Milestone 231

Version:	1.0
Date:	2014-05-29
Author:	MRAC
Document reference:	Milestone_MS231

Specifications of data sharing tools (M14)

STATUS: FINAL

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 Evolution and Biodiversity Science, Germany
Partners:	UTARTU, University of Tartu, Natural History Museum , Estonia
	UEF, University of Eastern Finland, Digitisation Centre, Finland
	GBIF, Global Biodiversity Information Facility, Denmark
	UniLeeds, University of Leeds, School of Biology, UK
	UFZ, Helmholtz Centre for Environmental Research, Germany
	CSIC, The Spanish National Research Council, Doñana Biological Station, Spain
	UCAM, University of Cambridge, Centre for Science and Policy, UK
	CNRS-IMBE, Mediterranean Institute of marine and terrestrial Biodiversity and Ecology, France
	Pensoft, Pensoft Publishers Ltd, Bulgaria
	SGN, Senckenberg Gesellschaft für Naturforschung, Germany
	VIZZUALITY, Vizzuality S.L., Spain
	FIN, FishBase Information and Research Group, Inc., Philippines
	HCMR, Hellenic Centre for Marine Research, Greece
	NHM, The Natural History Museum, London
	BGBM, Botanic Garden and Botanical Museum Berlin-Dahlem, Germany
	UCPH, University of Copenhagen: Natural History Museum of Denmark, Denmark
	RMCA, Royal Museum of Central Africa, Belgium
	PLAZI, Plazi GmbH, Switzerland
	GlueCAD, GlueCAD Ltd. – Engineering IT, Israel
	IEEP, Institute for European Environmental Policy, UK
	INPA, National Institute of Amazonian Research, Brazil
	NRM, Swedish Museum of Natural History, Sweden
	IBSAS, Slovak Academy of Sciences, Institute of Botany, Slovakia
	EBCC-CTFC, Forest Technology Centre of Catalonia, Spain
	NBIC, Norwegian Biodiversity Information Centre, Norway
	FEM, Fondazione Edmund Mach, Italy
	TerraData, TerraData environmetrics, Monterotondo Marittimo, Italy
	EURAC, European Academy of Bozen/Bolzano, Italy
	WCMC, UNEP World Conservation Monitoring Centre, UK

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

EU BON

EU BON: Building the European Biodiversity Observation Network

Project no. 308454

Large scale collaborative project

MS231

Specifications of data sharing tools

Milestone number	MS231
Milestone name	Specifications of data sharing tools
WP no.	WP2
Lead Beneficiary (full name and Acronym)	Royal Museum of Central Africa, RMCA
Nature	Written report_Data sharing tools specified
Delivery date from Annex I (proj. month)	2014-01-31 (M14)
Delivered	yes
Actual forecast delivery date	2014-05-29 (M18)
Comments	

Names of the Authors	Name of the Partner	Logo of the Partner
Hannu Saarenmaa	UEF	UNIVERSITY OF EASTERN FINLAND
Patricia Mergen, Kim Jacobsen, Larissa Smirnova, Franck Theeten	MRAC	ROYAL MUSEUM FOR CENTRAL AFRICA Afrika Afrique TERVUREN
Israel Pe'er	GlueCAD	GlueCAD Eng. & Biodiversity IT
Éamonn Ó Tuama	GBIF	GLOBAL BIODIVERSITY INFORMATION FACILITY
Lyubomir Penev	Pensoft	PENSOFT
Debora Drucker, Flávia Pezzini, William Magnusson	FDB-INPA	INPA INSTITUTO NACIONAL DE PESQUISAS DA AMAZÔNIA
Anton Güntsch	FUB-BGBM	BGBM
Sarah Faulwetter, Christos Arvanitidis	HCMR	hcmr ΕΛΚΕΘΕ
Urmas Kõljalg, Kessy Abarenkov	UTARTU	UNIVERSITY OF TARTU
Nils Valland	NBIC	NORWEGIAN BIODIVERSITY INFORMATION CENTRE
Donat Agosti, Terry Catapano, Robert Morris, Guido Sautter	PLAZI	PLAZI taking care of freedom
Names of the Authors	**Name of the Advisory**	**Logo of the Advisory**
Bruce Wilson (Member of the EU BON Informatics Task Force)	DataONE	DataONE

In case the report consists of the delivery of materials (guidelines, manuscripts, etc)

Delivery name	**Delivery name**	**From Partner**	**To Partner**

Contents

1. Introduction

This document gives an overview of information and knowledge sharing tools currently available for the biodiversity research community and makes recommendations towards the main requirements needed to build new releases of data sharing tools for EU BON data providers. The Description of Work defines the task 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)"*

Other existing tools which may be used for data sharing, such as those used by organisations to comply with the requirements of the INSPIRE[1] directive, other GIS related tools and crowdsourcing tools, contribute as well substantially to the community and have been included in this report. However, especially tools for sharing habitat data need to be further investigated and agreed upon.

