreted) and possibly integrated with other data. Once processed, the data may support particular types of information.

For example, an occurrence record for a certain species within a dataset is a "data". The interpreted contribution of one or a set of such records with its known attributes and relationships to other data, in term of scientific meaning, is "information".

The LifeWatch²⁶ information models, which aim to conform with the INSPIRE²⁷ Implementation Rules, address the differences between data and information (in accordance with Federal Standard 1037²⁸) in its 'Information View'.

- Data: representation of measurements, facts, concepts, or instructions in a formalized manner that can be processed by humans or by automatic means.
- Information: the meaning that a human assigns to data by means of the known conventions used in their representation.

The LifeWatch Reference Model²⁹ further distinguishes between two aspects of information:

- Primary and derived information (including metadata) related to biodiversity data.
- Meta-information, that is: descriptive information about available information and resources with regard to a particular purpose (i.e. a particular mode of usage). Examples of 'Purposes of data' that are handled by different meta-information models include: Discovery, Orchestration, Collaboration, Identification, Authentication and Authorization, Provenance, Quality evaluation, Indexing, Retrieving, and Integration.

²⁵ <http://www.eubon.eu/documents/1/>

²⁶ <http://www.lifewatch.eu/web/guest/home>

²⁷ <http://inspire.ec.europa.eu/>

²⁸ http://en.wikipedia.org/wiki/Wikipedia:Federal_Standard_1037C_terms

²⁹ http://www.eubon.eu/getatt.php?filename=LW-RMV0.5_4310.pdf

Processed and secondary data and information

Based on the increased availability of biological records, secondary information can be generated by processing and analyzing primary data using cutting-edge techniques for modelling, mapping, statistics, graphing and for visualization of data.

The non-exhaustive example products of secondary information and data products may include Red Lists, endangered species lists, observations that associate spatial coordinates, environmental data with habitat and landscape data, genetic data based on sequences and genes.

The need for definition of data for purpose

The discovery, analysis, and interpretation of data, particularly for the purposes of generating information, often requires an understanding of the semantic context for a particular term, which depends on the particular scientific community and the purpose for which the data was collected. For example, precipitation has a very different meaning in the context of a chemistry dataset than an ecological dataset. And within ecology, the concepts of rain, snow, and sleet are understood to be specific forms of precipitation.

Ontologies are structured way to organize the different meanings that a particular term can have in different contexts as well as to describe the relationships between different concepts. Well-structured ontologies can greatly assist both the discovery and interoperability of datasets, but the proper application of these ontologies requires an understanding of the context of the data, which should be provided by the metadata. One mechanism of providing that information is to explicitly specify that context, by referencing a particular term in a relevant ontology or from a specifically referenced controlled vocabulary of keywords.

Some recent developments regarding vocabularies and ontologies in biodiversity informatics are outlined in deliverable D2.1.

Data publishing

Biodiversity data can be made publicly available through the process of “publishing”. Data publishing makes the data accessible through the use of standard procedures and protocols. It implies the use of common practices and standards ensuring that data can be discovered and reused effectively, and that data owners and custodians get the recognition they deserve. These practices also apply for data sharing, when data are made fully publicly available.

GBIF³⁰ and Pensoft³¹ summarize the incentives to publish biodiversity data as follows:

- Data can be indexed and made discoverable, browsable and searchable through biodiversity infrastructures (e.g., GBIF, Dryad³² and others):

³⁰ <http://www.gbif.org/publishingdata/summary>

³¹ <http://www.pensoft.net/>

- Discoverable and accessible data contribute to global knowledge about biodiversity, and thus to the solutions that will promote its conservation and sustainable use.
- Data publishing enables datasets held all over the world to be integrated, revealing new opportunities for collaboration among data owners and researchers.
- Publishing data enables individuals and institutions to be properly credited for their work to create and curate biodiversity data, by giving visibility to publishing institutions through good metadata authoring.
- Collection managers can trace usage and citations of digitized data published from their institutions and accessed through GBIF and similar infrastructures.
- Data produced and collected using public funds can be published, cited, used and re-used, either as separate datasets or collated with other data. Indeed, some funding agencies now require researchers to make their data freely accessible.

To encourage the publishing of biodiversity data one should stress the importance of the use of the ‘Data papers’ concept (recently promoted for the biodiversity community by Chavan and Penev (2011), Chavan et al. (2013)).

A **data paper** is a searchable metadata document, describing a particular dataset or a group of datasets, published in the form of a peer-reviewed article in a scholarly journal. In contrast to the data sets published in conjunction with academic research papers, data papers may contain raw primary data, independent of a research hypothesis. This makes it uniquely adapted for the publication of biodiversity data from large collections, such as those curated by natural history museums.

Unlike a conventional research article, the primary purpose of a data paper is to describe data and the circumstances of their collection, rather than to report on hypotheses testing and to draw conclusions.

Key characteristics of the data-paper concept (Chavan et al., 2013) are that it:

- provides a citable journal publication that brings scholarly credit to data publishers;
- describes the data through structured, human-readable extended metadata;
- brings the existence of the data to the attention of the scholarly community.

Recent developments include the endorsement of the data paper concept by several EU-funded projects and the creation of the next-generation Biodiversity Data Journal³². Furthermore, Colombia’s Alexander von Humboldt Biological Resources and Research Institute is commissioning a journal dedicated to publishing data papers, and public repositories, such as Dryad and Scratchpads, are collaborating with academic publishers to encourage data-paper publishing (Chavan et al., 2013).

³² <http://www.datadryad.org/>

³³ <http://bdj.pensoft.net/>

Data sharing and open access

Wikipedia defines data sharing as “the practice of making data used for scholarly research available to other investigators”³⁴. It’s considered to be a part of scientific method together with documentation and archiving. A number of institutions, funding and publishing agencies have policies regarding data sharing. While data sharing for some is about validating results, for others, publishing data are about enabling big data solutions and approaches (Anderson, 2014).

The terms “data sharing” and “data publishing” are often used interchangeably. However, there are differences. Data that is shared may still be private and access to it can be controlled. Access to shared data can be revoked. (This was an important clause in the original GBIF Data Sharing Agreement, which placed emphasis in keeping the data owner in control.) However, when something is published, it has been made openly available for good, and access cannot be revoked anymore.

Shared data are useful only if they are searchable and usable. For both characteristics data must be formatted in a standard way, conform to standard structure and semantics and have appropriate metadata attached³⁵.

Despite the ongoing discussion how to share, what to share and on what conditions to share it’s almost impossible to imagine the modern science without data sharing initiatives emerging worldwide and in different disciplines.

Open access is an important principle in data sharing (although data can also be shared in restricted ways). Data sharing necessitates the use of an agreement or a license where the terms and conditions have been stated. When integrating data from thousands of sources, only open access and standardized licenses such as those of Creative Commons will work.

The important players in domains of earth and biodiversity observation, such as GEO BON, GEOSS, including EU BON, pursue strategic goals³⁶, among which data sharing is directly addressed:

- address the need for timely, global and open data sharing across borders and disciplines, within the framework of national policies and international obligations, to maximize the value and benefit of Earth observation investments,
- implement interoperability amongst observational, modelling, data assimilation and prediction systems.

The first 10-Year Implementation Plan of GEO stated that "The societal benefits of Earth observations cannot be achieved without data sharing", and set out the GEOSS Data Sharing Principles:³⁷

³⁴ http://en.wikipedia.org/wiki/Data_sharing

³⁵ <http://www.nature.com/nature/journal/v461/n7261/full/461171a.html>

³⁶ https://www.earthobservations.org/documents/geo_vi/12_GEOSS%20Strategic%20Targets%20Rev1.pdf

³⁷ https://www.earthobservations.org/geoss_dsp.shtml

- There will be **full and open exchange** of data, metadata and products shared within GEOSS, recognizing relevant international instruments and national policies and legislation;
- All shared data, metadata and products will be made available with minimum time delay and at minimum cost;
- All shared data, metadata and products being provided free of charge or no more than cost of reproduction will be encouraged for research and education.

EU BON Data Sharing Agreement

The EU BON project determined in 2013 the need to put in place a detailed Data Sharing Agreement³⁸, which follows the above GEOSS Data Sharing Principles, but also gives additional terms and conditions, which are relevant for the biodiversity community. These conditions include the need to hide potentially sensitive data on endangered species, and the need for an embargo on data release to support priority in scientific publishing, and to motivate data sharing. This agreement has yet to be tested in practical terms.

Other related initiatives include the revision of the GBIF Data Sharing Agreement to ensure that all data sets are associated with a standard, machine-readable Creative Commons equivalent license (i.e. CC-0, CC-BY, CC-BY-NC) that can be automatically processed to support data integration across large number of data sets, and the Bouchout declaration³⁹ that promotes licenses or waivers in support of open biodiversity knowledge management. The EU BON Data Sharing Agreement is in line with the main principles of the Bouchout declaration on open biodiversity knowledge management. Recommendations that are beyond the scope of the agreement are also promoted (e.g. the need for persistent identifiers for data, linking data using agreed vocabularies and sustaining identifiers in the long term) (Wetzel et al., 2015).

Moreover, EU BON adheres to the principles of free and open exchange of data and knowledge, in accordance with the “Joint Declaration on Open Science for the 21st Century”, presented by the European Federation of Academies of Sciences and Humanities and the European Commission on 11th April, 2012⁴⁰.

³⁸ http://www.eubon.eu/news/10954_EU%20BON%20Data%20Sharing%20Agreement

³⁹ <http://www.bouchoutdeclaration.org/declaration/>

⁴⁰

<http://www.allea.org/Content/ALLEA/General%20Assemblies/General%20Assembly%202012/Joint%20Declaration%20GA%20Rome%202012%20signed%20v2.pdf>

3. Inventory of tools

Types of tools

There is a number of specific tools for biodiversity data sharing, such as GBIF's Integrated Publishing Toolkit (IPT). However, there are also general purpose tools, such as MS-Access, that are widely used in data management. These general tools are often used to share tables of semi-structured data. Most of these tools are well known by the community. They are generally easy to use and do not require a steep learning curve or the assistance of an IT specialist. From a short term perspective these tools provide a quick win for data exchanges.

Neither spreadsheets, such as MS-Excel and comma/tab delimited files, should be ruled out as efficient means to share data and information. They are routinely used to transfer data among collaborators and to feed higher level data management systems or applications. While such systems are popular, using such tools, particularly without applying clear standards to the data does not promote larger scale data management nor interoperability between datasets. The use of proprietary systems forces data into particular formats and can become an additional barrier to data sharing, reuse and accessibility.

In order to overcome such barriers, the community has developed data sharing tools that assert common standards and structures on users. Some tools are more generic and data schema independent and thus can be used in multiple domains, while many other tools are targeted designed for selected data types, models, specific applications and purposes.

One can cite here tools to exchange geographic information such as background maps, sampling localities and coordinates. These tools are of general purpose and are not necessarily designed for biodiversity and habitat related data. However, they're still useful for the domain.

There are groups of tools that have been specifically designed for biodiversity data, environmental data and ecological data. They are often developed in the context of a project or of an application. Most are very useful but sometimes need adaptations or connector applications to become interoperable at larger scales. Typically these tools include data export functions, which allow deriving data into standardized formats.

Data publishing tools can process raw data into reports or publications to be further shared as information for educational, decision-making, policy-making purposes, which offer additional form of information sharing.

The distinction of tools for sharing and publishing is becoming less important. Technically they implement the same interoperability mechanisms. The distinction may lie in the ability of a tool to offer functions for embargo, hiding sensitive details, and access control. In general, data sharing tools aim at facilitate curating live data, while publishing tools are suited for making a frozen version of data permanently discoverable, and accessible.

We can also broadly group the tools into **distributed** and **centralized** categories. The distributed ones are being used and managed by the data custodians themselves. The

centralized ones are portals or shared repositories not managed by the data custodians, but by an aggregator or publisher.

Tools can also be categorized as **specialized** or **general** purpose. Specialized tools have built-in support for biodiversity data types and data standards, whereas general purpose tools, e.g. GIS tools and spreadsheets, can deal with more generic data.

The limitation and a possible problem of the approach outlined here is in the word “specialized”. There simply are not distributable data sharing tools specialized for each biodiversity data type (genomic, occurrence, species, habitat, ecosystem, ...), but rather only for occurrence and species level data. The question is whether specialized tools are needed at all for each data type.

Tools surveyed by EU BON

This report mainly focuses on data publishing and data sharing tools. As stated in the introduction, there are also other tools like storage tools, data management tools, data capture or portals/interfaces of some applications which the users can also consider as part of the data sharing process. These tools are out of the scope of this report.

About 30 existing data sharing tools, commonly used in the natural history domain, have been evaluated by the EU BON community and results of their assessment are presented in the following format:

- *Main usage, purpose, selected examples*
- *Pros and cons of the tool*
- *Recommendations*
- *Tool status*

A summarized overview of these tools is given in the **Annex 1** and is available for consultation online⁴¹. The list is not meant to be exhaustive, but rather as a snapshot of the current state of art and as knowledge of the community relevant to data sharing tools. The EU BON online repository is being regularly updated with additional tools, newly discovered and analyzed tools or newly developed tools. Therefore, this review can be used as a gap analysis on tools that are required. For instance, there seems to be an absence of tools for sharing habitat data.

This analysis did not start from scratch, but was based on a previous analysis of tools made in the framework of the projects EDIT (European Distributed Institute of Taxonomy)⁴² and SYNTHESYS (Synthesis of Systematic Resources)⁴³.

⁴¹ <http://eubon.cybertaxonomy.africamuseum.be/data-sharing-tools-repository>

⁴² <http://www.e-taxonomy.eu/>

⁴³ <http://www.synthesys.info/>

4. Requirements

User communities to be supported

To answer the needs of biodiversity observation network, the data sharing tools should be applicable to different types of data, for example tools specialized in species occurrence, that should, in turn, be combined or made interoperable with tools specialized on habitat data. To this end, aspects such as genetic and functional trait data should not be overlooked. In this regard the tools used by EEA (European Environment Agency)⁴⁴ and LTER (The Long Term Ecological Research Network)⁴⁵ are particularly useful. For species occurrence data the data sharing tools of GBIF⁴⁶ adhering to the TDWG standards are widely used and very relevant.

EU BON, as stated in its DoW and Data Sharing Agreement, has close ties to GEOSS (The Global Earth Observation System of Systems)⁴⁷. The data sharing tools to use should, to a large extent, be compatible with the GEOSS community tools, and support observational, quantitative data. The biodiversity community has some special requirements for data sharing, which have been noted in the EU BON Data Sharing Agreement. This applies for example to sensitive data that include localities of certain endangered species. Attention has to be drawn here to the fact that there are some requests on embargo periods before the data becomes publicly available. Care should be taken so that the tools used provide mechanisms to handle these special requirements.

In relation with the overall global GEO BON initiative, tools that are able to handle the Aichi Targets⁴⁸ and the Essential Biodiversity Variables (EBV)⁴⁹ are needed to make the EU BON / GEO BON platform for data sharing effective.

Different tools or sufficiently flexible tools will be needed to accommodate the different type of users and their anticipated needs in terms of access to data and information for further processing or decision making. These end users are for example test site managers, scientists engaged in monitoring programs, modelers, decision and policy makers, as well as interested citizens.

Another general requirement is that the metadata description (file or data) should automatically be assigned to all processed and published data or object. Thus a tool for data sharing should guarantee persistent link between the metadata and data/object. This is very important for the integrity of the information, to keep track of the origin of the data and respect IPR statements, for example.

⁴⁴ <http://www.eea.europa.eu/>

⁴⁵ <http://www.lternet.edu/>

⁴⁶ <http://www.gbif.org/>

⁴⁷ <https://www.earthobservations.org/geoss.shtml>

⁴⁸ <http://www.cbd.int/sp/targets/>

⁴⁹ https://www.earthobservations.org/geobon_ebv.shtml

Selected tools

Having thoroughly analyzed a range of modern biodiversity data landscape (D2.1) and the identified data gaps (D1.2), followed by a comprehensive review of existing data sharing tools, it was concluded that the choice of tools selected for implementation by the EU BON project should be determined by the needs of the end users involved in biodiversity observation. This is a large community – from amateur scientists (citizen scientists) exploring and recording life around them *via* their mobile devices, to researchers, and to decision makers looking for processable and reliable data to build their reports and forecasts upon.

During the assessment phase, the number of tools that were identified for the purpose of data handling and testing accumulated to the amount that could barely be handled or supported by the EU BON project alone. Instead, the EU BON consortium has identified the tools that will be adapted, supported and distributed targeting the different groups of data providers. The preference was given to distributed, controllable, and specialized tools as it's explained above in the chapter 3. This limits the choice of tools presented in the **Table 1**.

With this scope, the status of the selected tools was analyzed and recommendations were offered regarding e.g. required enhancements to support EU BON prioritized use cases (see D2.1), the GEOSS Data Sharing Principles, and the EU BON Data Sharing Agreement.

Table 1. List of selected tools.

Purpose	Data type	User group	Tool name	Description of tool	Operating Systems	Standard supported	Requirements for implementation	Testing results	Link to the source, tutorials, manuals
Data sharing, distributed	Occurrence data (collections, taxonomy), Monitoring data (including sample-based data)	Scientists, Monitoring sites	GBIF Integrated Publishing Toolkit (IPT)	<p>Tool to publish and share biodiversity data sets and metadata through the GBIF network. Allows publication of three types of biodiversity data:</p> <ul style="list-style-type: none"> primary occurrence data (specimens, observations), species checklists and taxonomies, sample-based data from monitoring programs 	Windows, MacOS, Linux	DwC, DwC-A, EML	<p>Enhancement with the Event core to handle sample-based data.</p> <p>Darwin Core standard enriched with quantitative measurements.</p>	<p>Tested by different partners.</p> <p>Several datasets from test sites are published: http://www.gbif.org/dataset/search?q=&type=SAMPLING_EVENT</p> <p>There is an ongoing discussion at GBIF community site on sample-based publishing.</p>	<p>Download: http://www.gbif.org/ipt</p> <p>User manual: https://github.com/gbif/ipt/wiki/IPT2ManualNotes.wiki</p> <p>Community site: http://community.gbif.org/pg/groups/47949</p>
Data sharing, centralized	Metadata (Monitoring, environmental science, ecology)	Monitoring sites	DEIMS (Drupal Ecological Information Management System)	Drupal open-source, collaborative platform, that provides a web interface for scientists and researchers' networks, projects and initiatives with a metadata management and data sharing system.	Windows, Linux	EML, ISO		Tested by CSIC. Datasets from Doñana LTER site are published.	<p>Repository: https://data.lter-europe.net/deims/</p> <p>EML handbook: https://data.lter-europe.net/deims/sites/data.lter-europe.net/deims/files/emlbestpractices-2.0-FINAL-20110801_0.pdf</p>
Data sharing and exchange, distributed	Data	Scientists, Monitoring sites, Citizen scientists	Spreadsheet processors (e.g. Excel, GBIF spreadsheet processor, DataUp, Dash)		Windows, MacOS		Explore ways to generate and deposit a metadata file (in EML) by DataUP and made data available for discovery and use (by GBIF) for the public.	DataUp is tested by Doñana site.	GBIF spreadsheet processor: http://tools.gbif.org/spreadsheet-processor/

Data publishing (Scholarly publishing), centralized	Data and metadata	Scientists, Monitoring sites	PWT or ARPHA Publishing Platform	Narrative (text) and data integrated publishing workflow, launched to mobilize, review, publish, store, disseminate, make interoperable, collate and re-use data through the act of scholarly publishing.	x	DwC, DwC-A, EML	<p>A new plugin developed which makes it possible to convert metadata into a manuscript for scholarly publications, with a click of a button.</p> <p>A possibility to easily import occurrence records into a taxonomic manuscript in ARPHA.</p> <p>An automatic export and integration of PlutoF data into Pensoft's ARPHA platform via API.</p>	The AWT is fully operational and currently used by three Pensoft journals – Biodiversity Data Journal, Research Ideas and Outcomes and One Ecosystem .	<p>AWT: http://arpha.pensoft.net/</p> <p>BDJ: http://bdj.pensoft.net/</p> <p>RIO: http://rio.pensoft.net</p> <p>One Ecosystem: http://oneecosystem.pensoft.net</p> <p>A tutorial for the use of ARPHA called “Trips and tricks” is available on the website at: http://arpha.pensoft.net</p>
Data mining	Historical data, data from publications	Scientists	GoldenGATE Imagine or TreatmentBank and DwC	A platform to store, annotate, access and distribute taxonomic treatments and the data objects within. It offers with GoldenGate ^[1] and respective XML schemas (TaxonX ^[2] , TaxPub ^[3]) tools to convert unstructured text into semantically enhanced documents with an emphasis on taxonomic data like treatments, scientific names, materials observation, traits or bibliographic references.	x	DwC, DwC-A	<p>Taxpub as domain specific extension of the Journal Article Tag Suite has been developed to model the semantic content of the biodiversity literature; RDF and a treatment ontology is under development. (https://github.com/plazi/TreatmentOntologies)</p>	<p>DwC-A are routinely used to transfer data from Plazi to GBIF since 2014;</p> <p>TaxPub is used to import publications from Pensoft of Plazi;</p> <p>GoldenGate conversion is operational and successfully used for conversions (Miller et al., 2015).</p>	<p>API: http://plazi.org/wiki/Treatment_Data_Access</p> <p>GoldenGate Imagine software and manual: http://plazi.org/wiki/GoldenGATE_Editor</p>

Data sharing, distributed	Metadata , ecological data	Scientists, Monitoring sites	Morpho Metadata Editor (KNB) and Metacat	Application designed to facilitate the creation of metadata so that scientist can easily locate and determine the nature of a wide range of data sets. It interfaces with the Knowledge Network for Biocomplexity (KNB) Metacat server.	Linux, PostGreS QL	EML	Explore using Morpho (editor) and Metacat (servers) for managing ecological metadata to access and expose LTER sites /datasets. Design feasibility test to clarify and document the requirements for implementation.	Tested by CSIC and INPA.	https://knb.ecoinformatics.org/#tools/morpho Morpho user guide: https://knb.ecoinformatics.org/software/dist/MorphoUserGuide.pdf https://knb.ecoinformatics.org/knb/docs/ Metacat Administrator's Guide: (http://knb.ecoinformatics.org/software/dist/MetacatAdministratorGuide.pdf)
Data sharing, centralized	Occurrence data (collections, observation, molecular), monitoring data, metadata	Scientists, Monitoring sites, Citizen scientists	PlutoF Platform, PlutoF-API, Mobile apps	Online service to create, manage, share, analyse and mobilise biodiversity data. Data types cover ecology, taxonomy, metagenomics, nature conservation, natural history collections, etc.	x Android	EML	Implementing use of high-end devices to mobilize data from the public, while focusing on quality of data.	Tested by UTARTU, INPA and in Israel.	PlutoF: http://plutof.ut.ee App: On Google Play

Enhancements required

In order to support the requirements of biodiversity observation, functional enhancements were made to several tools as indicated in **Table 1**. The GBIF IPT was enhanced by support to quantitative monitoring data. The ARPHA Publishing Tool was developed into a new version from the predecessor PWT. For the TreatmentBank tool (former GoldenGate Imagine tool), a domain specific extension has been developed to model the semantic content of the biodiversity literature. For the PlutoF platform extended support to mobile devices has been developed. These and other enhancements have been described below.