2. Definitions and concepts

2.1 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 formalise and categorise the multiple aspects of the notion of "data". Furthermore, the integration of biodiversity data, which may include at least formats of genetic sequences, species distribution (/abundance/biomass/production) values and habitat maps, requires clear unambiguous identifications of terms for data.

Data is a set of values of quantitative measurement of, or a qualitative fact about something 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 discreet 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 discreet units of data may be further organised into a dataset, which is an organisational 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 many thousands of files, and it is possible for a single database to contain many datasets. The organisation of data into files and

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

datasets is generally not standardised and depends on the particular needs of the individuals collecting the data and the anticipated uses of that data.

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." (ISO)

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.

Metadata is “data about other data”, based on standard specific to a particular discipline. Metadata is 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’). For instance, metadata description of a business shop would be the name, subject, nature or category of inventories, location, address and opening hours.

In the GBIF context, 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’."*

Within the biodiversity domain the metadata description (file or data) should automatically be assigned to all processed and published data or object. As a requirement for EU BON a tool for data sharing should guaranty that the link between the metadata and data/object cannot be lost. 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 EML[1], Darwin Core Archive[2], ISO 19115[3], and ABCD[4], to name a few. These data standards and others have been extensively presented and reported about in D2.1[5].

2.1.1 Data vs. information

Data or 'raw data' (also known as primary data) is a term of unit level collected from a source. From the perspective of the infrastructure service provider an important distinction between data and information is that (raw) data entities are provided, defined and described by an external source, outside of the scope of the infrastructure. That data doesn't yield much information until it is processed (hence interpreted). Once processed, the data may support particular types of *information.*

[1] Ecological Metadata Language (http://en.wikipedia.org/wiki/Ecological_Metadata_Language)

[2] http://en.wikipedia.org/wiki/Darwin_Core_Archive

[3] Geographic information – Metadata(http://www.iso.org/iso/catalogue_detail.htm?csnumber=26020)

[4] Access to Biological Collection Data (http://wiki.tdwg.org/ABCD)

[5] http://www.eubon.eu/documents/1/

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

The LifeWatch[1] information models, which aim to conform with the INSPIRE[2] Implementation Rules, address the differences between data and information (in accordance with Federal Standard 1037C[3]) in its 'Information View'.

- *Data: Representation of measurements, facts, concepts, or instructions in a formalised manner that can be process 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 Model[4] 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)*

2.1.2 Meta-information

While there is a clear understanding how to distinguish between "data" and "information", the terms *meta-data* and *meta-information* are often used interchangeably.

Meta-information is the descriptive information about resources within a required context of a particular purpose.

In ORCHESTA[5], meta-information is the kind of data needed (by the various meta-information models) for particular tasks *"where many different resources (services and data objects) must be handled by common methods and therefore have to have/get common attributes and descriptions (like a location or the classification of a book in a library)."*

Examples of 'Purposes of data' that are handled by different meta-information models include: Discovery, Orchestration, Collaboration, Identification, Authentication and Authorisation, Provenance, Quality evaluation, Indexing, Retrieving, Integration.

2.1.3 Processed and secondary data and information

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

[1] http://www.lifewatch.eu/web/guest/home

[2] http://inspire.ec.europa.eu/

[3] http://en.wikipedia.org/wiki/Wikipedia:Federal_Standard_1037C_terms

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

[5] Orchestra Networks - data management software provider (http://www.orchestranetworks.com/)

2.1.4 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 describe 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 explicitly referencing a particular term in a relevant ontology or from a specifically referenced controlled vocabulary of keywords.

Section 13 of D2.1 outlines some recent developments regarding vocabularies and ontologies in biodiversity informatics and provides some recommendations for EU BON on their adoption.

Recommendation: Within EU BON documentation, the term *data* should be associated with the purpose and the context in which this data is used whenever an ambiguous interpretation might arise.

2.2 Data publishing

Biodiversity data can be shared or made publicly available through the process of 'publishing'. Publishing makes the data accessible through the use of standard procedures and protocols. The term is often used interchangeably with 'data sharing'. However, data publishing additionally 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 for making datasets public.

GBIF and Pensoft[1] summarise the incentives to publish biodiversity data as follows[2]:

- Data can be indexed and made discoverable, browsable and searchable through biodiversity infrastructures (e.g., GBIF, Dryad[3] and others):
- Discoverable and accessible data contributes 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.

[1] http://www.pensoft.net/

[2] http://www.gbif.org/publishingdata/summary; Chavan, V., Penev, L. (2011) The data paper: a mechanism to incentivize data publishing in biodiversity science. BMC Informatics, 12 (Sppl 15): S2 (accessed online at http://www.biomedcentral.com/1471-2105/12/S15/S2 on 18/03/2014)

[3] http://www.datadryad.org/

- Collection managers can trace usage and citations of digitised 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.

The use of 'Data papers' was recently promoted for the biodiversity community by Chavan and Penev (2011), although the concept has been used in other domains and is not new: data papers have been published by the Ecological Society of America in Ecological Archives since 2000[1], Earth System Science Data[2], CMB data papers[3], BMC Data Notes[4] and the International Journal of Robotics Research[5] , to mention a few examples.

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.

Huang *et al.* (2013)[6] have challenged the applicability of the data paper stating that: (1) peer-reviewed data papers will draw too strongly on a limited resource base of peer-reviewers, (2) this format is a barrier to inclusion of citizen-science biodiversity data, (3) they are dependent on who will foot the bill for publication, (4) duplication of efforts, such as the use of online supplemental materials, already provided by many journals and (5) various issues related to the technical execution of data paper publications.

In response, Chavan *et al*. (2013)[7] state that data papers are fully compatible with the stable, long-term access to well-described, high-quality data sets through the implementation of a joint data-publishing and archiving policy by databases and journals (Huang *et al*., 2013). With a number of situations where a data paper may in fact be preferable as a means to help foster data mobilisation of and access to currently non-digitised or unpublished data sets, such as: (1) historical data sets that offer views of the past composition of biodiversity in specific locations or that offer time series for analysis; (2) to enable availability of a data set with all necessary documentation on methods and other details, while also giving the publishing researcher value in the form of a citable publication. One of the important aspects of a data paper is its subjection to peer-review, which dixit Chavan *et al.* (2013) will eventually lead to the emergence of a new breed of peer-reviewers over time: with an understanding of data collection, management, and publishing, as well of the potential uses of the data. Without such peer-review, however, controls for the

[1] Ecological Archives: Data Papers, Supplements, and Digital Appendices for ESA Journals (http://www.esapubs.org/archive/default.htm)

[2] Earth System Science Data (http://www.earth-syst-sci-data.net/)

[3] CMB data papers (http://lambda.gsfc.nasa.gov/outreach/recent_papers.cfm)

[4] BMC Research Notes: Data Notes (http://www.biomedcentral.com/bmcresnotes/ifora/?txt_jou_id=4005&txt_mst_id=104807)

[5] International Journal of Robotics Research (http://ijr.sagepub.com) and Editorial: Data Papers — Peer Reviewed Publication of High Quality Data Sets. Int J Robot Res 2009, 28:587, doi: 10.1177/0278364909104283.

[6] Huang, X., Hawkins, B.A., Qiao, G. (2013) Biodiversity data sharing: will peer-reviewed data papers work? BioScience, 63, 1, 5-6. doi:10.1525/bio.2013.63.1.2 Downloaded from http://bioscience.oxfordjournals.org/ on March 17, 2014

[7] 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

standards of data and metadata and their reuse can be problematic (Costello *et al*., 2012[1] cited in Chavan *et al*., 2013). Recent developments include the endorsement of the data paper concept by several EU-funded projects such as ViBRANT [2](Virtual Biodiversity Research and Access Network for Taxonomy) and BioFresh[3] (a program to support freshwater biodiversity) and the creation of the next-generation Biodiversity Data Journal. 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).

2.3 Data sharing

2.3.1 What is data sharing?

Wikipedia defines Data sharing as "the practice of making data used for scholarly research available to other investigators"[4]. 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 is about enabling big data solutions and approaches.[5]

But 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 attached.[6]

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.

One of the important areas of concern is a climate change where particular attention is paid to possible interactions between different sectors, e.g., agriculture, water, energy, hazards, and health. In the past research efforts were limited by the difficulty of assembling and integrating diverse data types coming from different platforms and collected with different instruments. Another challenge is the need for integration of data across scientific disciplines, especially across the natural and social sciences, which would help to better understand the interactions between climate and human activity[7].

A major player in this area is The Global Earth Observation System of Systems (GEOSS)[8] which is simultaneously addressing nine areas of critical importance to people and society. Its main role is the coordination and quality control of data gathered from different instruments and multiple observing platforms and the provision of an overall framework for rapid integration of both remote sensing and *in situ* datasets. By promoting interoperability

[1] Costello MJ, Michener WK, Gahegan M, Zhang Z-Q, Bourne P, Chavan V. 2012. Quality Assurance and Intellectual Property Rights in Advancing Biodiversity Data Publications, ver. 1.0. Global Biodiversity Information Facility.