Testing

The selected and enhanced tools were installed and extensively tested by EU BON partners and particularly by the test sites. They were evaluated during several training workshops and disseminated and discussed in the biodiversity informatics community *via* community sites and mailing lists.

Testing by test sites:

Several test sites were established by EU BON, each representing different geographical regions and ecosystems. Besides other functions they should play an important role in testing and validation of EU BON concepts, tools and services.

Documenting data sets is an essential part of data integration. By describing the contents and context of data files, metadata ensure the discoverability of data sets and allow early filtering options before data analysis. Based on the WP5 document (MS513) describing the kind of data the test sites are producing, the evaluation of the available tools for documenting data sets highlighted in MS231 “Specification of data sharing tools” were done. The tools (at least three promising alternatives) were deployed in the test sites own servers and were extensively tested with their own data. The feedback was reported back to the consortium (MS517).

The test sites are producing huge amount of information on the functioning of ecosystems on daily basis. These sites collect data of many different kinds (biotic, abiotic), different formats (haplotype frequencies, species observations, environmental parameters, media files), at various scales, and metadata (procedures, protocols, description of methods and campaigns). Therefore, the documentation of this information, its standardization and integration to the global observation network is an essential step in the workflow of any site, field station or nature reserve.

An assessment of the data sharing tools included:

- checks against a pre-defined list of essential data elements sets fits with actual data coming from the test sites;
- suitability of different data elements to the pre-defined EML/Darwin Core tags;

- check whether the tools have sufficient resolution to account for the data requirements previously made by EU BON partners;
- ease of installation, in case the tool will be available on the portal - does it require a registration, what are the conditions on data share and data use;
- upload process: what file formats are supported, what is the speed of data upload, can the database be connected;
- editing the information;
- how complete are the metadata and other additional information from the test site, can it be provided in standard format properly to document the data sets;
- limitations, e.g. tool does not allow automatic adding of taxonomic lists;
- defining the strengths and weaknesses of the tools to help new users to choose the one that best fit into their data set documentation process;
- user friendly;
- support: technical and scientific assistance;

Also, useful tools provided by different biodiversity information facilities that may help the process of data integration and analysis within the EU BON consortium were analyzed. It considers both mobile applications and online tools that are currently widespread within the biodiversity monitoring community. They are found under the umbrella of research projects, monitoring teams, NGOs and citizen scientists associations. They are mainly based on the number of technological advances implemented in smart devices such as PDAs, mobile phones and tablets that allow including information related to the observations we make in an automatic way (e.g. GPS position, geo-referenced pictures, date and time...). Similarly, a growing number of biodiversity information platforms offers the users a web portal (sometimes + mobile app) where they can share their observations.

For a proper evaluation, the following aspects were analyzed:

- easy to use,
- quick in recording,
- allow the user to keep track of his/her activity (list observations) and ideally group them to get a quick overview of the activity (individuals per species, per area, etc.).

Conclusions and lessons learned are summarized in the document MS517 and taken into account while implementing the tools.

Training:

To support biodiversity data mobilization and integration, EU BON pays attention to capacity building of biodiversity communities (e.g. researcher, citizen scientists, NGO's) that are

involved in collecting and providing biodiversity information, including monitoring initiatives (Wetzel et al., 2015)⁵⁰. To overcome existing limitations and improve data digestibility, EU BON has developed a training framework that includes supporting data mobilization and interoperability at the user and institutional level. A comprehensive training program was implemented with a focus on data and metadata integration strategies, use of standards and data sharing tools for institutional data and IT managers, researchers, citizen scientists and monitoring programs. Several technical (informatics) workshops have been held on data standards and prototypes, e.g. of data sharing tools and the biodiversity portal. In addition, interdisciplinary ‘task forces’ such as those on EBVs and remote sensing have been set up to foster capacity building.

In the framework of training preparation MRAC has tested EU BON IPT using different datasets. Therefore, the EU BON test sites were contacted and asked to provide their typical sampling protocols and monitoring data to be extensively tested. Preliminary results of this exercise were presented by MRAC and GBIF during the EU BON General meeting in Cambridge (1-4 June 2015), also resulting in fruitful discussions on how to improve the tool.

Also PlutoF and related citizen science applications were subject of the trainings and hands-on sessions. The training outputs and user feedback were considered during further tool/platform development and improvement.

In collaboration with the European Mediterranean Observation Network (EMODnet), GoldenGATE has been taught and evaluated (8-9 June 2015). User feedbacks have been integrated to improve the tools.

Community feedback:

At the preparation stage for the trainings and as a post-training discussion platform the GBIF Community Site⁵¹ was used. This is an open social networking platform targeted at GBIF stakeholders and the biodiversity informatics community at large. To discuss and promote the new IPT functionalities the sample-based publishing interest group⁵² was created aiming to gather people interested in the subject of publishing biodiversity data coming from biological sampling efforts. Groups like EU BON and GBIF have been pushing for a change in biodiversity standards and tools to enable a more faithful representation of these data online. The group also aims at facilitating the uptake of tools modifications by the community by means of discussions, trainings, online supporting material etc. The questions raised during the discussion are carefully investigated by tool developers and if possible the changes are implemented.

Partners of the EU BON are also subscribed to the public IPT mailing lists where users can share their experience with the tool, indicate the bugs, ask for help, and exchange ideas.

Such community feedback is an important source of information which helps both developers and users to solve many of the problematic issues, improve the product, transfer the

⁵⁰ [10.1080/14888386.2015.1075902](https://doi.org/10.1080/14888386.2015.1075902)

⁵¹ <http://community.gbif.org/>

⁵² <http://goo.gl/VCulg9>

knowledge and connect people all around the world, helping the scientific communication and data mobilization.

5. Implementation

This chapter provides a detailed description of the EU BON supported tools: specifications, adaptations made and results of the testing followed by the recommendations on the implementation of the tool in a real environment.

This description will be used to produce detailed workflows which should form an important part of the EU BON Helpdesk⁵³ aiming to support the users (data provider) by assisting them on the data mobilization road (from choosing the standard, monitoring scheme or data sharing tool to visualization and interpretation of published data).

GBIF Integrated Publishing Toolkit (IPT)

Tool description:

The GBIF IPT (Integrated Publishing Toolkit)⁵⁴ is open source software widely used to publish and share biodiversity datasets on the GBIF network and related networks. It uses the standards Darwin Core (DwC) and Ecological Metadata Language (EML). Currently the IPT support three core types of data: checklists, occurrences, and events, plus data set level metadata. It is a community-driven tool and the new enhancements sponsored by the EU BON project were widely discussed and assessed by the users⁵⁵. It has multilingual user interface and a very extensive supporting documentation⁵⁶. The IPT provides a service to convert data set metadata into a draft data paper manuscript for submission to a peer-review journal (see chapter on PWT)⁵⁷. Detailed information can be found at GBIF site⁵⁸.

Enhancements by EU BON:

The latest release from September 10th 2015 is the version 2.3. This version has been developed together with the EU BON, in the form of the first prototype to test the handling of sample-based data with several uses cases from the EU BON monitoring test sites. Sample-based data are a type of data available from thousands of environmental, ecological, and natural resource investigations. These can be one-off studies or continuous monitoring programs. Such data are usually quantitative, collected after carefully designed sampling, calibrated, and follow certain protocols so that changes and trends of populations can be detected (Ó Tuama, 2015).

⁵³ <http://eubon.cybertaxonomy.africamuseum.be/node/2812#overlay-context=>

⁵⁴ <http://www.gbif.org/ipt>

⁵⁵ <http://lists.gbif.org/mailman/listinfo/ipt>

⁵⁶ <https://code.google.com/p/gbif-providertoolkit/wiki/IPT2ManualNotes?tm=6>

⁵⁷ <http://www.gbif.org/publishingdata/datapapers>

⁵⁸ <http://www.gbif.org/ipt>

In version 2.3, a new core object, the sampling **Event** is introduced. The Event is defined as “an action that occurs at a certain location during a certain time “. Using a star schema (one-to-many relational model, where a row in a (central) *core* table can be linked to many rows in one or more (surrounding) *extension* tables, **Fig. 2**) should facilitate encoding sample-based data, and provide additional data types (biotic and abiotic) *via* associated extensions.

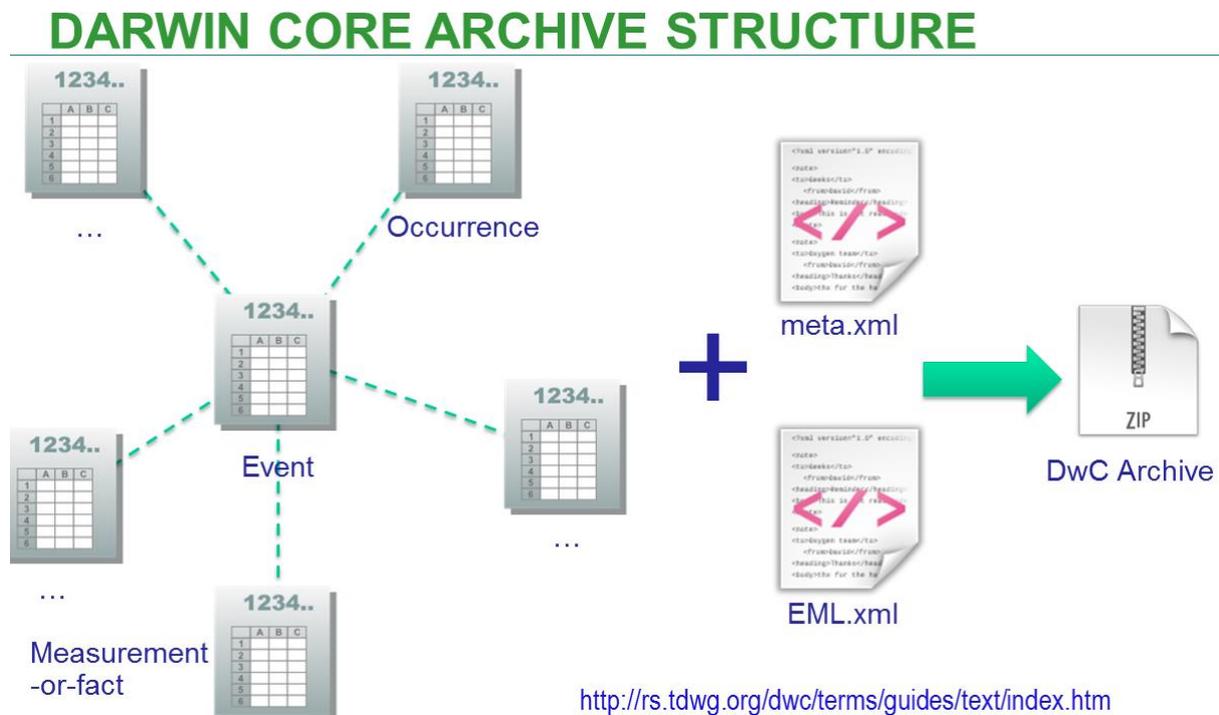


Figure 2. Darwin Core Archive star schema, with the Event Core configuration.

In the core table, each row is a sample identified by a unique eventID and other columns holding sampling protocol, sample size, date, location, etc. The rows in the “Occurrence” extension table refer to a sampling event in the core (via eventID) and list the taxa in the sample together with associated quantity measurement. It also allows the use of a “Measurement-or-facts” extension for the efficient expression of environmental information associated with the event.

The Darwin Core vocabulary already provides a rich set of terms, organized into several classes (e.g., Occurrence, Event, Location, Taxon, Identification). Many of these terms are relevant to describe sample-based data. Synthesizing several sources of input, a small set of terms relating to sample data were identified as essential, some of which are already present in the DwC vocabulary. Five new terms were ratified by TDWG (Biodiversity Information Standards) on 19 March 2015.

Darwin Core terms for sample-based data (*Indicates new terms):

- eventID
- parentEventID*

- `samplingProtocol`
- `sampleSize*`
- `sampleSizeUnit*`
- `organismQuantity*`
- `organismQuantityType*`

Detailed information on how to configure core types and extensions can be found on the wiki for IPT manual⁵⁹.

Testing and implementation:

Testing of the new IPT functionalities attempted by several EU BON partners, test sites themselves (including the comparison with other common data sharing tools), MRAC (in the training preparation stage using the data from the test sites) and GBIF (who has assisted test sites to actually publish several datasets).

Currently, all test sites manage to share information and data sets regarding biodiversity using different systems and platforms: Rhine-Main observatory (RMO), Sierra Nevada and Doñana belong to the LTER network, where biodiversity information coming from these sites is being uploaded⁶⁰; information coming from Amvrakikos National Park, as well as other data sets regarding marine biodiversity are being shared by HCMR via the MedOBIS regional node⁶¹. Additional information is being shared by other regional and national networks as well as own-developed/deployed systems such as Metacat (Sierra Nevada and INPA) or Institutional web portals.

For the IPT testing purposes some of the sites (Doñana, RMO) have used EU BON IPT prototype, Amvrakikos and INPA have used their own IPT instances.

Comparison and evaluation of tools (DEIMS, IPT, DataOne) done by WP5 is discussed in MS517. Regarding the IPT it is emphasized that it allows a very comprehensive documentation of data sets, including monitoring protocols, taxonomic coverage and many other details. Depending on whether the user is more or less reluctant to learn new tools and procedures (DEIMS is easier), and depending on the length of its taxonomic coverage (quicker in DEIMS), and whether taxonomic authority and checking are required (only available in the GBIF IPT) one or the other are advised. It should be noted here that access to a centralized DEIMS instance is not public, and a proper access should be obtained before starting sharing information. On the contrary, logging into a local instance of IPT is straightforward.

⁵⁹ <https://github.com/gbif/ipt/wiki/IPT2ManualNotes.wiki#configure-core-types-and-extensions>

⁶⁰ <http://data.lter-europe.net/deims/>

⁶¹ <http://lifewww-00.her.hcmr.gr:8080/medobis/>

One of the outputs of the testing phase was a list⁶² of fields for describing data sets and proposed correspondence to relevant EML tags. This list can help data providers to fill in the metadata properly and fully.

The testing carried out by using data sets coming from daily test site activities, taking into account data coming from different domains. Doñana biological station has for instance performed testing with survey data (Coastal birds), Israel Butterfly Monitoring scheme with data coming from their citizen science program, RMO has published terrestrial data (Macrophytes and fresh water invertebrates), Amvrakikos was responsible for marine data (benthos communities in lagoon environment), and Sierra Nevada tried out to publish vegetation data from the forest monitoring.

Initially used as test data to prepare the trainings and to test the EU BON IPT prototype, most of these data sets have been successfully published through GBIF⁶³ and thus enriched the biodiversity information landscape with monitoring data giving a new prospects to data use.

Future developments:

GBIF has defined next action points to enhance the latest developments, especially the introduction of the Event core (cited from the presentation of Donald Hobern at the GBIF Nodes Madagascar meeting, 2015):

- Monitor and report use of extension in network
- Develop visualizations to show temporal and geographic distribution of sample-based data
- Work with existing data publishers to expose extra elements from relevant datasets
- Develop filters to access data for sampling events
- Feasibility studies for further visualizations

Also tags as keywords for EBV classes are under consideration. There has been also discussion at TDWG how to develop the Darwin Core Archive (DwC-A) format further. Ontologies (such as BCO, OBOE) have been brought up as an alternative. Our working hypothesis is that in the long run ontologies may be the solution, but for concrete data exchange needs in near term sticking with the DwC-A format is the practical and affordable solution.

Tool status:

An EU BON instance of IPT is already in place at <http://eubon-ipt.gbif.org> together with a few test sample data sets expressed using an early iteration of the sample data model. The

⁶² <http://www.ebd.csic.es/documents/512028/0/Tabla+Consenso+Metadatos.pdf/911c0572-4418-41b8-8f0c-f0aff5b4060c>

⁶³ http://www.gbif.org/dataset/search?q=&type=SAMPLING_EVENT

latter is undergoing revision based on feedback from the EU BON partners. This instance serves as the EU BON IPT Data Repository, linked to the EU BON Portal prototype.

Version 2.3 of the IPT is available for download in both compiled⁶⁴ and source code⁶⁵ versions.

DEIMS: Drupal Ecological Information Management System

Tool description:

DEIMS, Drupal Ecological Information Management System, is a Drupal based tool to upload and share datasets providing their metadata. Basically, DEIMS is a Drupal installation profile (a set of modules and customizations) for storing, editing and sharing data about biological and ecological research, providing as well forms to describe metadata according to the EML model. DEIMS will help the user to fill in the metadata and provide external links to the data. Each provider is responsible for maintaining the data updated and publicly accessible, depending on the sharing agreements.

Developed in partnership between the US Long Term Ecological Research Network, the University of New Mexico, the University of Puerto Rico, the University of Wisconsin, and Palantir.net, DEIMS main objective is providing a unified framework for ecological information management for LTER sites, biological stations and similar research groups.

DEIMS is not strictly a data or metadata sharing tool, as far as it is not straightforwardly deployable in each provider's infrastructure. Rather than considering it as a tool, we can describe it as an ecological CMS, which needs a Drupal 7 instance deployed and configured properly before starting to install and configure DEIMS modules. This is indeed the main disadvantage in comparison to other metadata sharing tools: it is not easy to deploy and set up needing Drupal experts to configure the host Drupal 7 site according to the data provider requirements.

Testing and implementation:

In the particular case of LTER Europe, they host a Drupal 6 website with DEIMS installed⁶⁶, as well as documentation, guidelines and training resources, as main dataset repository. LTER-EU datasets are public, but the forms to create and share their metadata are only accessible to LTER sites. Some of the EU BON test sites are currently sharing datasets using LTER-EU DEIMS, which are being harvested by GI-cat using the DEIMS EML harvest list⁶⁷. A further upgrade to DEIMS + Drupal 7 is scheduled to start during March 2016, and the stability of both the entire DEIMS site and the harvest list are still not guaranteed, as far as the last versions of the modules are not strictly consistent with the previous DEIMS + Drupal 6 versions.

⁶⁴ <http://www.gbif.org/ipt/releases>

⁶⁵ <https://code.google.com/p/gbif-providertoolkit/source/checkout>

⁶⁶ <https://data.lter-europe.net/deims/>

⁶⁷ <https://data.lter-europe.net/deims/eml/harvest-list-all.xml>

Future developments:

As an alternative, but not accessible for the moment, DEIMS metadata could be translated into ISO-19139 metadata files and shared using a GeoNetwork repository, which could also generate CSW endpoints, consumable by GI-cat. Further tasks will be performed by LTER in reference to this alternative, in order to provide publicly accessible site for GeoNetwork, translation stylesheets and the service endpoints.