[2] http://vbrant.eu/

[3] http://www.freshwaterbiodiversity.eu/

[4] Wikipedia (http://en.wikipedia.org/wiki/Data_sharing)

[5] Kent Anderson. 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/)

[6] *http://www.nature.com/nature/journal/v461/n7261/full/461171a.html*: Standards and tool development

[7] http://www.spacelaw.olemiss.edu/jsl/pdfs/articles/35JSL201.pdf

[8] https://www.earthobservations.org/geoss.shtml

among many different data sources and systems from around the world, GEOSS will facilitate testing and inter-comparison of measurements and increase the representation and reliability of the results.

GEOSS Data Sharing Principles[1]

The vision of GEOSS is *"to realize a future wherein decisions and actions for the benefit of humankind are informed via coordinated, comprehensive and sustained Earth observations and information"*. Among other strategic goals these two are of particular interest for data sharing[2]:

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

To achieve its vision and strategic goals, GEOSS has adopted the *10-Year Implementation Plan* which acknowledges the importance of data sharing. The Plan, endorsed by nearly 60 governments and the European Commission, highlights the following **GEOSS Data Sharing Principles**:

- 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 being 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 the cost of reproduction will be encouraged for research and education.

EU BON data sharing agreement[3]

EU BON conceives its work on the basis of the GEOSS Data Sharing Principles, and in the EU BON Data Sharing Agreement it is defined as "full and open exchange of data, metadata, and products shared within GEOSS, recognising relevant international instruments and national policies and legislation. All shared data, metadata and products shall be made available with minimum time delay and at minimum cost. Availability of all shared data, metadata and products free of charge or at no more than cost of reproduction shall be encouraged for research and education"[4].

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",

[1] See also D2.1: http://www.eubon.eu/documents/1/

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

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

[4] http://www.earthobservations.org/geoss_dsp.shtml

presented by the European Federation of Academies of Sciences and Humanities and the European Commission on 11th April, 2012[1].

2.3.2 Challenges

Different studies (Tenopir *et al.,* 2011[2]; Hardisty *et al.,* 2013[3]) discuss 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.* 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 re-using data remains the lab and/or individual working environment. Main obstacles noticed are insufficient time and lack of funding. So sharing data is still a complex and challenging issue.

The authors[3] also give several recommendations which are necessary to reduce duplication and enhance collaboration, notably in bioinformatics:

“1. Open Data should be normal practice and should embody the principles of being accessible, assessable, intelligible and usable.

2. Data encoding should allow analysis across multiple scales. The encoding schema needs to facilitate the integration of various data sets in a single analytical structure.

3. Infrastructure projects should devote significant resources to market the service they develop, specifically to attract users from outside the project-funded community, and ideally in significant numbers. To make such an investment effective, projects should release their service early and update often, in response to user feedback.”

GEOSS has already identified several challenges they will be facing:

- “not only access to relevant data is important, but a clear understanding of how the data were collected, what quality control procedures were utilised, and what transformation and analysis techniques were applied. A basic step in obtaining such understanding is access to appropriate metadata, i.e., documentation that describes data sources and processing. Encouraging all data providers to provide adequate metadata for their data is therefore a key priority. Free and open access to this metadata is then necessary to ensure that all users can discover the data they may need.
- a second critical issue for both researchers and data sources is appropriate data attribution. For data providers to continue providing high quality data and metadata in the long term, they will need to receive appropriate recognition for the data they supply. From the viewpoint of the scientific community, being able to precisely trace data “provenance”— i.e., data sources and processing histories — is essential to the reproducibility of scientific research. From the viewpoint of commercial providers, identifying them as the data source can enhance the reputation of their products and provide a further incentive to provide access to their data” [4].

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

[2] http://www.plosone.org/article/info%3Adoi%2F10.1371%2Fjournal.pone.0021101

[3] http://www.biomedcentral.com/1472-6785/13/16

[4] http://www.spacelaw.olemiss.edu/jsl/pdfs/articles/35JSL201.pdf

- interoperability[1]: arrangements and standards for sharing and integrating data.
- "legal interoperability for data which means that the legal rights, terms, and conditions of databases from two or more sources are compatible and the data may be combined by any user without compromising the legal rights of any of the data sources used".[2]

Improving data access and sharing should significantly increase data utilisation by reducing the cost and reuse restrictions for the users. This should create innovative opportunities for new and existing players in the information sector to improve and expand their activities".

2.4 Data sharing tools

2.4.1 Introduction

There are quite a few data sharing tools out there. Some have also other purposes, but are used by the community to share data and information such as tables or even structure or semi-structures text documents. These tools are often easy to use, well known by the community and have the advantage that they do not require a long and steep learning curve, particular IT skills or the assistance of an IT specialist. On a short term perspective they give the impression a quick win for data exchange.

Spreadsheets (like Excel tables) or comma/tab separated well-structured files should not be ruled out as efficient means to share data and information. Indeed they can be very useful to transfer data among collaborators or to feed higher level data management systems or applications. The learning curve is often less steep and exchange of data can happen routinely without need of regular support of IT literate staff.

On the other hand, using such basic tools or using them without respecting a clear structure can create issues for larger scale use or interoperability. Importation into other applications can be difficult or even impossible and thus become a barrier to data sharing and later reuse or accessibility.

Accordingly, in order to overcome these barriers, data sharing tools that help or force the user to structure the data have been developed. Some are more generic and data schema independent and can be used in multiple domains.

Other tools, on the other hand, are very specific and designed for selected data types, models, specific applications or purposes.

One can cite here tools to exchange geographic information that is 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 designed in the context of a project or of an application. They are very useful but often need adaptations or connector applications to become interoperable, at larger scales.

[1] https://www.earthobservations.org/documents/geo_vi/07_Implementation%20Guidelines%20for%20the%20GEOSS%20Data%20Sharing%20Principles%20Rev2.pdf

[2] http://www.codata.org/GEOSS/EU.pdf

Data publishing tools can process for example raw data into reports or publications to be further shared as information for educational, decision making, policy making purposes, which is an additional form of information sharing (see part 2.2).

2.4.2 Tools surveyed by EU BON

In the context of EU BON, it is important to have a good overview of existing tools and how they can be useful for the domain. In **Annex 1**, there is a list of selected existing data sharing tools that have been assessed by the community. Summarised overview of these tools is given in the table in part 2.4.3.

Each tool is presented using the same structure:

Main usage, purpose, selected examples

Pros and Cons of the tool

Recommendations

Tool status

The list in Annex 1 is not meant to be exhaustive but rather as a snapshot of the current state of art and knowledge of the community relevant to data sharing tools. This report can thus be used also for a gap analysis on tools yet needed to be developed. For instance, there seems to be a void for tools for sharing habitat data. In practice, this report will be complemented by a dynamic list of tools on the helpdesk of the project[1] with regular updates of the status of the tools, their recommended usages in the context of EU BON. Likewise, additional tools, newly discovered and analysed tools or newly developed tools will be added to the list. The current analysis hasn't started from scratch, but it is based on previous analysis of tools made in the framework of the projects EDIT (European Distributed Institute of Taxonomy)[2] and SYNTEHSYS (Synthesis of Systematic Resources)[3] and which are taken into account by using the BDTracker system[4].

It is important for the needs of the data sharing tool features in EU BON to have different groups of tools, for example tools, such as those that are specialised in species occurrence type of data that should, in turn, be combined or made interoperable with tools specialised 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)[5] and LTER[6] are particularly useful. For species occurrence data the data sharing tools of GBIF[7] adhering to the TDWG[8] Biodiversity Information Standards are widely used and very relevant.

A data architecture, workflow and data standards to use have been defined in D2.1. In this context data sharing tools are to some extent defined as data providing tools. These tools should be taken into account by EU BON.

[1] EU BON helpdesk (http://eubon.cybertaxonomy.africamuseum.be/)
[2] http://www.e-taxonomy.eu/
[3] http://www.synthesys.info/
[4] The Biodiversity Service & Application Tracker (http://bdtracker.cybertaxonomy.africamuseum.be//)
[5] http://www.eea.europa.eu/
[6] The Long Term Ecological Research Network (http://www.lternet.edu/)
[7] Global Biodiversity Information Facility (http://www.gbif.org/)
[8] http://www.tdwg.org/

EU BON, as stated in its DoW and Data Sharing Agreement, has close ties to GEOSS (The Global Earth Observation System of Systems)[1]. The data sharing tools to use should, to a large extent, be compatible with the GEOSS community tools. 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 embargo periods or other related IPR requests.

In relation with the overall global GEO BON[2] initiative to which EU BON constitutes the European node, tools that are able to handle the Aichi Targets[3] and the EBVs[4] are needed in order to make the EU BON/GEO BON platform for data sharing effective.

Different tools or just flexible enough 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 sites managers, scientists engaged in monitoring programs, modellers, decision and policy makers as well as interested citizens. In this regard, the project DataONE has identified 'personas'[5] which could be re-used to identify which tool is suitable for which users.

This report mainly focuses on data sharing or also so called data providing 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 clearly needed for the overall workflow in the context of EU BON, but do not directly form part of the scope of this document. They have been partly covered in D2.1 and will be further addressed in other upcoming EU BON deliverables and activities.