After the joint workshop in Granada, both EU BON and LTER agreed to collaborate and share metadata among EU BON and LTER tools and sites. EU BON will provide feedback about the integration of DEIMS in the EU BON registry, taking into account that biodiversity-related metadata must not be degraded during the translation processes, and in fact may need to be expanded with more detailed taxa information. LTER will provide EU BON with feasible alternatives to extract metadata from DEIMS and related tools.

Tool status:

The platform is available at <https://data.lter-europe.net/deims/>. The datasets are public, but the possibility to create the forms and share the metadata is only open for LTER sites.

Spreadsheet tools

Fairly often scientists without technical expertise use spreadsheets as a database alternative. Tabular data provides a great deal of flexibility in how data can be structured. However, this flexibility also makes it easy to structure the data in a way that is difficult to reuse (White et al., 2013).

Microsoft Excel, DataUp, Dash, and open source tools such as Libre Office or Open Refine are software packages that enable the creation of spreadsheets or forms, provide simple data comparison and analysis tools, and create graphs.

Proprietary formats such as those used by Microsoft Excel (e.g., .xls, .xlsx) can be difficult to load onto other software or platform. In addition, these types of files can become obsolete, because of for instance more recent versions of the software that no longer support the original format (White et al., 2013). They lack reproducibility, version control and are in general not suitable for big data processing. These issues can be partly solved if data are stored in a format that can be opened by any type of software, i.e. text files.

Open Refine⁶⁸ could be recommended as a powerful desktop application for data cleanup and transformation to other formats. It has extended documentation and online supporting tutorials⁶⁹ and videos.

Data tables are ubiquitous in daily work of monitoring sites. This why the EU BON test sites were keen to test some of the tools and check whether these tools are able of properly map fields or terms required by test sites to document their data sets (see MS517). So, the DataUp

⁶⁸ <http://openrefine.org/>

⁶⁹ http://enipedia.tudelft.nl/wiki/OpenRefine_Tutorial

tool was tested with data sets coming from monitoring studies run in the Doñana National Park.

DataUp is the tool developed by DataOne to help environmental scientists to upload files to a repository for data management. It also includes a metadata editor. The tool allows to share data sets and document them in a very simple way. It is very friendly and allows the user to login by using Google, Facebook and Microsoft accounts. Afterwards, it gives the user the possibility of entering additional personal and professional information. Files of apparently any format can be uploaded either by drag and drop them into the web browser or using the file explorer. Documentation is very simple, including the name and e-mail of the provider, the file date, title, keywords, abstract, project title and data range description. An additional tab allows the user to load metadata from file, mapping the table name, table description, field name, field description, data type, and units. This is probably because it merely constitutes a hosting service where information is accessible. DataUp is friendly and easy-to-use application, however, the documentation is very basic, and it does not allow the sampling protocol associated to data gathering to be also documented.

Recently, the Data Up is merged with new data sharing platform Dash⁷⁰ from University of California to give support to the California Digital Library.

Recognizing that spreadsheets are common data capture/management tools for biologists and that the Darwin Core terms lend themselves to representation in the tabular format of spreadsheets, three organizations, GBIF, EOL, and The Data Conservancy (DataONE project), collaborated to develop the GBIF Darwin Core Archive Spreadsheet Processor⁷¹, usually just called "the Spreadsheet Processor."

The Spreadsheet Processor is a web application that supports publication of biodiversity data to the GBIF network using pre-configured Microsoft Excel spreadsheet templates. Two main data types are supported: i) occurrence data as represented in natural history collections or species observational data and ii) simple species checklists.

The tool provides a simplified publishing solution, particularly in areas where web-based publication is hampered by low-bandwidth, irregular uptime, and inconsistent access. It enables the user to convert local files to a well-known international standard using an asynchronous web-based process. The user selects the appropriate spreadsheet template (metadata (**Fig. 3**), species occurrence or checklist), completes it and then emails it to the processing application which returns the submitted data as a validated Darwin Core Archive, including EML metadata, ready for publishing to the GBIF or other network (**Fig.4**).

⁷⁰ <http://datapub.cdlib.org/2014/09/12/dataup-is-merging-with-dash/>

⁷¹ <http://tools.gbif.org/spreadsheet-processor/>

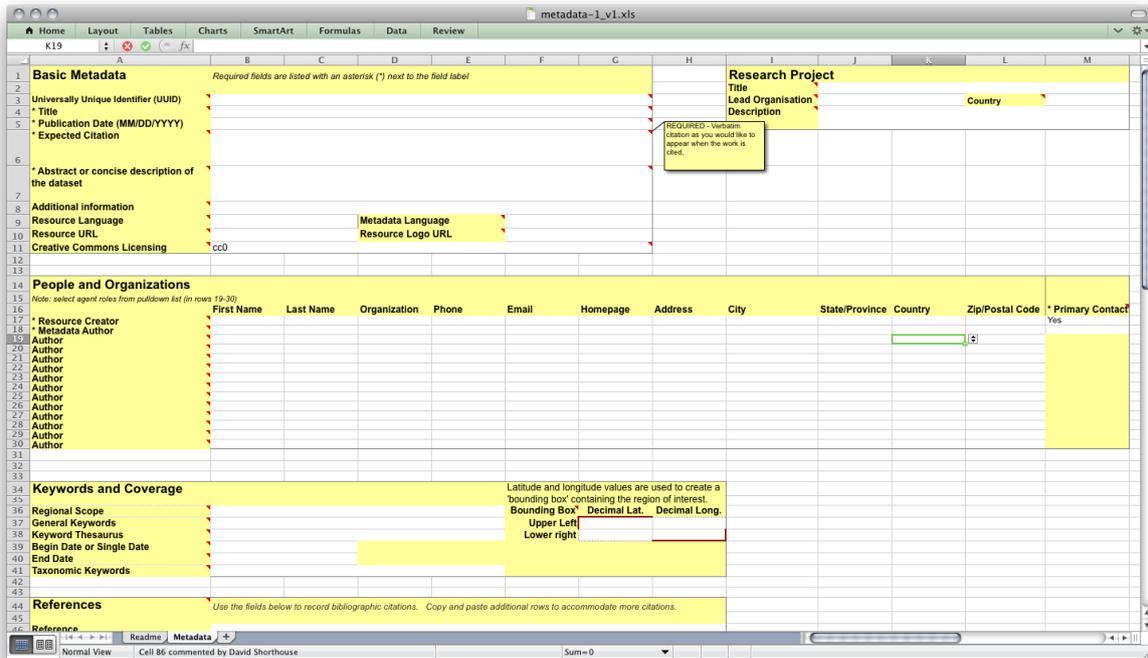


Figure 3. Example of Metadata template.

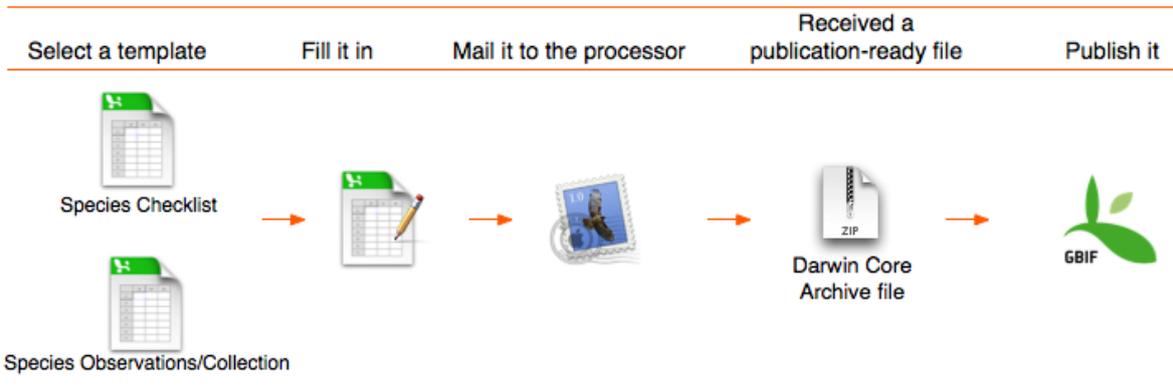


Figure 4. The web based processor ingests a spreadsheet and outputs of a validated Darwin Core Archive.

Future developments:

To extend number of templates for other data types (e.g. sample-based data) and adapt it to the new DwC terms.

Biodiversity Data Journal and ARPHA publishing platform

Tool description:

The Biodiversity Data Journal (BDJ)⁷² and associated ARPHA publishing platform⁷³ represent together a next-generation, narrative (text) and data integrated publishing workflow, launched to mobilize, review, publish, store, disseminate, make interoperable, collate and re-use data through the process of scholarly publishing. All these processes are realized for the first time within a single, authoring, peer-review and publishing, online collaborative platform.

The Biodiversity Data Journal is a novel, community peer-reviewed, open-access journal, launched to accelerate mobilization, dissemination and sharing of biodiversity-related data of any kind. All structural elements of the articles, that is text, descriptions, species occurrences, data tables, etc., are treated, stored and downloaded as DATA in both human and machine-readable formats. The journal will publish papers on any taxon of any geological age from any part of the world with no lower or upper limit to manuscript size, for example:

- new taxa and nomenclatural acts
- data papers describing biodiversity-related databases;
- local or regional checklists and inventories;
- ecological and biological observations of species and communities;
- identification keys, from conventional dichotomous to multi-access interactive online keys;
- descriptions of biodiversity-related software tools.

ARPHA⁷⁴ stands for Authoring, Reviewing, Publishing, Hosting and Archiving, all in one place. It is an innovative publishing solution developed by Pensoft that supports the full life cycle of a manuscript, from authoring and reviewing to publishing and dissemination. ARPHA consists of two interconnected workflows. A journal can use either of the two or a combination of both (**Fig. 5**): 1) ARPHA-XML web-based authoring, peer-review and publishing, and 2) ARPHA-DOC - Document-based peer-review and publishing. The XML-based workflow is currently used by three journals of Pensoft – Biodiversity Data Journal, Research Ideas and Outcomes and One Ecosystem. The second, file-based submission workflow, is currently used by 12 journals published by Pensoft.

⁷² <http://bdj.pensoft.net/>

⁷³ <http://arpha.pensoft.net/>

⁷⁴ <http://arphahub.com/>

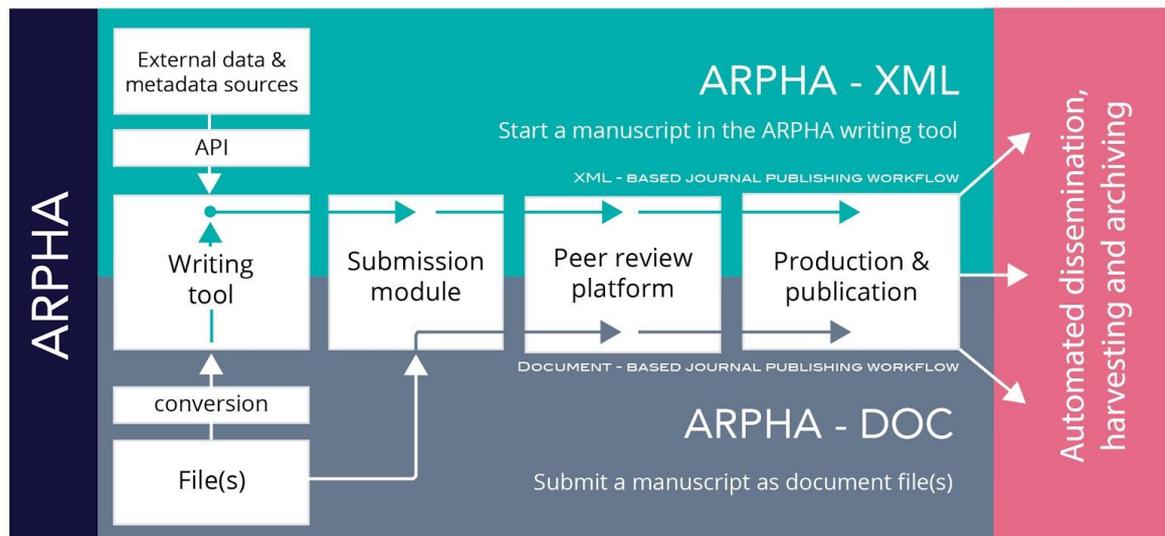


Figure 5. ARPHA consists of two integrated workflows: in ARPHA-XML, the manuscript is written and processed via the ARPHA Writing Tool, and in ARPHA-DOC, the manuscript is submitted and processed as document file(s).

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. ARPHA was successfully prototyped in 2013 by the Biodiversity Data Journal and the associated Pensoft Writing Tool. The latter, together with the document-based Pensoft Journal System (PJS), has since been upgraded, re-factored and re-branded into a generic ARPHA authoring, editorial and publishing platform. The core of this novel workflow is a collaborative online manuscript authoring module called ARPHA Writing Tool (AWT). AWT's innovative features allow for upfront markup, atomization and structuring of the free-text content already during the authoring process, import/download of structured data into/from human-readable text, automated export and dissemination of small data, on-the-fly layout of composite figures, and import of literature and data references from trusted online resources into the manuscript. ARPHA is also probably the world's first publishing system that allows submission of complex manuscripts via an API.

ARPHA provides:

- Full life cycle of a manuscript, from writing through submission, revisions and re-submission within a single online collaborative platform;
- Conversion of Darwin Core and other data files into text and vice versa, from text to data;
- Automated import of data-structured manuscripts generated in various platforms (Scratchpads, GBIF Integrated Publishing Toolkit (IPT), DataOne data base, authors' databases);
- Automated import of occurrence data from BOLD, iDigBio and GBIF platforms;

- A set of pre-defined, but flexible, Biological Codes and Darwin Core compliant, article templates;
- Easy online collaborative editing by co-authors and peers;
- A novel, community-based and public, pre-submission, pre-publication and post-publication peer-review processes.

Enhancement by EU BON:

The ARPHA Writing Tool was identified as one of the important EU BON products for data mobilization and will be incorporated into the data publishing toolbox of the EU BON Portal.

A number of improvements of the tool were implemented as part of the project. A new plugin developed as part of EU BON to a workflow previously developed by the GBIF and Pensoft, and tested with datasets shared through GBIF and DataOne, now makes it possible to convert metadata into a manuscript for scholarly publications, with a click of a button. Pensoft has currently implemented the feature for biodiversity, ecological and environmental data. Such records are either published through GBIF or deposited at DataONE, from where the associated metadata can be converted directly into data paper manuscripts within the ARPHA Writing Tool, where the authors may edit and finalize it in collaboration with co-authors and peers and submit it to the Biodiversity Data Journal with another click.

Another new feature developed makes it possible to easily import occurrence records into a taxonomic manuscript in ARPHA. This streamlines the authoring process and significantly reduces the time needed for creation of a manuscript. Substantial amount of documented occurrence records awaiting publication are stored in repositories and data indexing platforms, such as the Global Biodiversity Information Facility (GBIF), Barcode of Life Data Systems (BOLD Systems), or Integrated Digitized Biocollections (iDigBio). A new upgrade of ARPHA now allows by simply specifying an identifier (ID) in the relevant box, occurrence data, stored at GBIF, BOLD systems, or iDigBio, to be directly inserted into the manuscript. It all happens in the user-friendly environment of the AWT, where the imported data can be then edited before submission to the Biodiversity Data Journal or other journals using ARPHA. Not having to retype or copy/paste species occurrence records, the authors save a lot of effort. Moreover, they automatically import them in a structured Darwin Core⁷⁵ format, which can be easily downloaded from the article text into structured data by anyone who needs the data for re-use after publication.

Furthermore, a technical workshop on development of automated workflow between PlutoF and ARPHA to streamline publication of PlutoF data through Biodiversity Data Journal was held in November 2015, in Bulgaria. The workshop was attended by representatives of Pensoft and UTARTU. PlutoF is a biological data management system maintained by the University of Tartu consisting of several modules/data objects: ecological molecular projects, genomic data, citizen science, taxon occurrences, these projects, natural history collections, etc. The purpose of the workshop was to find technical solution for automatic export and

⁷⁵ <http://arphahub.com/>

integration of PlutoF data into Pensoft's ARPHA platform via API and its subsequent publication in Biodiversity Data Journal. Furthermore, the meeting aimed at discussing the publication of >400,000 fungal Species Hypotheses in MycoKeys. Pensoft and UNITE team have also discussed how to extend Pensoft taxon profile with information from PlutoF.

Testing and implementation:

Since its launch on 16th of September 2013 until February 2016, the journal has published altogether more than 250 articles, of which 34 data papers and 10 software descriptions. The journal has got more than 1,500 users and their number increases on a daily basis.

One of the major data mobilization initiatives realized by ARPHA and BDJ is the publication of data papers on the largest European animal data base 'Fauna Europaea'. A new series 'Contributions on Fauna Europaea' was launched at the beginning of 2014. This novel publication model was aimed to assemble in a single collection 57 data-papers on different taxonomic groups covered by the Fauna Europaea project and 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 2015, 11 articles altogether have been published in BDJ (de Jong et al., 2014).

A tutorial for the use of ARPHA called "Trips and tricks" is available on the website at: <http://arpha.pensoft.net>.

Tool status:

The AWT is fully operational and currently used by three Pensoft journals – Biodiversity Data Journal, Research Ideas and Outcomes⁷⁶ and One Ecosystem⁷⁷. New functionalities are added continuously in line with the increased interest in publishing scientific data.

Future developments:

Enhancement of AWT and BDJ for traits data, and sample based Darwin Core compliant data sets is envisaged for the near future, as well as development and implementation of tools for visualization of genomic data. New article type templates are also scheduled, for instance IUCN compliant species conservation profile. Also, currently, the BDJ and AWT are constrained to be used mostly by the biodiversity community, so expansion to other scientific domains is in the forthcoming tasks of Pensoft IT department.

⁷⁶ <http://rio.pensoft.net>

⁷⁷ <http://oneecosystem.pensoft.net>

Plazi TreatmentBank and DwC

Tool description:

Plazi's⁷⁸ TreatmentBank⁷⁹ provides access to, and makes taxonomic treatments and included data of taxa citable by minting persistent identifiers. Taxonomic name usages refer implicitly or explicitly to a particular underlying concept of the applied name. In the latter case, a specific section includes a documentation of the traits and distribution of a related group of group of organisms (taxon)⁸⁰, called taxonomic treatment. There are millions of treatments in the scientific literature, which form an extremely valuable source of information. These treatments are increasingly linked to their underlying data, such as observation data, keys for identifications or other digital objects, and very often they cite each other. Once semantically enhanced, the data is a powerful source for analyses and visualizations at any given level (Miller et al., 2015). Often these are the only records of rare species and thus contribute substantially to uncover the vast majority of biodiversity (Miller et al., 2015). There are two bottlenecks to providing semantically useful modern Internet access at this level. The first is that the vast majority are not even digitally available, or at most are parts of semantically unstructured PDF-formatted documents. The second is that a substantial amount of the literature is only accessible through a paywall or comes with restrictions on their use. With the increasing wealth of digitized observation records, upon which most of the publications are based, it becomes imperative to provide retro access to the treatments, to link to them, and to enhance them with links to the material referenced in them. The Plazi workflow (**Fig. 6**) is a tool to achieve this conversion within a legal framework (Agosti & Egloff, 2009).

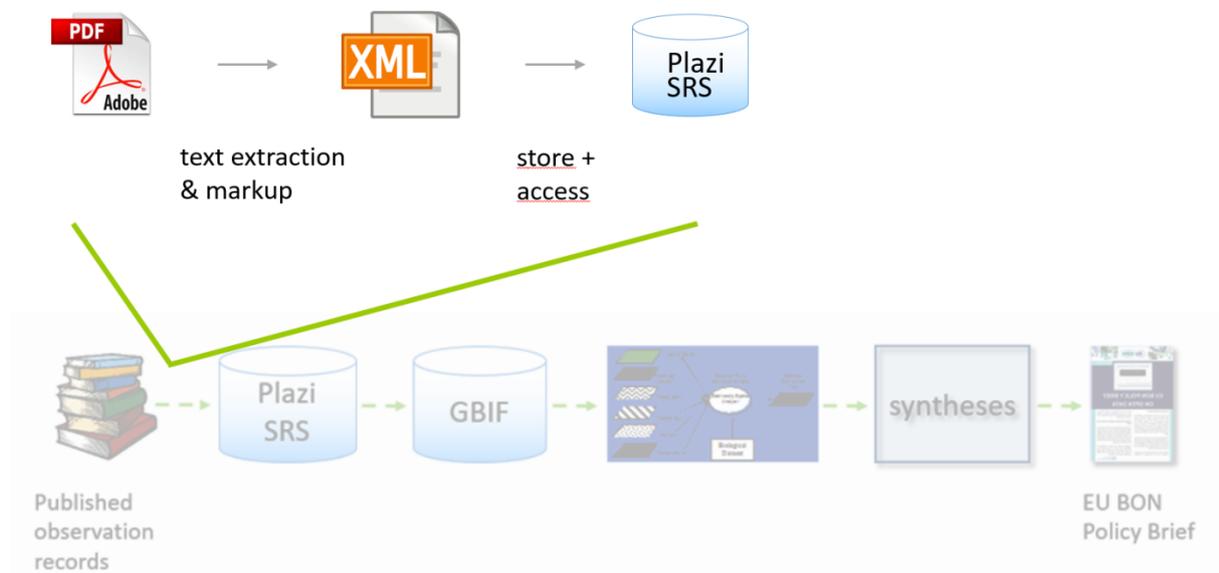


Figure 6. The Plazi workflow (green) within EU BON.