[1] https://www.earthobservations.org/geoss.shtml

[2] GEO BON Biodiversity Observation Systems(https://www.earthobservations.org/geobon.shtml)

[3] http://www.cbd.int/sp/targets/

[4] Essential Biodiversity Variables (https://www.earthobservations.org/geobon_ebv.shtml)

[5] http://dataone-sc.wikispaces.com/Personas

2.4.3 Summary of tools from Annex 1.

The abbreviations are explained in the full text of tool specification.

Tool	**OS***	**Short description**	**Standard supported****	**EU BON relevance**		**Recommendations towards EU BON objectives**
				yes	no	
Tools to share biodiversity data						
GBIF Integrated Publishing Toolkit (IPT)	x	Tool to publish and share biodiversity data sets and metadata through the GBIF network. Allows publication of two types of biodiversity data: i) primary occurrence data (specimens, observations), ii) species checklists and taxonomies	DwC , DwC-A, EML	x		Enhance IPT for sample-based data sets. Extend GBIF's IPT to handle sample based datasets and cooperate with EuMon to harvest datasets from biodiversity monitoring.
GBIF Spreadsheet-Processor	x	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.	DwC, DwC-A EML	x		EU BON should work with projects to promote the use and structuring of Spreadsheet processors with existing publishing tools to encourage inputs from the huge communities of Excel based data providers.
Biodiversity Data Journal[1] and Pensoft Writing Tool	x	Narrative (text) and data integrated publishing workflow to mobilise, review, publish, store, disseminate, make interoperable, collate and re-use data through the act of scholarly publishing. Two types of biodiversity data supported: (i) primary occurrence data (specimens,	DwC, DwC-A EML			Enhance PWT and BDJ for traits data, and sample based DwC compliant data sets. Use the technologies invented by BDJ to re-publish legacy literature.

[1] http://biodiversitydatajournal.com

		observations), (ii) Species checklists and taxonomies				
Bibliography of Life	x	A platform consisting of three integral tools, RefBank[1] and ReFindit[2] and Biosystematics Literature Repository based at ZENODO/CERN. 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.	MODS OAI-PMH	x		Enhance Bibliography of Life to domains other than biodiversity through amendment of new searched platforms and harvesting mechanisms to enrich the content of RefBank.
Metacat: Metadata and Data Management Server	x	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.	EML ISO 19139 FGDC Biological Data Profile	x		Enhance GBIF capability to bridge and interoperate with existing data-providers/repositories for environmental data and LTER, namely DataONE/KNB, thus exposing Metacat datasets through the EU BON portal.
DataONE Generic Member Node	X	Is a python reference implementation of a complete (Tier 4) member node to DataONE. 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.		x		The GMN should be investigated as an option for standing up a data sharing environment for partners and national organisations supporting WP4 and 5, particularly for data that is not suitable for inclusion in GBIF.
DataONE Slender Node		Software stack 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.				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

[1] http://refbank.org
[2] http://refindit.org

						projects, as well as citizen science projects.
Morpho Metadata Editor		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.	EML	x		Allows ecological data curation, assuring that data tables are correctly built. Means to relate taxonomic coverage with DwC 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.
GeoServer	X	Server software written in Java that allows users to share and edit geospatial data. GeoServer is the reference implementation of the WMS, WFS and WCS standards of the OGC, as well as a high performance certified compliant WMS.	WMS WFS WCS + CS-W (INSPIRE)			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.
GeoNetwork	x	Software server allows users to share and edit geospatial metadata and to link them to on maps that are available on line in a search interface. Metadata are based on the ISO 19115 and ISO 19139. It is interoperable with any maps server provided in WMS and CSW formats. Compliant with the Z39.50 and OAI-PMH protocols.	CS-W + ISO 19115 ISO 19139 (INSPIRE) + OAI-PMH			Using GeoNetwork would allow a good interoperability with ISO, OGC and INSPIRE standards. It allows linking together metadata, data, maps and thesaurus.

Data Access Protocol-compliant servers (DAP)		REST web service based protocol designed for science data. There are multiple software packages which implement DAP, e.g. OPeNDAP Hyrax and THREDDS. There is current development to make Hyrax and THREDDS DataONE-enabled.				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.
DiGIR	x	XML-based protocol to implement queries to distributed data providers. It is modelled after the Z39.50 protocol. Supports several operations such as inventory of information resources on a provider, download to resource metadata, and queries to the full data.	DwC, ABCD		x	
TAPIRlink	x	TDWG Access Protocol for Information Retrieval. Its purpose was to unify the DiGIR and BioCASe protocols and make the protocol independent of certain schemas.	DwC, ABCD	x		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.
BioCASE	x	A software and 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	ABCD, DwC, DwC-A			Collection and observational data not yet available to biodiversity informatics infrastructures such as EU BON could be exposed via the BPS tool.

		standards and protocols				
Scratchpads	x	Virtual research environments — a web-based content management software (based on Drupal) which facilitates the organisation and publication of biodiversity data (mobilisation, structuring, linking and dissemination of taxon-centric information)	DwC-A, ABCD			Scratchpads are targeted towards managing and sharing small pieces of data pertaining to taxa / biodiversity. 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. Given the increased introduction of biodiversity 'Data Paper' into the publication domain, EU BON should seek to enhance the integration of Scratchpads, e.g. importing 'taxon pages' that are already DwC Archives enabled.
PlutoF		Online service to create, manage, share, analyse and mobilise biodiversity data. Data types cover ecology, taxonomy, metagenomics, nature conservation, natural history collections, etc.				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.
DSpace	x	Digital object management system, useful for managing arbitrary digital objects, such as data files. There is current work to DataONE-enable DSpace.				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.

Dryad Digital Repository		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.				
Species Observation System	x	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.				Major potential tool for broader European citizen science involvement in species mapping, surveillance and monitoring. Open the door for citizen science based portals such as Species Observation Service, while seeeking to standardise quality management of SC data.
DEIMS	x	The International Ecological Information Management System (DEIMS) 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.	EML ISO	x		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, and to provide the implementation facility for the CSW.
Plazi TaxonomicTreatment Server		A platform to store, annotate, access and distribute treatments and the data objects within. It offers with GoldenGate[1] and respective XML schemas (TaxonX[2], TaxPub[3]) tools to convert	DwC	x		The project needs to invest in human-machine interfaces, documentation and training, and tools that allow the easiest possible way to annotate the

[1] http://plazi.org/?q=GoldenGATE

[2] http://plazi.org/?q=taxonx

[3] https://github.com/tcatapano/TaxPub/releases

		unstructured text into semantically enhanced documents with an emphasis on taxonomic data like treatments, scientific names, materials observation, traits or bibliographic references.				treatments. Specific services, such as bibliographic name provision and materials examined parsing need to become standalone applications. Trait extraction needs be developed.
Spreadsheet tools		Microsoft Excel, DataUp and open source tool Libre Office are a software packages that enables the creation of spreadsheets or forms, provides simple data comparison and analysis tools, and creates graphs.		x		EU BON should help advance the use of best practices for data in Excel as well as advancing the education of other options for data analysis tools.
Database packages		Commercial software storing data, such as Microsoft Access, Microsoft SQL Server, and Oracle, and open source tools such as MySQL, PostgreSQL, and SQLite.	compatibility with standards via above mentioned software	x		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.
Tools to share molecular data						
European Nucleotide Archive (ENA)		Captures and presents information relating to experimental workflows that are based around nucleotide sequencing. INSDC forms the most comprehensive database for all molecular data types and linked metadata.				
The Barcode of Life Data Systems (BOLD)		Supports 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/PlutoF		An online resource for regularly updated, quality checked and annotated ribosomal DNA sequence data for kingdom Fungi.				
SILVA		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)		provides access to the 16S rRNA gene sequence alignment for browsing, blasting, probing, and downloading.				
Sequence Read Archive (SRA)		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)		Standardising the description, exchange and integration of molecular/genomic data.				

*Open Source

3. Conclusions and specification for further work

In conclusion, it can be said that understanding the needs of the end users and having a good overview of existing data sharing tools is essential to the EU BON project in order to address the data sharing needs of the widest community, and to handle the broadest range of data and information within the context of earth observation.

We can broadly group the tools into distributed and centralised categories. The distributed ones are being used and managed by the data custodians themselves. The centralised ones are shared repositories not managed by the data custodians, but by an aggregator or publisher.

The distinction of tools for sharing and publishing is also important. Data that is shared can still be private and access to it can be controlled. Such access can be revoked. When something is published, it is openly available, and access cannot be revoked anymore.

The tools can also be categorised as specialised or general purpose. Specialised 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.

In this report we have identified so many tools that they cannot possibly all be used or supported by the EU BON project. The purpose of this document is to provide a specification for the work that the EU BON project will be doing with the data sharing tools. Those tools will then be supported, distributed, and used to feed data into the EU BON Portal and to GEOSS. This means that we will have to make choices and action lists.

The minimum set of data sharing tools on which the EU BON project focuses is the first named group in each of the above categories. That is, distributed, controllable, and specialised. This limits the choice of tools in those in the table below: GBIF IPT, Morpho/Metacat, and spreadsheet processors.

Other tools may still find their place as components of the EU BON Portal where aggregated data can be presented and the data publishing process is being managed.