⁷⁸ <http://plazi.org>

⁷⁹ <http://bdj.pensoft.net/articles.php?id=5063>

⁸⁰ <http://www.ncbi.nlm.nih.gov/books/NBK47081/>

TreatmentBank covers this niche. It offers with GoldenGate⁸¹ and respective XML schemas (TaxonX⁸², TaxPub⁸³) open source tools to convert unstructured text into semantically enhanced documents with an emphasis on taxonomic data like treatments, scientific names, materials observation, traits or bibliographic references (Miller et al., 2015; Catapano, 2010). A complementary source is the automatic, daily import of treatments from TaxPub based publications (i.e. Pensoft family journals). Within EU BON, for a number of ongoing Open Access journals GoldenGate versions will be produced allowing automatic preprocessing the conversion to minimize a human operator input. It provides a platform that can store, annotate, access and distribute treatments and the data objects within.

Within TreatmentBank annotations of literature to provide links to external resources, such as specimens, related DNA samples on GenBank, or literature can be stored. Annotation can be done at any level of granularity, from a materials citation to detailed tagging of specimens, provision of details of the collectors, or provision of morphological descriptions even to the tagging of individual traits and their states.

The use of persistent resolvable identifiers and the treatment ontology allows provision of RDF that supports machine harvest and logical analysis data, within and between taxa.

TreatmentBank provides access to data aggregators or other consuming external applications and human users, including entire treatments to the Encyclopedia of Life⁸⁴, and observation records to GBIF using Darwin Core Archives (**Fig. 7**). The latter is implemented, whereby for each new upload in TB, an update in GBIF is triggered.

Within EU BON, the GBIF pathway is the input of publication based data, specifically observation records that are linked to a treatment within an article, for EU BON's modeling activities (**Fig. 7**).

A unique value of TreatmentBank to GBIF and EU BON is that approximately half of the taxa are not covered within GBIF, and thus it is contribution to the vast majority of the rare or little covered species (Miller et al., 2015).

⁸¹ <http://plazi.org/?q=GoldenGate>

⁸² <http://plazi.org/?q=taxonx>

⁸³ <https://github.com/tcatapano/TaxPub/releases>

⁸⁴ <http://eol.org>

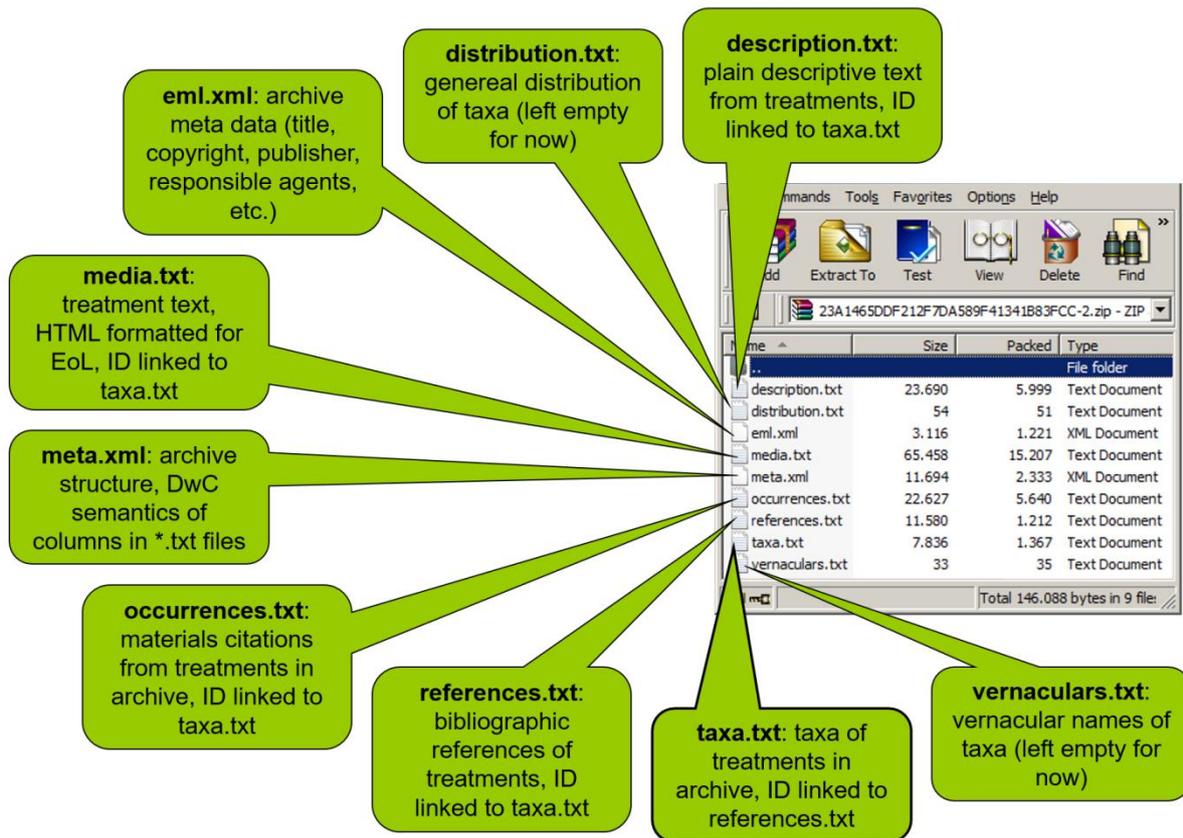


Figure 7. The implementation of Darwin Core Archive in Plazi to transfer treatment data. Observation data described with Darwin Core terms.

TreatmentBank is a one of its kind. With the US ETF⁸⁵ project, there is one complementary workflow known that focuses on traits, that collaborates with Plazi. TreatmentBank is built and maintained by highly skilled personnel, it is growing through regular input from Pensoft, synchronization with Zoobank and in-house processing of articles. It is part of Plazi 1 Million Treatment project to establish Open Access to the content of taxonomic publications by developing various tools to convert new treatments.

TreatmentBank is complemented by activities regarding legal status of treatments and other scientific facts, semantic developments, especially linking to external vocabularies and resources, and use by a number of high profile operations (GBIF, EOL, EU BON, Pro-iBiosphere⁸⁶, domain specific web sites). Currently 93000 treatments from 7633 articles are available.

New technical requests can be met quickly, and Plazi has in recent years been on the forefront to build interfaces to import data into GBIF, EOL or Map of Life (i.e. DwC A). Plazi uses RefBank⁸⁷ as a reference system for bibliographic references and is working in close collaboration with Zenodo (Biosystematics Literature Community, BLC)⁸⁸ to build a

⁸⁵ http://biowikifarm.net/v-botknow-test/web/About_BKP

⁸⁶ <http://www.pro-ibiosphere.eu/>

⁸⁷ <http://refbank.org/>

⁸⁸ <https://zenodo.org/collection/user-biosyslit>

repository for articles that are not accessible in digital form. To discover bibliographic references, Refindit⁸⁹ is used and developed.

Future developments:

- TreatmentBank is not yet industrial strength and will need in its next phase to assess how to move from a research site to a service site.
- GoldenGate, the TreatmentBank's central tool is powerful, but a more intuitive human-machine interface needs be developed.
- Customized versions of GoldenGate for taxonomic journals should be increased and crawlers to discover new issues to be harvested installed.
- Specific services, such as bibliographic name provision and materials examined parsing need to become standalone applications.
- Trait extraction needs be developed.
- TreatmentBank should become part of the LifeWatch IT infrastructure.
- In the short term, it is important to build a critical corpus of domain specific treatments to allow scientifically meaningful data mining and extraction. This may require extensive data be gathered from treatment authors.
- Make Plazi TreatmentBank a contributor to the EU BON taxonomic backbone.

The project in general is underfunded and understaffed. It needs to invest in human-machine interfaces, documentation and training, and tools that allow the easiest possible way to annotate the treatments, and especially to increase the daily conversion rate. Training activities need to be resumed and a proper training curriculum for users implemented.

Tool status:

This tool is ready to be used.

⁸⁹ <http://refindit.org/>

Metacat and Morpho

Tool description:

Metacat: Metadata and Data Management Server:

Metacat⁹⁰ is an open source metadata catalog and data repository that targets scientific data, particularly from ecology and environmental science. It is a key infrastructure component for the NCEAS data catalog, the Knowledge Network for Biocomplexity (KNB) data catalog, and for the DataONE system, among others.

The information is available through the data packages, which consists of the data set associated with its corresponding metadata. It can be easily searched, compared, merged, or used in other ways⁹¹ (for more information, see Annex 1).

Metacat is a Java servlet application that runs on Linux, Mac OS, and Windows platforms in conjunction with a database, such as PostgreSQL (or Oracle), and a Web server. The Metacat application stores data in an XML format using Ecological Metadata Language (EML) or other metadata standards such as ISO 19139 or the FGDC Biological Data Profile⁹².

Metacat's user-friendly Registry application allows data providers to enter data set documentation into Metacat using a Web form. Metacat users can also choose to enter metadata using the Morpho application, which provides data entry wizards that guide information providers through the process of documenting each data set¹. A data centre using Metacat can become DataONE member node with a relatively simple configuration.

Flexibility that allows organising and preserving heterogeneous datasets comes together with the drawback that it is not possible to query the data tables directly. PPBio found that it was necessary to provide auxiliary tables (<http://ppbio.inpa.gov.br/repositorio/dados>) to allow sampling effort to be evaluated effectively in most situations.

Morpho Metadata Editor:

Morpho is a user-friendly application designed to facilitate the creation of metadata⁹³. Morpho interfaces with the Knowledge Network for Biocomplexity (KNB) Metacat server. After the data are annotated with metadata, the user can choose to upload the data or just the metadata to the Metacat server, where they can be accessed from the web by selected colleagues or by the public. Metadata are stored in a file that conforms to the Ecological Metadata Language (EML) specification. Data can be stored with the metadata in the same file. Morpho allows the user to create a local catalogue of data and metadata that can be queried, edited and viewed¹.

Morpho has an advantage relate to the registry shipped within Metacat which is the Data Table description. Users need to install the tool in their local machines.

⁹⁰ <https://www.dataone.org/software-tools/metacat>

⁹¹ information provided by Metacat Administrator's Guide: <http://knb.ecoinformatics.org/software/metacat>

⁹² information provided by Metacat Administrator's Guide: <http://knb.ecoinformatics.org/software/metacat>

⁹³ information provided by Morpho User Guide:

<https://knb.ecoinformatics.org/software/dist/MorphoUserGuide.pdf>

Testing and implementation by EU BON:

Some EU BON test sites (such as Sierra Nevada Observatory and Brazilian Research Program in Biodiversity) are using Metacat and Morpho for data management. PPBio (INPA) set up and tested the Metacat metadata catalogue and data repository system that runs on the PELD Data Repository⁹⁴, which is a DataONE node/deployment in Manaus Brazil. Some negative points were noted:

- Flexibility that allows organizing and preserving heterogeneous datasets comes together with the drawback that it is not possible to query the data tables directly. PPBio found that it was necessary to provide auxiliary tables⁹⁵ to allow sampling effort to be evaluated effectively in most situations.
- Effective installation can require fairly advanced knowledge of TI, and the documentation is sometimes out-of-date. However, the backup provided by their help-desk is very good.
- Lack of github repository (turn contributions to be a bit more slow).
- No way to explore ecological data besides points in map.

Morpho is the default interface to upload data from desktops and is mainly used because it's necessary check the metadata/data sent in by the researchers before it gets uploaded to Metacat. Interface is really buggy and not user friendly to setup. Morpho is not currently undergoing development.

GBIF is collaborating with DataONE in developing a data accessor to allow a GBIF IPT to operate independently in the DataONE network, thus bridging Metacat based datasets to EU BON Portal. Major issues to deal with are cross mapping between metadata and preventing data replications, given data sets are available through multiple providers.

Future developments:

The main context for use is to match the needs of EU BON as a repository for tabular data. If there are specific projects that deal with tabular data at a standardized perspective – spatial, temporal or taxonomic, it is recommended, based on PPBio experience, to build standardized data tables that will facilitate further integration. Additional development to extend the tool in order to provide a customized data-entry interface that suits the particular requirements of each project can be considered.

The Metacat tool manages to consume the same EML harvest list endpoint as DEIMS provides, but with some small differences, maybe because of the specific version of the harvest list schema (DEIMS harvest list: <https://data.lter-europe.net/deims/eml/harvest-list-all.xml>; Metacat harvest list (from Sierra Nevada): <http://linaria.obsnev.es/panel/harvestlist>).

⁹⁴ <http://www.massapeld.ufc.br/repository/>

⁹⁵ <http://ppbio.inpa.gov.br/repositorio/dados>

During Seville hackathon⁹⁶ (26-28 January 2016), the test harvest of Granada's Metacat using its harvest list ended without success. The harvest list was compliant with EML 2.0.0 whilst GI-cat needs EML 2.1.1 compliant endpoints. The translation between both formats is feasible, e.g. using XSLT translate stylesheets, however those metadata files uploaded directly to Metacat, but not harvested, would not be published.

As a feasible alternative to retrieve metadata from Metacat Instances, the optional Metacat OAI-PMH data provider⁹⁷ could be installed in each test site instance, As far as GI-cat manages OAI-PMH endpoints as metadata providers, Metacat instances would be directly harvested by the GI-cat registry periodically.

Because Morpho doesn't recognize a multi-domain SSL certificate it would be logic to replace Morpho (or having as a backup method) with Metacat's optional web-based interface for uploading data.

During the Manaus workshop it was also discussed about metadata mapping (Morpho vs. IPT). Within that context it may worth considering if and what metadata fields related to systematic - monitoring schemes should be account for mapping Metacat/LTER datasets to EML/DwC.

Tool status:

Tools are ready to be used.

⁹⁶

<http://www.eubon.eu/show.php?storyid=13309&title=Shaping%20the%20EU%20BON%20Biodiversity%20Portal>

⁹⁷ <https://knb.ecoinformatics.org/knb/docs/oaipmh.html>

PlutoF

Tool description:

The PlutoF cloud⁹⁸ provides online service to create, manage, share, analyze, and mobilize biodiversity data. Data types cover ecology, taxonomy, metagenomics, nature conservation, natural history collections, etc. Common platform aims to grant the databases with professional architecture, sustainable developing and persistence. It provides synergy through common modules for the classifications, taxon names, analytical tools, etc. Common taxonomy module is based on available sources (e.g. Fauna Europaea, Index Fungorum) and may be developed collectively further by the users. Currently there are more than 1500 users who develop their private and institutional databases or use analytical tools for biodiversity data. PlutoF cloud also provides data curation, possibilities, including third party annotations to the data from external resources, such as genetic data from GenBank⁹⁹. PlutoF is developed by the IT team of Natural History Museum (University of Tartu, Estonia).

Curated datasets hosted by PlutoF cloud can be made available through public web portals. Examples include the UNITE community which curate DNA based fungal species and provide open access to their datasets through UNITE portal¹⁰⁰. Another example is eBiodiversity portal¹⁰¹ that includes taxonomical, ecological and genetics information on species found in Estonia. Any public dataset in PlutoF cloud that includes information on taxa found in Estonia will be automatically displayed in this portal. This enables to discover biodiversity information for Estonia in one portal.

Implementation of mobile app tools for citizen science sighting reports with PlutoF API:

Community-based data generated through collaborative tools and resources increasingly becomes a serious approach for mobilizing and generating biodiversity data for assessment and monitoring.

PlutoF API provides a structured system that eases the implementation of citizen-science based mobile app reporting schema, thus facilitating community-based tools for data sharing. Building on the PlutoF API tools supports the primary challenge of the EU BON project to make citizen-science data qualified, available, discoverable and publicly shared.

a. Mobile App tools to support and encourage public sighting reports

Beyond the attractiveness of using state-of-the-art tools to activate the public, mobile app tools empower citizen science recording schemes and support public participation in science with a range of advantages:

- Many people across the world own mobile phones and tablets, enjoy using them, and use a range of apps (applications).

⁹⁸ <http://plutof.ut.ee>

⁹⁹ <http://www.ncbi.nlm.nih.gov/genbank/>

¹⁰⁰ <http://unite.ut.ee>

¹⁰¹ <http://elurikkus.ut.ee>

- Among the young generation, the potential of using high-end IT tools attracts attention and interest.
- Apps are handy and easy to develop and use.
- Devices offer advanced technologies to collect and communicate valuable data in the field for enhancing data-accuracy and accurate spatial precision.
- Apps minimize effort of the user, thus, they offer an excellent tool to enhance public, voluntary participation (experts and hobbyists alike) in biodiversity reporting.
- From a policy perspective (Habitat and Birds' Directives, EBVs), apps can facilitate rapid reporting, validation, analyses and inform policy-makers in near-real time.
- Apps broaden the range of data by allowing the collection of other types of information (photos, sounds) which may provide and reveal additional data and metadata, such as habitat, behavior.

b. A Citizen Science based approach for collection and qualification of biodiversity data

The design concept of the two mobile apps developed by GlueCAD (a. for sporadic observations, b. for transects based systematic monitoring) and sound recording app by University of Tartu Natural History Museum, is based on a citizen science approach aimed at getting data that: (1) takes advantage of the device technology, (2) relies less on the skill of the user, (3) supports data with fields for efficient validation and qualification.

In practical terms it means relying on high-end IT devices to obtain the **maximum amount of data with the minimum of typing**, allowing volunteers to concentrate on observing, rather than data entry. The concept involved getting automatic and implied data rather than relying on the skills of the user.

Some practical examples:

- Getting GPS information on spatial location (coordinates) as well as altitude, coordinates-accuracy and date/time for every reported species.
- Weather data can be extracted, mostly online, from nearby meteorological stations.
- Using Standard Species lists to select from.
- The speed of movement can be measured to estimate sampling-effort.
- Activation of the camera adds documents the record; may improve validation capacity and may further contribute to information about the host plants and habitat.
- Using sound recording capability of mobile devices can add multimedia content for validating observations of vocally active animals - birds, frogs, insects etc.
- A registered observer is given a user ID (which is kept in the device memory) so that there is no need to retype user details.

- Facilitation of quality control by providing information to assist validation, e.g. source of data, identified by.
- Offer observers to the option of different identification methods such as identify by list, by pictures or by voice).

c. Relying on PlutoF Taxonomic DB

Observations reported through GlueCAD's apps rely on ad-hock querying of the API for taxon IDs, thus provides a dynamic adaptation to PlutoF, namely the standard, taxonomic backbone.

It also enables future extensions to support the downloading of other taxon lists to be used for sighting reports.

d. Managing observation data with PlutoF workbench

PlutoF system allows for the support of observation moderation for any project. Observation data will be then moderated by assigned expert, before going on public display. Expert can use PlutoF messaging interface to ask for additional information from user to accept or decline taxon identifications for a specific observation. They can also use added multimedia content (photos, videos, sound recordings) for taxon identification. Every change in taxon identification will be recorded and can be traced within the system.

Future developments:

University of Tartu Natural History Museum will continue the development of PlutoF services, partly linked to developments of national science infrastructure. New modules include water ecosystems, environmental samples and NGS, plant and forest pathology, governmental module, and LTER module (if collaboration will continue).

Tool status:

- Web-based services are available for individual users, workgroups and institutions. New infrastructure based on different technologies is under development and its beta version is available. PlutoF Platform is developed by a team of eight software engineers.
- The mobile app for sporadic observations reporting, called “I Saw a Butterfly” is out, free, on Google Play (**Fig. 8**). Observations are reported to PlutoF.
- The second app from GlueCAD for systematic observations (“BMSapp”) is currently being tested by INPA with Amazon’s frog list (100) and by the Israeli group of the butterflies monitoring scheme.

- Based on the range of taxonomic groups supported by PlutoF API, it is possible to upgrade and facilitate the mobile app with extended lists of taxa groups for biodiversity observations recording and data sharing.

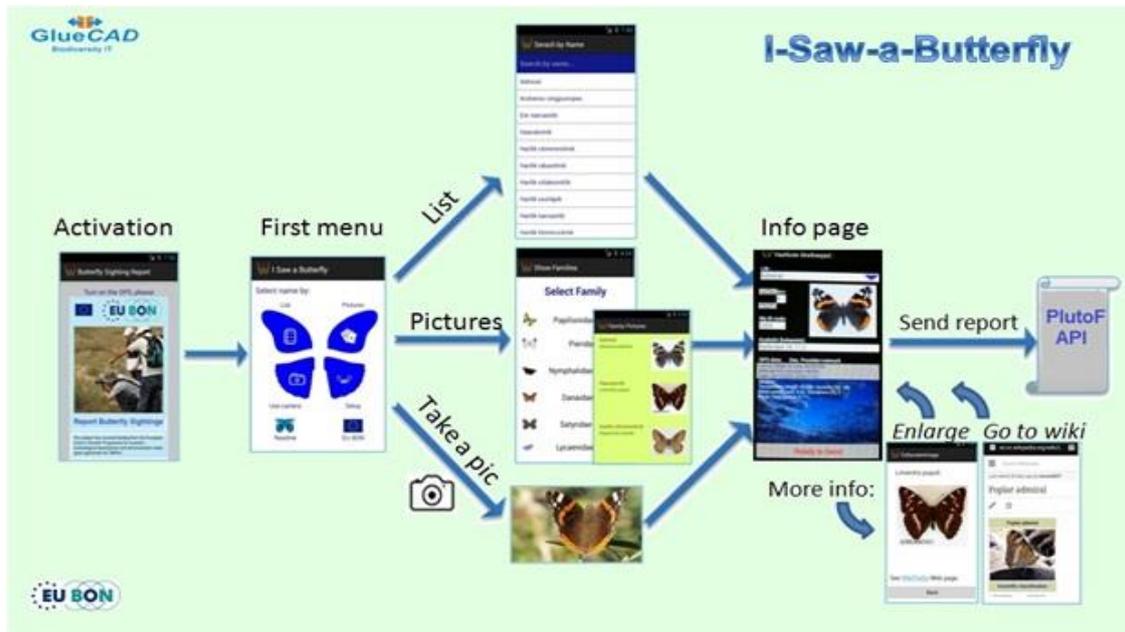


Figure 8. Mobile app for sporadic observations reporting.

6. Future developments and conclusions

Challenges

Different studies have (Tenopir et al., 2011; Hardisty et al., 2013) discussed the results of surveys conducted to understand how data are treated by scientists across different disciplines. From these surveys it can be deduced that, contrary to expectations, in our modern digital age data are not often shared openly. Hardisty et al. (2013) show that only between 6-8% of the researchers deposit datasets in an external archive of the research domain! The most common environment for storing, managing and reusing data remains the lab and/or individual working environment, including the desktop PCs. The main obstacles identified are insufficient time, reluctance in learning new approaches and lack of funding. So sharing data is still a complex and challenging issue.

Based on these studies, challenges identified by GEOSS¹⁰² and our own experience, several focus points can be picked up:

- Open Data
- Data standardization
- Data mobilization

¹⁰² <http://www.spacelaw.olemiss.edu/jsl/pdfs/articles/35JSL201.pdf>

Open Data should be normal practice and should embody the principles of being discoverable, accessible, intelligible and usable. This concerns also appropriate metadata describing data sources and processing. We are well advanced in this aspect by working together with GBIF on enhancement of their IPT tool for publishing sample-based data and by promoting the data paper concept together with Pensoft allowing easier and faster publishing of research data and metadata.

Of paramount importance here is to foster data mobilization in collaboration with such endeavors GEO BON, LTER, etc. to provide more diverse data to GBIF and to test advantages and disadvantages of the new functionalities of the IPT. These may result in further recommendations and updates/new releases of the current IPT. The usage of the new Darwin Core terms will also have to be followed up and feedback from the community (the TDWG Darwin Core working group and the GBIF IPT mailing list¹⁰³) will be taken into account.

For instance further recommendations on which core to choose when providing data to GBIF, depending on the structure of the data, are crucial points currently under discussion. The “Measurement Or Facts” extension was previously linked to the Occurrence Core and was used primarily to provide facts or measurements about the specimens and/or observations. The same extension linked to the Event Core allow the provision of habitat variables, parameters and descriptions. This leads to discussion on the pros and cons of the star schema approach *versus* using flat files and on how to interlink the different tables, as it is possible to map the same information with several core concepts. Feedback from training events shows that potential providers have a hard time to decide which core to use, as the data are often at the borderline between Occurrences or Event Core centered datasets.

The metadata are another point of attention for the future. The sampling protocols and procedures are stored in fields in the metadata part. Providers are encouraged to fill them in thoroughly. By fully completing these fields, this should simplify the publishing of the dataset as a data paper. However these fields are not part of the DarwinCore terms and remain simple text boxes with recommendations on information to be added and are only meant to be human-readable. Controlled vocabularies for some of these terms within the data itself (e.g. sampling protocol) are also needed for machine-readable metadata for instance.

Controlled vocabularies can be used within the DarwinCore terms to provide information on the “Gathering Event” such as including sampling methods, equipment used, information on the vessels used, the expeditions the participating actors and the funding bodies. Data providers should be encouraged to complete both data and metadata and not to consider the human-readable metadata as a substitute of the machine-readable data which may also be needed. Thus the need for controlled vocabularies and additional terms describing the Gathering Event should be further investigated.

Interesting questions were raised during last training events on the IPT and on the various mailing lists on how to provide data from a monitoring scheme, where different sampling protocols were used during a same campaign in a same area. Should it be provided as several

¹⁰³ <http://lists.gbif.org/mailman/listinfo/ipt/>

datasets each with its specific sampling protocol or can they be provided in form of one dataset listing the different sampling protocols to which the corresponding occurrences, checklists and measurement or facts should be linked to? Having a repeatable “Gathering Event” concept with associated terms as it is the case for example in the TDWG ABCD (Access to Biological Collection Data) schema could further looked into to answer these questions.

Last but not least, questions were asked during different discussions on how to make the sample-based datasets directly discoverable when searching from the GBIF data portal, as it seems that the new terms are currently not yet indexed and thus not searchable.

In conclusion, providers and users, should be encouraged to be active in mobilizing sample-based data and to give feedback to GBIF, so that they can be further adapted and triggered to meet the needs and expectations of the community. They should also learn and be trained to provide adequate metadata for their data. Free and open access to it should be widely promoted.

Data standardization or encoding should allow analysis across multiple scales. The arrangements and standards for data access and sharing will facilitate the integration of various data sets.

There still is no central entry point for the dispersed and heterogeneous biodiversity data (Wetzel et al., 2015). In order to enhance data discoverability and accessibility, EU BON has chosen to implement on its portal different tools compatible to the majority of standardized metadata formats (e.g. ISO 19115, EML and OGC CSW standards) which will allow the discovery and access of data sets stored in a range of biodiversity registries and catalogues. The developed software components and tools will be freely available in order to provide other BONs with a basic technological framework for their data mobilizing approaches.

Different sites are using different systems for sharing information and the challenge is trying to integrate all this information in a single metadata repository where all biodiversity information regarding EU BON appears. Future developments together with WP5 should deal with this issue, hopefully solving the limitations found in the tools that are being currently used (MS517).

Although the metadata language, EML, guarantees data discoverability, the raw data must also be accessible for automated data integration. Data mining tools (e.g. GoldenGate Image and Scratchpads) and further knowledge discovery would certainly help to make additional data available.

EU BON supported data sharing tools only cover a part of biodiversity data types which are relevant for earth observation. Most notably, specialized tools for sharing habitat data are not covered. There are very few such tools, as habitat data is not shared very much, and such data can rather easily be exchanged using general purpose GIS and database tools. Nevertheless, the EBONE project¹⁰⁴ did develop a specialized tool for habitat data, based on

¹⁰⁴ <http://www.wageningenur.nl/en/Expertise-Services/Research-Institutes/alterra/Projects/EBONE-2.htm>

Microsoft Access. We have evaluated this tool, but have chosen not to take further action, because the needs for sharing habitat data beyond what EBONE has already achieved have not yet been articulated.

There is an agreement between EU BON and LTER to collaborate further on sharing the metadata among EU BON and LTER tools and sites (**Fig.9**). EU BON will provide feedback about the integration of DEIMS in the EU BON registry, taking into account that biodiversity-related metadata must not be degraded during the translation processes, and in fact may need to be expanded with more detailed taxa information. LTER will provide EU BON with feasible alternatives to extract metadata from DEIMS and related tools.

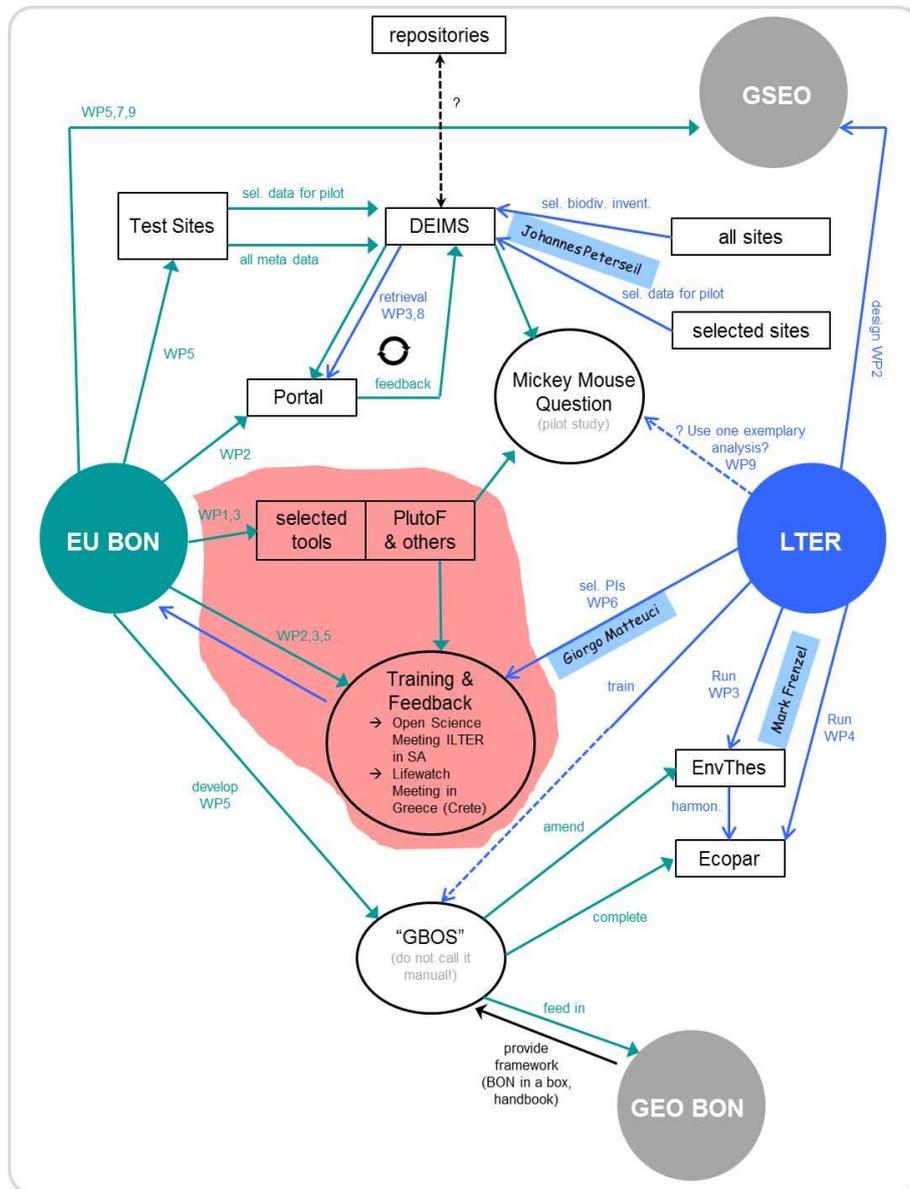


Figure 9. Information flows between EU BON and LTER Europe, as envisaged on the 3rd EU BON Stakeholder Roundtable in Granada on 9-11 December 2015.

Data mobilization. EU BON is not an infrastructure project, but does have a significant infrastructure development component. As such, EU BON should devote significant resources to promote the tools and services they develop and to attract users from outside the project-funded community. The thorough gap assessment conducted by EU BON shows the most obvious temporal, spatial and taxonomic biodiversity data gaps (**Fig.10**). These are largely due to lack of data sharing practices.



Figure 10. A typical distribution map from GBIF in 2014. Gaps in distribution are clearly visible for almost any species.

To this end the EU BON commenced ongoing campaigns which should gradually lead to mobilize biodiversity data across borders, e.g. by fostering citizen science awareness and activities enforcing with guidelines towered communities that can assemble and upload their data (Wetzel et al., 2015). Special focus is on approaching systematic monitoring schemes, promoting the newly extended standards for quantitative data, which builds on the developments made in EU BON Tasks T2.2 and T2.3 for standards development and upgraded tools for sample-based data. EU BON is working with the legacy of the EuMon project¹⁰⁵ to approach all quantitative biodiversity monitoring schemes in Europe for mobilizing their data. The EuMon metadatabase currently contains 639 descriptions of monitoring schemes, but the real number of them is probably about three-fold. Mobilizing this huge wealth of data will be a major achievement. In the remaining project time, in the

¹⁰⁵ <http://eumon.ckff.si/>

best case, EU BON can only get this process started. It remains for GEO BON, GBIF, the EuMon legacy, and future projects to bring this process to a completion.

For mobilizing data and promoting data sharing, EU BON has developed comprehensive training program, with a focus on data and metadata integration strategies, use of standards and data sharing tools for institutional data and IT managers, researchers, citizen scientists and monitoring programs. Several technical (informatics) workshops have been held on data standards and prototypes, e.g. of data sharing tools and the biodiversity portal. More are planned for biologists and for other life scientists from Eastern Europe who are actively involved in monitoring and managing biodiversity data. Results of that activity will be reported in the deliverable D2.4.

7. Bibliographic references

1. Agosti D, Egloff W (2009) Taxonomic information exchange and copyright: the Plazi approach. *BMC Research Notes* 2009, [2:53]. DOI: [10.1186/1756-0500-2-53](https://doi.org/10.1186/1756-0500-2-53) Anderson K (2104) Data sharing and science — Contemplating the value of empiricism, the problem of bias, and the threats to privacy. <http://scholarlykitchen.sspnet.org/2014/03/05/data-sharing-and-science-contemplating-the-value-of-empiricism-the-problem-of-bias-and-the-threats-to-privacy/>.
2. Catapano T (2010) TaxPub: An extension of the NLM/NCBI journal publishing DTD for taxonomic descriptions. Proceedings of the Journal Article Tag Suite Conference 2010 (<http://www.ncbi.nlm.nih.gov/books/NBK47081/>).
3. Chavan V, Penev L (2011) The data paper: a mechanism to incentivize data publishing in biodiversity science. *BMC Informatics*, 12 (Suppl 15): S2 (accessed online at <http://www.biomedcentral.com/1471-2105/12/S15/S2> on 18/03/2014).
4. Chavan V, Penev L, Hobern D (2013) Cultural change in data publishing is essential. *BioScience*, 63, 6, 419-420. doi:10.1525/bio.2013.63.6.3.
5. de Jong Y, Verbeek M, Michelsen V, Bjørn P, Los W, Steeman F, Bailly N, Basire C, Chylarecki P, Stloukal E, Hagedorn G, Wetzel F, Glöckler F, Kroupa A, Korb G, Hoffmann A, Häuser C, Kohlbecker A, Müller A, Güntsch A, Stoev P, Penev L (2014) Fauna Europaea – all European animal species on the web. *Biodiversity Data Journal* 2: e4034. doi: 10.3897/BDJ.2.e4034.
6. Hardisty A, Roberts D, the Biodiversity Informatics Community (2013) A decadal view of biodiversity informatics: challenges and priorities. *BMC Ecology*, 13:16, DOI: 10.1186/1472-6785-13-16.
7. Hoffmann A, Penner J, Vohland K, Cramer W, Doubleday R, Henle K, Kõljalg U, Kühn I, Kunin WE, Negro JJ, Penev L, Rodríguez C, Saarenmaa H, Schmeller DS, Stoev P, Sutherland WJ, Ó Tuama É, Wetzel FT, Häuser CL (2014) Improved access to integrated biodiversity data for science, practice, and policy - the European Biodiversity Observation Network (EU BON). URL/DOI: <http://www.pensoft.net/journals/natureconservation/article/6498/the-need-for-an-integrated-biodiversity-policy-support-process-%E2%80%93-building-the-european-contribution-to-a-global-biodiv>.
8. Marx V (2013) Biology: The big challenges of big data. *Nature* 498, 255–260 doi:10.1038/498255a (<http://www.nature.com/nature/journal/v498/n7453/full/498255a.html>).
9. Miller J, Agosti D, Penev L, Sautter G, Georgiev T, Catapano T, Patterson D, King D, Pereira S, Vos R, Sierra S (2015) Integrating and visualizing primary data from prospective and legacy taxonomic literature. *Biodiversity Data Journal* 3: e5063. doi: [10.3897/BDJ.3.e5063](https://doi.org/10.3897/BDJ.3.e5063).

10. Ó Tuama É (2015) Publishing sample data using the GBIF IPT: http://www.gbif.org/sites/default/files/gbif_IPT-sample-data-primer_en.pdf.
11. Tenopir C, Allard S, Douglass K, Aydinoglu AU, Wu L, Read E, Manoff M, Frame M (2011) Data sharing by scientists: practices and perceptions, PLOS, DOI: 10.1371/journal.pone.0021101.
12. Wetzel FT, Saarenmaa H, Regan E, Martin CS, Mergen P, Smirnova L, Ó Tuama É, García Camacho FA, Hoffmann A, Vohland K, Häuser CL (2015) The roles and contributions of Biodiversity Observation Networks (BONs) in better tracking progress to 2020 biodiversity targets: a European case study, Biodiversity, DOI: 10.1080/14888386.2015.1075902
13. White EP, Baldrige E, Brym ZT, Locey KJ, McGlinn DJ, Supp SR (2013) Nine simple ways to make it easier to (re)use your data. Peer J PrePrints 1:e7v2 <https://doi.org/10.7287/peerj.preprints.7v2>.

Annex 1: Non-exhaustive list of tools

Below are the specifications of the tools surveyed, as to February 2015. Updates for the selected tools are available in the main text of the report. This list is also available on the EU BON Helpdesk website¹⁰⁶, where it will be updated as needed. Additional lists are available through the GBIF resources page¹⁰⁷, the DataONE software tools catalogue¹⁰⁸, and the BDTracker¹⁰⁹.

A.1 GBIF Integrated Publishing Toolkit (IPT)

Main usage, purpose, selected examples

The Integrated Publishing Toolkit is a free open source software tool written in Java that is used to publish and share biodiversity data sets and metadata through the GBIF network. Designed for interoperability, it enables the publishing of content in databases or text files using open standards, namely, the Darwin Core and the Ecological Metadata Language. It also provides a 'one-click' service to convert data set metadata into a draft data paper manuscript¹¹⁰ for submission to a peer-reviewed journal. Currently, the IPT supports two core types of data: checklists and occurrence data sets (plus data set level metadata).

The IPT is a community-driven tool. Core development happens at the GBIF Secretariat but the coding, documentation, and internationalisation are a community effort. New versions incorporate the feedback from the people who actually use the IPT. In this way, users can help get the features they want by becoming involved. The user interface of the IPT has so far been translated into six languages: English, French, Spanish, Traditional Chinese, Brazilian Portuguese, Japanese. New translations into other languages are welcomed.

The IPT is available for download in both compiled¹¹¹ and source code¹¹² versions.

As of September 2013, there are 104 IPT installations located in 87 countries serving 131 checklists published by 18 different publishers and 799 occurrence data sets published by 76 different publishers totalling 117.5 million records.

Examples of use of IPT

Darwin Core Archives are required for data harvest to the new VertNet¹¹³ portal and the IPT is seen as a great tool to facilitate the creation of these files and to provide hosting of them for participating institutions.

¹⁰⁶ <http://eubon.cybertaxonomy.africamuseum.be/data-sharing-tools-repository>

¹⁰⁷ <http://www.gbif.org/resources/summary>

¹⁰⁸ https://www.dataone.org/software_tools_catalog

¹⁰⁹ <http://bdtracker.cybertaxonomy.africamuseum.be/>

¹¹⁰ <http://www.gbif.org/publishingdata/datapapers>

¹¹¹ <http://www.gbif.org/ipt/releases>

¹¹² <https://code.google.com/p/gbif-providertoolkit/source/checkout>

¹¹³ <http://vertnet.org/>

*INBO (The Research Institute for Nature and Forest)*¹¹⁴ and *Canadensys*¹¹⁵ use the IPT as basis for a complete data mobilisation workflow from in-house data management systems to GBIF. The tool has been instrumental in the growth of the Canadensys network.

SiB¹¹⁶ Colombia uses the IPT as a central part of their data publishing model¹¹⁷ in which it has facilitated publication of primary data.

Pros and Cons of the tool

Pros

1. Publication of two types of biodiversity data: i) primary occurrence data (specimens, observations), ii) species checklists and taxonomies.
2. Integrated metadata editor for publishing data set level metadata.
3. Internationalisation: user interface available in six different languages: English, French, Spanish, Traditional Chinese, Brazilian Portuguese, Japanese; instructions are available for translating the interface¹¹⁸.
4. Data security: controls access to data sets using three levels of dataset visibility: private, public and registered; controls which users can modify data sets, with four types of user roles.
5. Integration with GBIF Registry: can automatically register data sets in the GBIF Registry; registration enables global discovery of data sets in both the GBIF Registry, and GBIF Data Portal.
6. Support for large data sets: can process ~500,000 records/minute during publication; disk space is the only limiting factor; for example, a published dataset with 50 million records in DwC-A format is 3.6 GB.
7. Standards-compliant publishing: publishes a dataset in Darwin Core Archive (DwC-A) format, a compressed set of files based on the Darwin Core terms, and the GBIF metadata profile based on the Ecological Metadata Language standard.
8. The tool is supported by good documentation and mailing list¹¹⁹; the User Manual is also available in both English¹²⁰ and Spanish¹²¹.

Cons

1. Currently [February 2015], the IPT can only be used for occurrence data sets and checklists

¹¹⁴ <https://www.inbo.be/>

¹¹⁵ <http://www.canadensys.net/>

¹¹⁶ <http://www.sibcolombia.net/web/sib/home>

¹¹⁷ <http://www.sibcolombia.net/web/sib/acerca-del-sib>

¹¹⁸ <https://code.google.com/p/gbif-providertoolkit/wiki/HowToContribute>

¹¹⁹ <http://lists.gbif.org/mailman/listinfo/ipt>

¹²⁰ <https://code.google.com/p/gbif-providertoolkit/wiki/IPT2ManualNotes?tm=6>

¹²¹ <https://code.google.com/p/gbif-providertoolkit/wiki/IPT2ManualNotes?wl=es>

2. The IPT lacks built-in data validation. Since the IPT is designed to run effectively on a common computer, validating extremely large data sets (+100 million records) becomes an impractical operation. GBIF has been working with its partners, however, to provide pluggable remote validation services on performant data architecture to fill this gap.
3. The IPT depends on server administrators to backup its data. There are plans to address this problem by adding long-term data storage and redundancy to the IPT this year.

Recommendations

Standards used: Darwin Core, Darwin Core Text Guidelines, Ecological Metadata Language.

Suggested improvements: enhance IPT for sample-based data sets.

Tool status

The IPT is currently used to publish occurrence data sets and checklists and associated metadata (or metadata documents alone). Work is underway to enhance it for publication of sample-based data. This requires developing a data model for sample-based data that is compatible with the DwC-A model. This will include a new core and extension and a modified instance of the IPT that recognises the new core/extension. A prototype IPT (**Fig. A1**) is already in place at <http://eubon-ipt.gbif.org> together with a few test sample data sets expressed using an early iteration of the sample data model. The latter is undergoing revision based on feedback from the EU BON partners.

Hosted resources available through this IPT

Logo	Name	Organisation	Type	Subtype	Records	Last modified	Last publication	Next publication
--	BMS-IL, Israel Butterflies Monitoring Scheme, by Israel Peer	Not registered	Occurrence	Observation	7,455	2014-04-12	2014-04-11	--
--	congoriver2010-limnology	Not registered	Occurrence	--	86	2014-04-16	2014-04-16	--
--	Diatoms of the Congo 2010 expedition	Not registered	Occurrence	--	0	2014-05-05	Not published	--
--	Lepidoptera observations by Hannu Saarenmaa	Not registered	Occurrence	Observation	7,289	2014-04-11	2014-04-11	--
--	Rhine Main Observatory Aquatic Invertebrates Biodiversity	Not registered	Occurrence	Observation	9	2014-04-15	2014-04-15	--

The most recently updated resources are also available as an RSS feed.

All data available in this installation is dedicated to the public domain. All content on this website is licensed under CC-BY

Figure A1. An instance of the IPT adapted for use with sample based data within EU BON.

A.2 GBIF Spreadsheet-Processor

Recognising that spreadsheets are a common data capture/management tool for biologists and that the Darwin Core terms lend themselves to representation in the tabular format of spreadsheets, three organisations, GBIF, EOL, and The Data Conservancy (DataONE project), collaborated to develop the GBIF Spreadsheet-Processor¹²², a web application that supports publication of biodiversity data to the GBIF network using pre-configured Microsoft Excel spreadsheet templates. Two main data types are supported: i) occurrence data as represented in natural history collections or species observational data and ii) simple species checklists.

The tool provides a simplified publishing solution, particularly in areas where web-based publication is hampered by low-bandwidth, irregular uptime, and inconsistent access. It enables the user to convert local files to a well-known international standard using an asynchronous web-based process. As illustrated in **Fig. A2**, the user selects the appropriate spreadsheet template, completes it and then emails it to the processing application which

¹²² <http://tools.gbif.org/spreadsheet-processor/>

returns the submitted data as a validated Darwin Core Archive, including EML metadata, ready for publishing to the GBIF or other network.

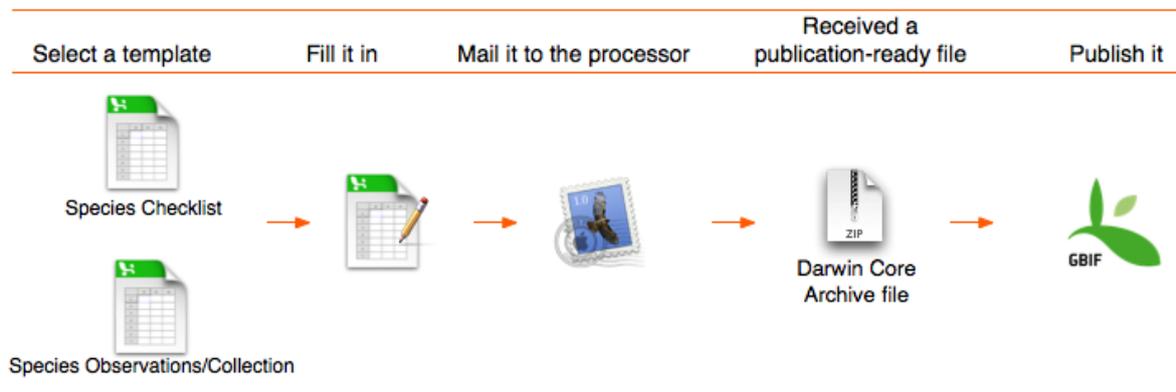


Figure A2. The web based processor ingests a spreadsheet and outputs a validated Darwin Core Archive.

Pros and Cons of the tool

The spreadsheet processor shares some of the pros & cons of the GBIF IPT above. Its chief advantage is its suitability for use in regions with low-bandwidth, irregular uptime, and inconsistent access.

A.3 Biodiversity Data Journal¹²³ and Pensoft Writing Tool¹²⁴

Main usage, purpose, selected examples

The Biodiversity Data Journal (BDJ) and associated Pensoft Writing Tool (PWT) represent together a next-generation, narrative (text) and data integrated publishing workflow, launched to mobilise, review, publish, store, disseminate, make interoperable, collate and re-use data through the act of scholarly publishing. All these processes are realised for the first time within a single, authoring, peer-review and publishing, online collaborative platform.

The Biodiversity Data Journal is a novel, community peer-reviewed, open-access journal, launched to accelerate mobilisation, dissemination and sharing of biodiversity-related data of any kind. All structural elements of the articles – text, descriptions, species occurrences, data tables, etc. – are treated, stored and downloaded as DATA in both human and machine-readable formats. The journal will publish papers on any taxon of any geological age from any part of the world with **no lower or upper** limit to manuscript size, for example:

- new taxa and nomenclatural acts
- data papers describing biodiversity-related databases;
- local or regional checklists and inventories;

¹²³ <http://biodiversitydatajournal.com>

¹²⁴ <http://pwt.pensoft.net>

- ecological and biological observations of species and communities;
- identification keys, from conventional dichotomous to multi-access interactive online keys;
- descriptions of biodiversity-related software tools.

The Pensoft Writing Tool is a manuscript authoring online collaborative platform. It is integrated with peer-review and editorial manager, publishing and dissemination tools, currently realised through the Biodiversity Data Journal. PWT can be integrated with any journal publishing platform that is able to accept XML-born manuscripts.

The Pensoft Writing Tool provides:

- Full life cycle of a manuscript, from writing through submission, revisions and re-submission within a single online collaborative platform;
- Conversion of Darwin Core¹²⁵ and other data files into text and vice versa, from text to data;
- Automated import of data-structured manuscripts generated in various platforms (Scratchpads¹²⁶, GBIF Integrated Publishing Toolkit (IPT)¹²⁷, authors' databases);
- A set of pre-defined, but flexible, Biological Codes and Darwin Core compliant, article templates;
- Easy online collaborative editing by co-authors and peers;
- A novel, community-based, pre-publication peer-review.

Examples of use of BDJ and PWT

During the first two months after its launch on 16th of September 2013, BDJ published some 50 articles (taxonomic, data papers, software descriptions, general research articles), including the landmark *Beyond dead trees: integrating the scientific process in the Biodiversity Data Journal*¹²⁸ and *Eupolybothrus cavernicolus* Komerički & Stoev sp. n. (Chilopoda: Lithobiomorpha: Lithobiidae): the first eukaryotic species description combining transcriptomic, DNA barcoding and micro-CT imaging data¹²⁹. The journal has already ca. 1500 users and this number increases daily.

Darwin Core Archives are generated automatically for all occurrence data and taxon treatments in each separate published paper. The DwC-A formats follow the standards used for harvesting by GBIF and Encyclopedia of Life (EOL)¹³⁰.

The journal accepts manuscripts generated by the Scratchpads Publication Module in XML format through the Pensoft Writing Tool, at the “click of a button”.

¹²⁵ <http://rs.tdwg.org/dwc/>

¹²⁶ <http://scratchpads.eu/>

¹²⁷ <http://ipt.pensoft.net/ipt/>

¹²⁸ <http://biodiversitydatajournal.com/articles.php?id=995>

¹²⁹ <http://biodiversitydatajournal.com/articles.php?id=1013>

¹³⁰ <http://eol.org/>

Pros and Cons of the tool

Pros:

- Integrated text (narrative) and data publication of two types of biodiversity data: (i) primary occurrence data (specimens, observations), (ii) Species checklists and taxonomies
- Occurrence data published in the different papers can be shared and collated together
- Can be used to publish in the form of “data papers” of any kind of biodiversity-related data.
- Data and content are archived in PubMedCentral after publication
- Small datasets are downloadable straight from the article text
- Standards-compliant publishing: export automatically taxon treatments and occurrence data into Darwin Core Archive (DwC-A) format, a compressed set of files based on the Darwin Core terms, and the GBIF metadata profile based on the Ecological Metadata Language standard
- Provides a publication venue for software and tools descriptions

Cons:

- Currently, the BDJ and PWT are constrained to be used mostly in the biodiversity domain.
- Data sharing tools can only be used for occurrence data sets and checklists.

Recommendations

Standards used: Darwin Core, Darwin Core Archive, Ecological Metadata Language.

Suggested improvements: enhance PWT and BDJ for traits data, and sample based Darwin Core compliant data sets. Use the technologies invented by BDJ to re-publish legacy literature (e.g., historical floras and faunas for example and mobilise data included in them).

Tool status

The PWT and BDJ can be used to publish biodiversity-related data and associated metadata.

A.4 Bibliography of Life

Main usage, purpose, selected examples

The Bibliography of Life¹³¹ platform was developed within the EU FP7 project ViBRANT and consists of three integral tools, RefBank¹³² and ReFindit¹³³ and Biosystematics

¹³¹ <http://biblife.org>

¹³² <http://refbank.org>

¹³³ <http://refindit.org>

Literature Repository based at ZENODO/CERN¹³⁴. Currently the platform is being maintained by Plazi and Pensoft.

While RefBank is the place to store, parse, edit, and download bibliographic references, ReFindit is designed to discover and download references from a wide range of open access online bibliographies, such as CrossRef, PubMed, Mendeley, Biodiversity Heritage Library (BHL), RefBank, Global Names Usage Bank (GNUB) and others (**Fig. A3**).

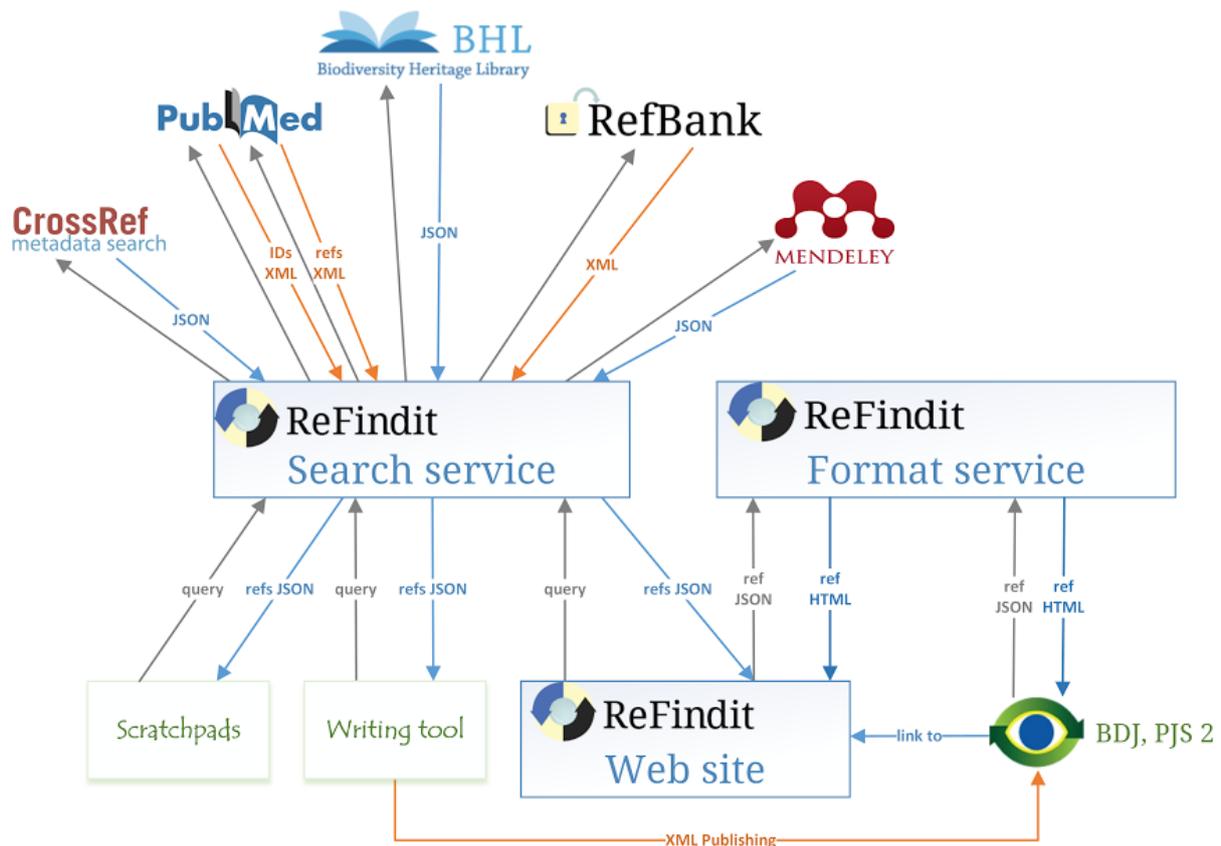


Figure A3. RefBank and ReFindit workflow .

RefBank is an open, coordinator-free network of independent nodes that replicate bibliographic references on each node, eliminating any single point of failure. This architecture further prevents any single entity from governing the data because everyone can set up a node and participate in the network with their own full copy of the whole data set. Pull-based replication prevents erroneous data from being actively pushed into the network. Contributing to RefBank is easy: everyone can upload individual bibliographic references or entire bibliographies. ReCAPTCHA protects the upload forms without the need for login or user accounts; API-based upload only requires a node-specific pass phrase. RefBank embraces near duplicate references, exploiting their inherent redundancy for automated reconciliation. The web interface further supports manual curation.

¹³⁴ <http://zenodo.org>

ReFindit provides an easy search function, based on a simple interface, which collates and sorts the results from the search engines for presentation to the user to read and with the option to refine the results presented or submit a new search. The searched references may be used for different purposes, e.g. conversion in some 600 citation styles and download in widely accepted bibliographic metadata standards. The tool is available through the Bibliography of Life as a standalone application at www.refindit.org, and is integrated as a search interface in Scratchpads, Pensoft Writing Tool (PWT)¹³⁵ and the Biodiversity Data Journal (BDJ)¹³⁶.

Pros and Cons of the tool

Pros

- Federated, open source infrastructure
- Community ownership of open data
- Service-oriented infrastructure with APIs available
- Unlimited number of style versions of a reference
- The ReFindit tool open to add new online databases for searching and browsing
- Services for handling of a bibliographic reference
- DOIs assigned to legacy publications stored at ZENODO.

Cons

- Currently, Biodiversity of Life is focusing mostly on the biodiversity domain, although technologically it is not constrained to that.
- The Bibliography of Life still lacks intensive promotional campaign to broad the scope and range of users.

Recommendations

Standards used: MODS, OAI-PMH

Suggested improvements: enhance Bibliography of Life to domains other than biodiversity through amendment of new searched platforms and harvesting mechanisms to enrich the content of RefBank.

Tool status

RefBank and ReFindit tool are fully operable. The Biosystematics Literature Repository is currently at beta testing stage.

¹³⁵ <http://pwt.pensoft.net>

¹³⁶ <http://biodiversitydatajournal.com>

A.5 Metacat: Metadata and Data Management Server

Main usage, purpose, selected examples

Metacat is a repository for data and metadata (descriptions of data) that helps scientists find, understand and effectively use the data sets they manage or those created by others. The information is available through the data packages, which consists of the data set associated with its corresponding metadata. Thousands of data sets are currently documented in a structured way and stored in Metacat systems, providing the scientific community with a broad range of science data that – because the data are consistently described – can be easily searched, compared, merged, or used in other ways¹³⁷.

Not only is the Metacat repository a reliable place to store metadata and data (the database is replicated over a secure connection so that every record is stored on multiple machines and no data are ever lost to technical failures), it provides a user-friendly interface for information entry and retrieval. Scientists can search the repository via the Web using a customisable search form. Searches return results based on user-specified criteria, such as desired geographic coverage, taxonomic coverage, and/or keywords that appear in places such as the data set's title or owner's name. Users need only to click on a linked search result to open the corresponding data-set documentation in a browser window and discover whom to contact to obtain the data themselves or how to immediately download the data via the Web¹. All the data packages can be provided with the proper data set usage rights to guarantee that proper recognition is given to the involved parties.

Metacat is a Java servlet application that runs on Linux, Mac OS, and Windows platforms in conjunction with a database, such as PostgreSQL (or Oracle), and a Web server. The Metacat application stores data in an XML format using Ecological Metadata Language (EML) or other metadata standards such as ISO 19139 or the FGDC Biological Data Profile¹³⁸.

Metacat is being used extensively throughout the world to manage heterogenic and complex environmental data. It is a key infrastructure component for the NCEAS data catalog, the Knowledge Network for Biocomplexity (KNB) data catalog, and for the DataONE system, among others¹. Metacat was adopted by the Brazilian Research Program in Biodiversity – PPBio in 2010 and currently stores data collected in 24 different field stations in Brazil. Currently there are more than 400 data packaged available to users in <https://ppbiodata.inpa.gov.br/metacatui/#data/page/0>. All the data from PPBio is curated and validated by a data manager.

The metadata stored in Metacat includes all of the information needed to understand what the described data are and how to use them: a descriptive data set title; an abstract; the temporal, spatial, and taxonomic coverage of the data; the data collection methods; distribution information; and contact information. Each information provider decides who has access to this information (the public, or just specified users), and whether or not to upload the data set

¹³⁷ information provided by Metacat Administrator's Guide: <http://knb.ecoinformatics.org/software/metacat>

¹³⁸ information provided by Metacat Administrator's Guide: <http://knb.ecoinformatics.org/software/metacat>

itself with the data documentation. Information providers can also edit the metadata or delete it from the repository, again using Metacat's straightforward Web interface¹.

Pros and Cons of the tool

Pros: Metacat's user-friendly Registry application allows data providers to enter data set documentation into Metacat using a Web form. When the form is submitted, Metacat compiles the provided documentation into the required format and saves it. Information providers need never work directly with the XML format in which the metadata are stored or with the database records themselves. In addition, the Metacat application can easily be extended to provide a customised data-entry interface that suits the particular requirements of each project. Metacat users can also choose to enter metadata using the Morpho application, which provides data entry wizards that guide information providers through the process of documenting each data set¹. A data center using Metacat can become DataONE member node with a relatively simple configuration.

The metadata stored in Metacat includes all of the information needed to understand what the described data are and how to use them: a descriptive data set title; an abstract; the temporal, spatial, and taxonomic coverage of the data; the data collection methods; distribution information; and contact information. Each information provider decides who has access to this information (the public, or just specified users), and whether or not to upload the data set itself with the data documentation. Information providers can also edit the metadata or delete it from the repository, again using Metacat's straightforward Web interface¹.

Cons: Flexibility that allows organising and preserving heterogeneous datasets comes together with the drawback that it is not possible to query the data tables directly. PPBio found that it was necessary to provide auxiliary tables (<http://ppbio.inpa.gov.br/repositorio/dados>) to allow sampling effort to be evaluated effectively in most situations.

Recommendations

Main context for use in to match the needs of EU-BON is as a repository for tabular data. If there are specific projects that deal with tabular data at a standardised perspective – spatial, temporal or taxonomic, it is recommended, based on PPBio experience, to build standardised data tables that will facilitate further integration. Additional development to extend the tool in order to provide a customised data-entry interface that suits the particular requirements of each project can be considered.

Tool status

This tool is ready to be used.

A.6 DataONE Generic Member Node

Main usage, purpose, selected examples

The DataONE Generic Member Node (GMN) is a python reference implementation of a complete (Tier 4) member node to DataONE. It can be freely downloaded from the

DataONE source code repository¹³⁹. The software is designed to be used from the command line and via REST API calls – there is no graphical user interface.

Pros and Cons of the tool

The GMN is a complete implementation of the DataONE member node stack in a language commonly used for a wide range of scientific purposes. This software is regularly updated and maintained by DataONE as part of their tools for testing during development. Lacking a GUI, however, the GMN is not appropriate for direct use by most scientists. It can, however, be an effective tool for constructing a data sharing site which is compatible with DataONE. Note, however, that Morpho (next section) can be used to package and upload data to either Metacat or to a GMN installation. As such, Morpho provides a data submission tool with ONEMercury providing a data search and delivery infrastructure.

Recommendations

Where an existing data repository wishes to become a DataONE member node, the GMN is a tool that can be used to adapt the repository's existing software. The GMN should be investigated as an option for standing up a data sharing environment for partners and national organisations supporting Work Packages 4 and 5, particularly for data that is not suitable for inclusion in GBIF.

Tool status

This tool is ready to be used.

A.7 DataONE “Slender Node”

Main usage, purpose, selected examples

The DataONE Slender Node software stack is designed to provide a lightweight means to create a Tier 1 (public read, no authentication) DataONE member node based on a collection of data and metadata files on a server file system. The software periodically crawls this file system, processes commonly understood metadata formats for links to the underlying data files, and constructs the necessary packages to expose this data via DataONE.

Pros and Cons of the tool

The Slender node is intended to be extremely easy to deploy and adding/updating of data is simply a matter of updating files on a file system. It does not provide any means for enabling authenticated access to data – it only supports public readable data and metadata.

Recommendations

¹³⁹ https://repository.dataone.org/software/cicore/trunk/mn/d1_mn_generic/

Depending on the timing of the software release and the timing of EU BON needs, this may be an option for enabling access to data from allied projects and smaller national data projects, as well as citizen science projects.

Tool status

This tool is in active development with release in mid-2014 expected.

A.8 Morpho Metadata Editor

Main usage, purpose, selected examples

Created for scientists, Morpho is a user-friendly application designed to facilitate the creation of metadata (information that describes your data) so that you and others can easily locate and determine the nature of a wide range of data sets. By specifying some basic information (a title and abstract, for example) about your data in a uniform, standardised way, you or any one you have granted permission to access your data will be able to find and view the data. When you create a metadata file that explains what your data represent and how they are organised, you are not only better able to manage the data, you help other scientists discover and understand them, too¹⁴⁰.

Morpho interfaces with the Knowledge Network for Biocomplexity (KNB) Metacat server. Once you have annotated your data with metadata, you can choose to upload your data—or just your data description (the metadata)—to the Metacat server, where they can be accessed from the web by selected colleagues or by the public if you so choose. Metadata are stored in a file that conforms to the Ecological Metadata Language (EML) specification. Data can be stored with the metadata in the same file. Morpho allows the user to create a local catalog of data and metadata that can be queried, edited and viewed¹.

Pros and Cons of the tool

Morpho is a user-friendly tool that allows researchers to easily create metadata, (i.e. describe their data in a standardised format), and create a catalog of data & metadata upon which to query, edit and view data collections. In addition, it also provides the means to access network servers - like the KNB Metacat server - in order to query, view and retrieve all relevant, public ecological data. Morpho has an advantage relate to the registry shipped within Metacat which is the Data Table description. Users need to install the tool in their local machines.

Recommendations

PPBio's experience shows that Morpho is a tool that allows ecological data curation, assuring that data tables are correctly built. Controlled vocabularies and standardised terms to describe field sites can be used to avoid ambiguity. Means to relate taxonomic coverage with DwC

¹⁴⁰ information provided by Morpho User Guide:

<https://knb.ecoinformatics.org/software/dist/MorphoUserGuide.pdf>

standard is desirable. Having Morpho wizard accessible through the web, without the need to have it installed in local machines would be desirable to implement within the context of EU BON.

Tool status

This tool is ready to be used.

A.9 GeoServer

Main usage, purpose, selected examples

GeoServer is an open source software server written in Java that allows users to share and edit geospatial data. Designed for interoperability, it publishes data from any major spatial data source using open standards. Being a community-driven project, GeoServer is developed, tested, and supported by a diverse group of individuals and organisations from around the world. GeoServer is the reference implementation of the Open Geospatial Consortium (OGC) Web Feature Service (WFS) and Web Coverage Service (WCS) standards, as well as a high performance certified compliant Web Map Service (WMS).

Pros and Cons of the tool

GeoServer enables the publishing of data using OGC web services, which is important for a variety of modeling and workflow applications. It has an active development community and has significant use in the ecological and environmental science community. GeoServer is not currently DataONE-enabled and there are no active plans for such development.

Recommendations

EU BON should investigate the level of use of GeoServer within the partner and allied organisations to understand the potential need for interoperability with this package and what EBV-relevant data may need to be exposed from relevant GeoServer repositories. It is likely that interoperability can be achieved through the OGC web services.

Tool status

This tool is ready to be used.

A.10 GeoNetwork

Main usage, purpose, selected examples

GeoNetwork¹⁴¹ is an open source software server written in Java and using LUCENE or SQL, that allows users to share and edit geospatial metadata and to link them to on maps that are available on line in a search interface. It is designed for interoperability. Metadata are based on the ISO 19 115 and ISO 19 139 metadata profile. It is interoperable with any maps server provided in the WMS (Web Map Server) and CSW (Catalogue Service for the Web) formats. It is also compliant with the Z39.50 and OAI-PMH protocols (to synchronise the

¹⁴¹ <http://geonetwork-opensource.org/>

replication of metadata coming from external sources), and with GeoRSS to publish information as well as with the GEMET (General Multilingual Environmental) thesaurus.

Being a community-driven project, GeoNetwork is developed, tested, and supported by a diverse group of individuals and organisations from around the world. It also feature a lot of input from the FAO and the community of institutions working with INSPIRE data. GeoNetwork complete WMS server by creating of catalogue of maps and documents dealing with spatial information searchable by keyword

Pros and Cons of the tool

Good integration with WMS servers, in particular GeoNetwork. Using GeoNetwork would allow a good interoperability with ISO, OGC and INSPIRE standards. It allows linking together metadata, data, maps and thesaurus. Open Source, but used by major institution (Food and Agriculture Organization of the United Nations (FAO)¹⁴² initiator of the project) and projects (OneGeology¹⁴³).

Recommendations

We would recommend to test GeoNetwork and evaluate the released versions, as it is one of the most advance GIS available in the market in term of compliance with the OGC and INSPIRE standards. Most of the projects related to INSPIRE ad OGC use it for their reference implementation of the standards. This tool can act as an intermediate layer to allow other tools publishing maps (WMS, WFS, like the above mentioned GeoServer) to be compliant with INSPIRE and to link their data and metadata with thesauri. It can be part of a public portal gathering and publishing data from one or several projects, with full text and geographical search engine. The mailing list of GeoNetwork is also very active, the community being placed at an intermediate cross-road position between the technical aspects of GIS, the scientific issues and the issue related to data management policies at nation and regional level, EU BON could benefit from following and intervening in those discussion.

A.11 Data Access Protocol-compliant servers

Main usage, purpose, selected examples

The Data Access Protocol (DAP¹⁴⁴) is a REST web service based protocol designed for science data. There are multiple software packages which implement DAP, with OPeNDAP Hyrax¹⁴⁵ and THREDDS¹⁴⁶ being the most widely deployed. THREDDS and OPeNDAP provide tools for enabling access to data in a variety of formats, including netCDF, HDF, HDF-EOS, and GRIB. These formats are more widely used in the climate and ecological forecasting communities than for species occurrence, though netCDF is seeing increased use

¹⁴² <http://www.fao.org/home/en/>

¹⁴³ <http://www.onegeology.org/>

¹⁴⁴ http://www.opendap.org/pdf/dap_2_data_model.pdf

¹⁴⁵ <http://www.opendap.org/>

¹⁴⁶ <https://www.unidata.ucar.edu/software/thredds/current/tds/>

by groups that create gridded output of species occurrence. These formats and server tools are also relevant to gridded habitat data.

Pros and Cons of the tool

DAP-compliant servers are highly relevant to modellers and are an efficient way to expose gridded data, with sub setting and time-slicing capabilities. There is current development to make Hyrax and THREDDS DataONE-enabled.

Recommendations

Where gridded data are to be used in the development of EBVs or as a gridded data product derived from species observation data, DAP-compliant servers may be an appropriate choice, particularly where making this data available to the modelling communities is concerned.

Tool status

These tools are available and ready for use.

A.12 DiGIR

Main usage, purpose, selected examples

Distributed Generic Information Retrieval (DiGIR) is a protocol developed by the biodiversity informatics community in 2000-2002. First deployed in MaNIS and VertNet, its purpose is to implement queries to distributed data providers. It is modelled after the Z39.50 protocol, which was used in the REMIB network – one of the first data sharing networks of the biodiversity community. When GBIF started operations in 2002, it adopted DiGIR and BioCASE as the interoperability mechanisms. Today, DiGIR is being replaced by other mechanisms, but is still in wide use.

Unlike Z39.50, DiGIR is XML-based, which was the main reason to develop it. The DiGIR protocol supports several operations such as inventory of information resources on a provider, download to resource metadata, and queries to the full data. The latter is restricted to Darwin Core.

There are several DiGIR implementations in different languages, such as PHP, Java, Python, and Microsoft .net. These are basically software wrappers for SQL databases. The GBIF Data Repository Tool is a Zope-based tool that supports upload and download of CSV documents from a hierarchical folder structure with Dublin Core metadata, and bundles the Python DiGIR provider. The tool is now discontinued, but served as a prototype for the IPT.

Pros and Cons of the tool

DiGIR offers a simple way to query remote databases. It also has simple metadata, and a DiGIR provider can describe its resources. Although the DiGIR protocol was deployed widely, it was never standardised by TDWG. Resource metadata are very basic and non-standard. Queries are restricted to Darwin Core. There is no harvesting mechanism for entire resources.

Recommendations

Phase out. Use TAPIR instead where distributed queries are needed.

Tool status

The PHP reference implementation is still available, see <http://digir.sourceforge.net/>.

A.13 TAPIRlink

Main usage, purpose, selected examples

TAPIR - TDWG Access Protocol for Information Retrieval, was developed in 2005-2008 as the successor of DiGIR. Its purpose was to unify the DiGIR and BioCAsE protocols and make the protocol independent of certain schemas. Otherwise TAPIR follows the same ideas as DiGIR. TAPIR became a TDWG standard in 2008, see <http://www.tdwg.org/activities/tapir/>.

Pros and Cons of the tool

TAPIR offers a simple way to query remote databases. Its resource metadata are more elaborate than DiGIR, but still non-standard. TAPIR providers cannot describe their resources, which is a setback from DiGIR. TAPIR has not been deployed widely. There is no harvesting mechanism for entire resources.

Recommendations

A TAPIR wrapper might be a good choice in front of large databases which must be queried, and not harvested. Capability of describing resources could be added to the protocol. EML-based metadata could be added, or replace the current resource metadata specification.

Tool status

TAPIRlink is the PHP reference implementation of the protocol, see <http://sourceforge.net/projects/digir/files/TapirLink/>.

A.14 BioCAsE

Main usage, purpose, selected examples

The Biological Collection Access Service, BioCAsE, is a transnational network of biological collections of all kinds. BioCAsE enables widespread unified access to distributed and heterogeneous European collection and observational databases using open-source, system-independent software and open data standards and protocols¹⁴⁷.

An important component of the BioCAsE infrastructure is the BioCAsE Provider Software (BPS), an xml data binding middleware, which is used as an abstraction layer in front of a database. After local configuration the database is accessible as a BioCAsE service - as defined by the BioCAsE protocol - and can be used to create distributed heterogeneous

¹⁴⁷ http://www.biocase.org/whats_biocase/unit_net.shtml

information systems. The BPS is agnostic to the kind of data being exchanged and any conceptual schema, such as ABCD (Access to Biological Collection Data)¹⁴⁸ for the BioCASE network¹⁴⁹, can be used to set up distributed networks.

In its latest Version, the BioCASE provider software provides a function for exporting data sets into ABCD-Archives so that portals can harvest entire databases without the need for visiting individual records.

Apart from its role as a data publishing tool in BioCASE and GBIF, the BPS is used in several Special Interest Networks such as the Global Genome Biodiversity Network (GGBN)¹⁵⁰, the Australian Virtual Herbarium (AVH)¹⁵¹, and GeoCASE¹⁵².

Pros and Cons of the tool

The BPS is based on stable data definitions and protocol specifications. The software itself is successfully used in more than 10 international index and actively supported by the BioCASE helpdesk). One of the outstanding capabilities is the ability to serve both access to full data sets and individual records via the same installation. However, compilation of very large datasets (> 1 million records) can be time consuming and needs improvement.

Recommendations

Collection and observational data not yet available to biodiversity informatics infrastructures such as EU BON could be exposed via the BPS tool. The standardized BPS interfaces ensure that the data will be understood in different contexts and become useful for a wide scientific audience.

Tool status

The BPS is actively maintained and developed by the Informatics research Group of the Botanic Garden and Botanical Museum Berlin-Dahlem¹⁵³. With more than 100 installations worldwide it has a broad user-base. New versions and the documentation can be downloaded from http://www.biocase.org/products/provider_software/index.shtml.

A.15 Scratchpads

Main usage, purpose, selected examples

Scratchpads¹⁵⁴ are virtual research environments — a web-based content management software (based on Drupal) which facilitates the organisation and publication of biodiversity data. The focus lies on the mobilisation, structuring, linking and dissemination of taxon-centric information, although the software can be used for any other type of web publishing

¹⁴⁸ <http://www.tdwg.org/standards/115/>

¹⁴⁹ http://www.biocase.org/whats_biocase/unit_net.shtml

¹⁵⁰ http://www.ggbn.org/ggbn_portal/

¹⁵¹ <http://avh.chah.org.au/>

¹⁵² <http://www.geocase.eu/>

¹⁵³ <http://www.bgbm.org/en/biodiversity-informatics>

¹⁵⁴ <http://scratchpads.eu>

(e.g. to create project websites, literature databases, etc.). Data are organised into different types of information — e.g. images, videos, specimen information, literature, species descriptions, occurrences, etc. — and are organised around a biological classification. Each piece of information can be tagged with a taxon name, and thus the information can be browsed either by navigating the biological classification or by searching for the taxon name. All information pertaining to a taxon is then displayed on so-called “taxon pages”. It is also possible to integrate information from other sources (e.g. EOL, GBIF, NCBI, Google Scholar, BHL...) into the system, many APIs are already available and can be activated with a single click. The system is easy to use and for the average user no special technical knowledge is required. Its communal design allows groups of researchers to use the system simultaneously, to collaboratively work on a project and to share data, either publicly or privately within virtual research groups. Where applicable, data can be exported as Darwin Core Archives. Scratchpads are maintained and hosted by the Natural History Museum in London and users can simply apply for a Scratchpads hosted on the Museum's servers, alternatively, the source code is available for download via a git repository.

Pros and Cons of the tool

Scratchpads provide a very easy tool to organise, publish and share taxon-centric information. There is an extensive documentation on the website and regular training courses are organised. No special technical knowledge is required to use the software. Hosting can either be provided by the NHM London or the software can be downloaded and hosted locally. Data can be exported as standard-conform DarwinCore Archives, facilitating information sharing with other databases and systems using DarwinCore. If hosted by the museum, users have restricted rights, so the possibilities of customising the software are limited. If downloaded, some technical knowledge is required, but then the software offers almost unlimited possibilities for modification for own purposes.

Recommendations

Scratchpads are targeted towards managing and sharing small pieces of data pertaining to taxa / biodiversity. They are not intended towards sharing huge occurrence records files or for metadata management of datasets. However, the system does have batch import functions and can read *.csv files of classifications, bibliographies, taxon descriptions, etc. and readily integrate them into the system. Collaboration with peers is made very easy through the system, allowing groups of researchers to contribute and share information among each other or with the public.

A.16 PlutoF

Main usage, purpose, selected examples

The PlutoF cloud¹⁵⁵ provides online service to create, manage, share, analyse, and mobilise biodiversity data. Data types cover ecology, taxonomy, metagenomics, nature conservation, natural history collections, etc. Common platform aims to grant the databases with

¹⁵⁵ <http://plutof.ut.ee>

professional architecture, sustainable developing and persistence. It provides synergy through common modules for the classifications, taxon names, analytical tools, etc. Common taxonomy module is based on available sources (e.g. Fauna Europaea, Index Fungorum) and may be developed collectively further by the users. Currently there are more than 1500 users who develop their private and institutional databases or use analytical tools for biodiversity data. PlutoF cloud also provides data curation, possibilities, including third party annotations to the data from external resources, such as genetic data from GenBank¹⁵⁶. PlutoF is developed by the IT team of Natural History Museum (University of Tartu, Estonia).

Curated datasets hosted by PlutoF cloud can be made available through public web portals. Examples include the UNITE community which curate DNA based fungal species and provide open access to their datasets through UNITE portal¹⁵⁷. Another example is eBiodiversity portal¹⁵⁸ that includes taxonomical, ecological and genetics information on species found in Estonia. Any public dataset in PlutoF cloud that includes information on taxa found in Estonia will be automatically displayed in this portal. This enables to discover biodiversity information for Estonia in one portal.

Pros and Cons of the tool

The web workbench allows to manage all personal biodiversity data (including private or locked data) in one place and share them with selected users. It is also possible to manage and analyse your own, institutional or workgroup data at the same time. Datasets on any taxon in any location can be created and stored in the system.

Recommendations

PlutoF cloud can be utilised by the EU BON project as one possible platform where Citizen Scientists can create, manage and share their biodiversity datasets.

Tool status

Web based service is available for all the individual users, workgroups and institutions. New infrastructure based on different technologies is under development and its beta version will be available in autumn 2014. Platform is developed by the team of eight IT workers.

A.17 DSpace

Main usage, purpose, selected examples

DSpace is an open source digital object management system, useful for managing arbitrary digital objects, such as data files. As distinct from Fedora Commons (managed by the same organisation – DuraSpace), DSpace comes with a usable user interface and is relatively usable “out of the box”. A wide range of institutions have implemented institutional repositories using DSpace. The Dryad Data Project (see next chapter) is based upon DSpace as a platform.

¹⁵⁶ <http://www.ncbi.nlm.nih.gov/genbank/>

¹⁵⁷ <http://unite.ut.ee>

¹⁵⁸ <http://elurikkus.ut.ee>

Pros and Cons of the tool

DSpace is a fairly complex tool with a broad range of capabilities. There is current work to DataONE-enable DSpace.

Recommendations

EU BON should investigate the level of use of DSpace (and Fedora Commons) within the partner and allied organisations to understand the potential need for interoperability with this package and what EBV-relevant data may need to be exposed from relevant repositories.

Tool status

The tool is available and ready for use, although a major rewrite is in progress as of this writing.

A.18 Dryad Digital Repository

Main usage, purpose, selected examples

The ‘Dryad Digital Repository’ is a curated resource providing a general-purpose location for a wide diversity of data types. Dryad's mission is to make the data underlying scholarly publications discoverable, accessible, understandable, freely reusable, and citable for all users. Dryad originated from an initiative among a group of leading journals and scientific societies in evolutionary biology and ecology to adopt a joint data archiving policy for their publications. Dryad is governed by a non-profit membership organisation. Membership is open to any stakeholder organisation, including but not limited to journals, scientific societies, publishers, research institutions, libraries, and funding organisations¹⁵⁹.

Pros and Cons of the tool

The data hosted by Dryad have been dedicated to the public domain under the terms of Creative Commons Zero (CC0) license, in order to minimise legal barriers and maximise the impact on research and education, the terms of reuse are explicit and have some important advantages¹⁶⁰:

- **Interoperability:** Since CC0 is both human and machine-readable, other people and indexing services will automatically be able to determine the terms of use.
- **Universality:** CC0 is a single mechanism that is both global and universal, covering all data and all countries. It is also widely recognised.
- **Simplicity:** There is no need for humans to make, or respond to, individual data requests, and no need for click-through agreements. This allows more scientists to spend their time doing science.

Dryad is based on the DSpace repository software with built-in internationalisation (i18n), automatically translating DSpace text based on the default language of the web browser. The

¹⁵⁹ <http://datadryad.org/pages/organization>

¹⁶⁰ <http://datadryad.org/pages/faq>

Dryad Repository does not impose any file format restrictions. As a result, Dryad cannot guarantee that all files in all data packages are accessible.

Dryad complies with Section 508 of the Rehabilitation Act of 1973. This is a United States federal law, while also being recognised as an international best practice. The Dryad website uses HTML by Section 508 standards and accessibility testing tools to ensure issues are found and fixed when new content features are added¹.

A full overview of integrated journals and costs for submission is provided here: <http://datadryad.org/pages/integratedJournals>

Recommendations

Dryad hosts research data underlying scientific and medical publications. Most data are associated with peer-reviewed journal articles, but data associated with non-peer reviewed publications from other reputable sources (such as dissertations) is also accepted. At this time, all Dryad submissions must be in English. Most types of files can be submitted (e.g., text, spreadsheets, video, photographs, software code) including compressed archives of multiple files. Ordinarily, no more than 10 GB of material are submitted for a single publication; larger data sets are accepted but will be subject to additional charges².

Tool status

This tool is ready to be used.

A.19 Species Observation System

Main usage, purpose, selected examples

Species Observation System¹⁶¹, is a web-based, freely accessible reporting system and data repository for species observations, used by citizen scientists, scientists, governmental agencies and county administrations in Sweden and Norway. The system handles reports of geo-referenced species observations of almost all major organism groups from all environments, including terrestrial, freshwater and marine habitats.

Species Observation System has an increasingly growth since its launch in year 2000 in Sweden, year 2008 in Norway and currently holds more than 40 million recorded observations in Sweden and 10,5 million in Norway (May 2014), including totally almost 1 million species documentation pictures. Thus, Species Observation System is by no comparison the largest data provider for biodiversity and conservation related science in Sweden and Norway. All data (except detailed location on a few sensitive species) is freely available in GBIF. The portals has about 600 000 unique visitors every year – in two countries with totally 14,5 million inhabitants.

The first generation of Species Observation System was launched in Sweden in year 2000, developed and hosted by the Swedish Species Information Centre at the Swedish University of Agricultural Sciences SLU. The Norwegian version was launched in 2008, adapted and

¹⁶¹ www.artportalen.se and www.artsobservasjoner.no

hosted by the Norwegian Biodiversity Information Centre (NBIC). The two organisations have developed and are managing this citizen science system in close cooperation with national biodiversity NGOs.

Pros and Cons of the tool

The tool is very efficient and due to the fact that the user friendliness and rich functionality encourages citizen scientist to use the system as their personal digital field diary. No anonymous sightings are allowed, and the user interface promotes extensive informal and voluntary quality control and annotation. Formal validation by about 300 expert users on important species is performed currently to achieve high data quality. A crucial feature of Species Observation System is that all data are openly shared in the society nationally and internationally.

The system is large and demanding (organisational foundation, ICT-competence/capacity, technical infrastructure and financial) to implement, manage, maintain and support.

Recommendations

Species Observation Service is considered as a major potential tool for broader European citizen science involvement in species mapping, surveillance and monitoring. In European countries or regions lacking efficient and open data species reporting systems, Species Observation System is recommended for European institutions, agencies and organisations to consider the system with the purpose of filling such tool gaps.

Tool status

Currently the Swedish Species Information Centre and the Norwegian Biodiversity Information Centre, together with environmental agencies in Sweden and Norway, are developing a common new version based on cutting edge technology. An optional English user interface is included. This version is partly launched in Sweden and a full version with reporting on all species groups will be launched in both countries at the end of the year 2014. During 2015 reporting apps for mobile devices will be available.

The system owners have not yet decided on conditions for sharing the system with other countries, the process will not start and decisions not taken before the new version is launched.

A.20 DEIMS: Drupal Ecological Information Management System

Main usage, purpose, selected examples

The International Ecological Information Management System (DEIMS)¹⁶² is a Drupal open-source, collaborative platform, that provides a web interface for scientists and researchers' networks, projects and initiatives with a metadata management and data sharing system. This system has been developed for and is particularly used within the Long-term ecological

¹⁶² <https://drupal.org/project/deims>

research (LTER)¹⁶³ domain, which aims at detecting environmental change and the associated drivers.

DEIMS is currently composed by the following components:

(a) the metadata editor, a web-based client interface to enter, store and manage metadata of three types of information sources: datasets, persons and research sites. Therefore, this editor provides the following interfaces: (i) dataset metadata editor, which provides entry forms for authorised users to create metadata description in compliance with the EnvEurope¹⁶⁴ (LTER-Europe¹⁶⁵) / ExpeERMetadata Specification for Dataset Level, based on EML (Ecological Metadata Language); (ii) site information metadata editor, which again allows authorised users to create metadata description for sites in the ILTER, ExpeER¹⁶⁶, and GEO BON networks; (iii) personnel database metadata editor for the creation or editing of the information, relevant to the scientists' contact details and research expertise;

(b) Discovery: allows multiple search profiles for all of the above types of information sources, as well as from external resources that are based on several search patterns, ranging from simple full text search and glossary browsing to categorised faceted search;

(c) Geoview (EnvEurope project), is a mapping component that provides a data portrayal on a map and view attributes of individual features (research sites, data sets) and portrays boundaries and centroids of the research sites, which are provided as Web Map Service (OGC-WMS) layers. These layers are directly linked to both Metadata editor and Discovery components so that the relevant metadata to be created and subsequently used for discovery.

Pros and Cons of the tool

The sharing of the dataset metadata collected by the DEIMS is implemented in two ways:

(a) periodic harvesting of metadata records according to the EML (Ecological Metadata Language) schema by Metacat. This is further used in order to create a data catalogue, which can in turn, be used by international or European initiatives (e.g. DataOne, GBIF) and projects (e.g. LifeWatch);

(b) periodic harvesting of metadata into the GeoNetwork catalogue, thus providing a catalogue service for web (OGC-CSW). The latter can be called for metadata collection by remote SDI catalogues, e.g. by the INSPIRE Geoportal.

The major advantage of the platform is its capacity to bridge the ecological domain with other global, European or national environmental geospatial information infrastructures as the INSPIRE, SEIS, GEOSS, through the transformation of the EML metadata to ISO/INSPIRE, and to provide the implementation facility for the CSW.

Recommendations

¹⁶³ <http://www.lternet.edu/>

¹⁶⁴ <http://www.enveurope.eu/>

¹⁶⁵ <http://www.lter-europe.net/>

¹⁶⁶ <http://www.expeeronline.eu/>

Although the original DEIMS started in 2008, based in Drupal 6, with UMBS, a handful of LTER sites, and Oak Ridge National Lab, it is only recently that the LTER network started to develop its current version (March 2013). Therefore, the platform is new and awaits the users to identify potential problems or obstacles but also directions for its potential development and expansion. Currently, DEIMS offers better and more metadata and data services using an adaptive/responsive interface.

Tool status

Among the projects which currently use DEIMS, the following are included: (a) International Long Term Ecosystem Research (ILTER) network; (b) LTER – Europe; (c) EnvEurope project; (d) EnvEurope.

This tool is ready to be used.

A.21 Plazi Taxonomic Treatment Server

Main usage, purpose, selected examples

Plazi's Taxonomic Treatment Server¹⁶⁷ provides access to the treatments of taxa. Each taxonomic usage is accompanied minimally by a text that describes the taxon or at least offers some further references, and thus defines the concept in a scientist's mind. There are millions of treatments in the scientific literature, which form an extremely valuable source of information. These treatments are increasingly linked to their underlying data, such as observation data, keys for identifications or other digital objects. There are two bottlenecks to providing semantically useful modern internet access. The first is that a huge number are not even digitally available, or at most are parts of semantically unstructured PDF-formatted documents. The second is that a substantial amount of the literature is only accessible through a paywall or comes with restrictions on their use. With the increasing wealth of digitised observation records, upon which most of the publications are based, it becomes imperative to provide access to the treatments, to link to them, and to enhance them with links to the material referenced in them.

The treatment repository fulfills this niche. It offers with GoldenGate¹⁶⁸ and respective XML schemas (TaxonX¹⁶⁹, TaxPub¹⁷⁰) tools to convert unstructured text into semantically enhanced documents with an emphasis on taxonomic data like treatments, scientific names, materials observation, traits or bibliographic references. It provides a platform that can store, annotate, access and distribute treatments and the data objects within. The Plazi approach also allows the legal extraction of uncopyrightable content from copyrighted material¹⁷¹.

The repository also can store annotations of literature to provide links to external resources, such as specimens, related DNA samples on GenBank, or literature. Annotation can be done

¹⁶⁷ <http://plazi.org>

¹⁶⁸ <http://plazi.org/?q=GoldenGATE>

¹⁶⁹ <http://plazi.org/?q=taxonx>

¹⁷⁰ <https://github.com/tcatapano/TaxPub/releases>

¹⁷¹ Agosti, D., W. Egloff. 2009. Taxonomic information exchange and copyright: the Plazi approach. BMC Research Notes 2009, 2:53 (<http://www.biomedcentral.com/1756-0500/2/53/abstract>)

at any level of granularity, from a materials citation to detailed tagging of specimens, provision of details of the collectors, or provision of morphological descriptions even to the tagging of individual traits and their states.

The use of persistent resolvable Identifiers allows smf option provision of RDF supports machine harvest and logical analysis data, within and between taxa.

The treatment server provides its content to aggregators or other consuming external applications and human users, including entire treatments to the Encyclopedia of Life¹⁷², and observation records to GBIF¹⁷³ using Darwin Core Archives. The latter will also be a base to harvest data for EU BON's modelling activities.

Pros and Cons of the tool

Pros

1. The Plazi Treatment Server is a one of its kind. With the US ETF¹⁷⁴ project, there is one complementary workflow known that focuses on traits, that collaborates with Plazi. The Plazi Treatment Server is built and maintained by highly skilled personnel, it is growing through regular input from Pensoft, whose treatments it stores. It is part of Plazi 1 Million Treatment project to establish open access to the content of taxonomic publications by developing various tools to convert new treatments.
2. The Plazi Taxonomic Treatment Server is complemented by activities regarding legal status of treatments and other scientific facts, semantic developments, especially linking to external vocabularies and resources, and use by a number of high profile operations (GBIF, EOL, EU BON, Pro-iBiosphere¹⁷⁵, domain specific web sites)
3. Currently 34000 treatments from 2700 documents are available.
4. New technical requests can be met quickly, and Plazi has in recent years been on the forefront to build interfaces to import data into GBIF and EOL.
5. Plazi uses RefBank¹⁷⁶ as a reference system for bibliographic references and is working in close collaboration with Zenodo (Biosystematics Literature Community, BLC)¹⁷⁷ to build a repository for articles that are not accessible in digital form. To discover bibliographic references, Refindit¹⁷⁸ is used and developed.

Cons

1. The Plazi Treatment Server is not yet full industrial strength and will need in its next phase to assess how to move from a research site to a service site.

¹⁷² <http://eol.org>

¹⁷³ <http://gbif.org>

¹⁷⁴ http://biowikifarm.net/v-botknow-test/web/About_BKP

¹⁷⁵ <http://www.pro-ibiosphere.eu/>

¹⁷⁶ <http://refbank.org/>

¹⁷⁷ <https://zenodo.org/collection/user-biosyslit>

¹⁷⁸ <http://refindit.org/>

2. GoldenGate, the Treatment Server's central tool is powerful, but a more intuitive human-machine interface needs be developed. Trait extraction needs further development.
3. The project is underfunded and staffed.

Recommendations

The project needs to invest in human-machine interfaces, documentation and training, and tools that allow the easiest possible way to annotate the treatments.

Specific services, such as bibliographic name provision and materials examined parsing need to become standalone applications.

Trait extraction needs be developed.

The Plazi Treatment Repository should become part of the IT infrastructure.

In the short term, it is important to build a critical corpus of domain specific treatments to allow scientifically meaningful data mining and extraction. This may require extensive data be gathered from treatment authors.

Develop a set of use cases to insure that the service requirements are complete.

Develop collaborations with treatment service projects outside the EU.

Tool status

This tool is ready to be used

A.22 Spreadsheet tools

Main usage, purpose, selected examples

Microsoft Excel is a software package, included in the Microsoft Office Suite that enables the creation of spreadsheets or forms, provides simple data comparison and analysis tools, and creates graphs. Data are captured in workbooks, which can be composed of a single or several sheets. Simple sort and filtering tools allow data to be queried. QA/QC can be performed using built-in tools that can find values and replace them with other values, remove duplicates, find missing values, characterise column data types, etc. Built-in or user-defined formulas can be used for calculations or transformations. Excel can also utilise Visual Basic for Applications (VBA) or .NET framework programming. Excel can also be used to create tables and visualisations. Other objects, such as photos and other images, text boxes, and clip art can be inserted into a spreadsheet.

Pros and Cons of the tool

Microsoft Excel is extremely widely used and it is possible to construct best practices that improve the reusability and machine processability of data stored and analysed using Excel. Such practices include having a single table per sheet, putting graphs on separate sheets from the data tables, and using named cells and ranges in formulas. However, those practices are not well known and are rarely followed. Complex formulas using cell references can be

extremely difficult for data generators to document and data consumers to comprehend. There are some known inaccuracies in statistical functions for data with larger dynamic ranges¹⁷⁹. Excel is a proprietary tool, and users in economically disadvantaged areas may not be able to afford a copy. Excel formatted files are generally not considered archive stable, but conversion to archive stable formats may result in loss of information. Open Source tools (e.g. Libre Office) are available and can read at least most Excel files, though there is occasional loss of fidelity. By itself, Excel has minimal capabilities for creating and managing metadata, and users almost never accurately populate those document properties.

By itself, Microsoft Excel is limited for data sharing. Groups often use Excel as a data storage and data analysis tool, and then rely on other tools to share these files. Examples include ftp sites, content management system (e.g. Drupal or SharePoint), file synchronisation tools (e.g. Dropbox), and simply sending files as email attachments.

GBIF has a spreadsheet processor¹⁸⁰ which provides a means to create structured output in formats which are suitable for publishing species occurrence data into GBIF.

The California Digital Library (CDL), in collaboration with Microsoft Research and DataONE, has created DataUP¹⁸¹ which allows Excel users to document data in Excel (including at least populating standard Dublin Core metadata fields and checking Excel documents for compliance with best practices). DataUP works as an ActiveX add-in for Excel on Windows and is available as a web site for all Excel users. DataUP can also upload data to the ONEShare member node of DataONE. In principle, a version of DataUP can be created which enables upload to another data repository which implements the DataONE Tier 3 (authenticated write) member node API.

Recommendations

Microsoft Excel is an extremely broadly used tool and relevant data will certainly be in Excel. EU BON should work with other relevant projects to help advance the use of best practices for data in Excel as well as advancing the education of other options for data analysis tools. EU BON should work with projects and test sites to ensure that species occurrence data in Excel is structured in ways that are compatible with the GBIF spreadsheet processor. Within this context EU BON should investigate ways to help ensure consistency in Darwin Core field usage to maximise the discoverability and semantic interoperability of GBIF-relevant data.

Tool status

The tool is available and ready for use.

¹⁷⁹ http://en.wikipedia.org/wiki/Numeric_precision_in_Microsoft_Excel

¹⁸⁰ <http://tools.gbif.org/spreadsheet-processor/>

¹⁸¹ <http://dataup.cdlib.org/>

A.23 Database packages

Main usage, purpose, selected examples

There are multiple database packages that are used for the organisation, analysis, and sharing of data, particularly data which is more complex than can be handled by typical spreadsheets and by projects which expect to share data. Examples include commercial software, such as Microsoft Access, Microsoft SQL Server, and Oracle, and open source tools such as MySQL, PostgreSQL, and SQLite. So-called “no SQL” databases are also relevant, such as MongoDB and CouchDB, as are data frameworks designed for large data, such as Hadoop and BigTable. PostgreSQL merits specific mention and relevance to EU BON as an open-source database with strong geospatial data management and analysis capabilities through the PostGIS package.

By themselves, databases have limited ability to share data. Exposing a database directly to the Internet (e.g. allowing inbound port 3306 to MySQL) is ill-advised due to security concerns. As such, some type of interface is needed to validate incoming data and commands. Ideally, that interface should also expose the data to people (e.g. a graphical user interface) and computer software (an application programming interface).

Pros and Cons of the tools

Database packages can be an important part of good data management practices. They can provide important methods for validation of data, automatic computation, and the normalisation of data is a best practice. Database transactions are a key tool for ensuring consistency of data during complex update operations. Care must be taken in the development of the underlying data model, as the data collected by a research project often evolves over time. As noted above, a database by itself is likely not sufficient as a data sharing tool, though automated tools do exist for providing at least read-only REST interfaces for reading data from a broad range of databases.

A key question in the use of databases for management of data, as opposed to file-based data management, is the definition of the atomic unit of data or the least addressable unit of data. Put it on another way, when files are used to manage and share data, each file can be given a unique identifier and each file can be addressed individually. Where databases are used, a broad range of choices are available. For GBIF, the observation is the atomic unit of data and each observation can be given a unique identifier. For a field site recording meteorological conditions, the data for one site for one day may be a natural choice for the atomic unit of data.

Recommendations

GBIF is exploring the use of Hadoop, in particular, and the ways which this could be enabled as a means to provide some of the data manipulation and extraction services needed to expand the applicability and usability of GBIF data. In general, EU BON should encourage the use of open source database tools. EU BON should consider the use of test sites and test packages using databases as means to demonstrate best practices.

Tool status

These tools are available and ready for use.

A.24 Tools to share molecular data

Sanger sequences:

European Nucleotide Archive (ENA)¹⁸² – captures and presents information relating to experimental workflows that are based around nucleotide sequencing. ENA forms part of the International Nucleotide Sequence Database Collaboration (INSDC)¹⁸³ and exchanges data between the collaboration partners (NCBI¹⁸⁴, DDBJ¹⁸⁵). INSDC forms the most comprehensive database for all molecular data types and linked metadata.

The Barcode of Life Data Systems (BOLD)¹⁸⁶ - designed to support the generation and application of DNA barcode data. Accepts new submissions (incl. submission of primary specimen data, images, trace files, and nucleotide sequences) and provides tools for third-party annotations to DNA barcodes by tagging and commenting options.

UNITE/PlutoF¹⁸⁷ – an online resource for regularly updated, quality checked and annotated ribosomal DNA sequence data for kingdom Fungi. UNITE keeps a local copy of INSD fungal rDNA sequences and provides tools for third-party annotations. UNITE also accepts new submissions and makes data available for browsing, blasting, and downloading on public homepage and identification tools. UNITE is currently specialised on fungal nucleotide sequences but there are no limits on organism group or DNA sequence type that can be submitted or stored for annotating.

SILVA¹⁸⁸ – a comprehensive online resource for regularly updated, quality checked and aligned ribosomal RNA sequence data for all three domains of life (Bacteria, Archaea and Eukarya).

The 16S rRNA Gene Database and Tools (Greengenes)¹⁸⁹ - provides access to the 16S rRNA gene sequence alignment for browsing, blasting, probing, and downloading.

NGS sequences:

Sequence Read Archive (SRA)¹⁹⁰ – stores raw sequencing data from the next generation of sequencing platforms (e.g. Roche 454 GS System, Illumina Genomy Analyzer, etc.).

Genomic Standards Consortium (GSC)¹⁹¹ – standardising the description, exchange and integration of molecular/genomic data.

¹⁸² <http://www.ebi.ac.uk/ena/>

¹⁸³ <http://www.insdc.org/>

¹⁸⁴ <http://www.ncbi.nlm.nih.gov/>

¹⁸⁵ <http://www.ddbj.nig.ac.jp/>

¹⁸⁶ <http://www.boldsystems.org/>

¹⁸⁷ <http://unite.ut.ee/>

¹⁸⁸ <http://www.arb-silva.de/>

¹⁸⁹ <http://greengenes.secondgenome.com/downloads>

¹⁹⁰ <http://www.ncbi.nlm.nih.gov/sra>

Recommendations

- Enhance the GBIF IPT for publishing sample based data by developing a prototype at <http://eubon-ipt.gbif.org> together with a sample data model for use with Darwin Core Archives.
- Enable harvesting and indexing of the Knowledge Network for Biocomplexity (KNB) metadata catalogue by the GBIF registry so that KNB resources are discoverable through EU BON.

¹⁹¹ <http://gensc.org/>