The limitation and possible problem of the approach outlined here is in the word “specialised”. There simply are not distributable data sharing tools specialised for each biodiversity data type (genomic, occurrence, species, habitat, ecosystem, …), but rather only for occurrence and species level data. The question is whether specialised tools are needed at all for each data type. For example, there may actually be potential users for specialised tools for sharing habitat data, which could be explored.

Now that the scope has been defined, we can look at the state of the chosen tools and see what enhancements may be required to support EU BON priority use cases (see D2.1), and the GEOSS Data Sharing Principles, and the EU BON Data Sharing Agreement.

The GBIF IPT needs to be enhanced to support a sample-based data model. This also includes revising the Darwin Core standard so that quantitative measurements have broader support. GBIF is already working with TDWG to make that happen. The IPT may also need to have support for an embargo period, although this activity can be handled manually for now. Furthermore, many ecological datasets are much richer than what the Darwin Core standard can or ever will support. It should still be possible to include such data as “payload” in the Darwin Core Archive documents that are being used to exchange data.

Further work on generalising the Darwin Core Archive will be needed in this and probably also in other areas. In summary, we could say that the GBIF IPT must learn from Metacat, and become more like it.

Metacat, on the other hand, is a mature product that does not need many changes. It can be used to host any kind of data and it supports the embargo functions. However, the shortcomings arise from the prevailing use of Metacat. Data files can be uploaded in any format without regard to any data standards. So, correspondingly, Metacat could perhaps learn from DwC-A by enabling at least a subset of an EML data package to be mapped to well-known vocabularies like Darwin Core. The goal would be to allow automatic processing of data from Metacat repositories, for instance, in production of the Essential Biodiversity Variables.

List of actions

Subject	**Aim**	**Action(s)**	**Comments**
Sample-based datasets	Enhance GBIF Integrated Publishing Toolkit (IPT) with the sample-model	1. Extend the IPT to handle Sample-model 2. Discuss ways to harmonise sample-base metadata with EuMON (biodiversity monitoring). 3. Investigate generalising the DwC-A format to support any files.	A specially adapted version of IPT is currently undergoing testing.
DataONE Network	Integration with EU BON Portal	1. Install and Implement DataONE Member Node on EU BON test site. 2. Investigate (with the assistance of DataONE/KNB) means to share services that will enable queries for data within DataONE repositories. 3. Discuss and provide documented requirements and use-cases (e.g. EBV) for the implementation and testing of such queries within the EU BON portal.	Investigate tool (e.g. GMN) that can be used to adapt a repository's existing software.
Publishing tools: Morpho Metadata Editor (KNB)	Enhance the integration with publishing tools	1. Explore using Morpho (editor) and Metacat (servers) for managing ecological metadata to access and expose LTER sites /datasets. 2. Design feasibility test to clarify and document the requirements for implementation.	Define use cases for tests.

Publishing tools: Excel files	Spreadsheet processors (e.g. DataUp)	1. 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.	

Following the above recommendations, combined with those of D2.1, and having identified case studies and test sites in the context of the other WPs, the EU BON community is now ready to concretely test and implement those tools. Special attention must be paid to the GEO BON expectations and particularly the specific needs of the Aichi Targets and of the EBVs. However, it is clear that the community is still a long way from having the required data standards available – further support and community discussion is needed. The Gene Ontology bioinformatics[1] initiatives provide a good example of how parallel development of tools and standards generates added value. Dedicated funding is needed to develop key elements of database infrastructure, including interoperability and data integration.

[1] http://www.geneontology.org/

Annex 1: Non-exhaustive list of tools:

Additional lists are available through the GBIF resources page[1] , the DataONE software tools catalog[2] and the BDTracker[3].

A.1 GBIF Integrated Publishing Toolkit (IPT)[4]

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 manuscript[5] 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.

Version 2.0.5 of the IPT is available for download in both compiled[6] and source code[7] 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 VertNet[8] 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.

INBO (The Research Institute for Nature and Forest)[9] and Canadensys[10] 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.

SiB[11] Colombia uses the IPT as a central part of their data publishing model[12] in which it has facilitated publication of primary data.

[1] http://www.gbif.org/resources/summary
[2] https://www.dataone.org/software_tools_catalog
[3] http://bdtracker.cybertaxonomy.africamuseum.be//
[4] http://www.gbif.org/ipt
[5] http://www.gbif.org/publishingdata/datapapers
[6] http://www.gbif.org/ipt/releases
[7] https://code.google.com/p/gbif-providertoolkit/source/checkout
[8] http://vertnet.org/
[9] http://www.inbo.be/content/homepage_en.asp
[10] http://www.canadensys.net/
[11] http://www.sibcolombia.net/web/sib/home
[12] http://www.sibcolombia.net/web/sib/acerca-del-sib

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 interface[1].
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 list[2]; the User Manual is also available in both English[3] and Spanish[4].

Cons

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

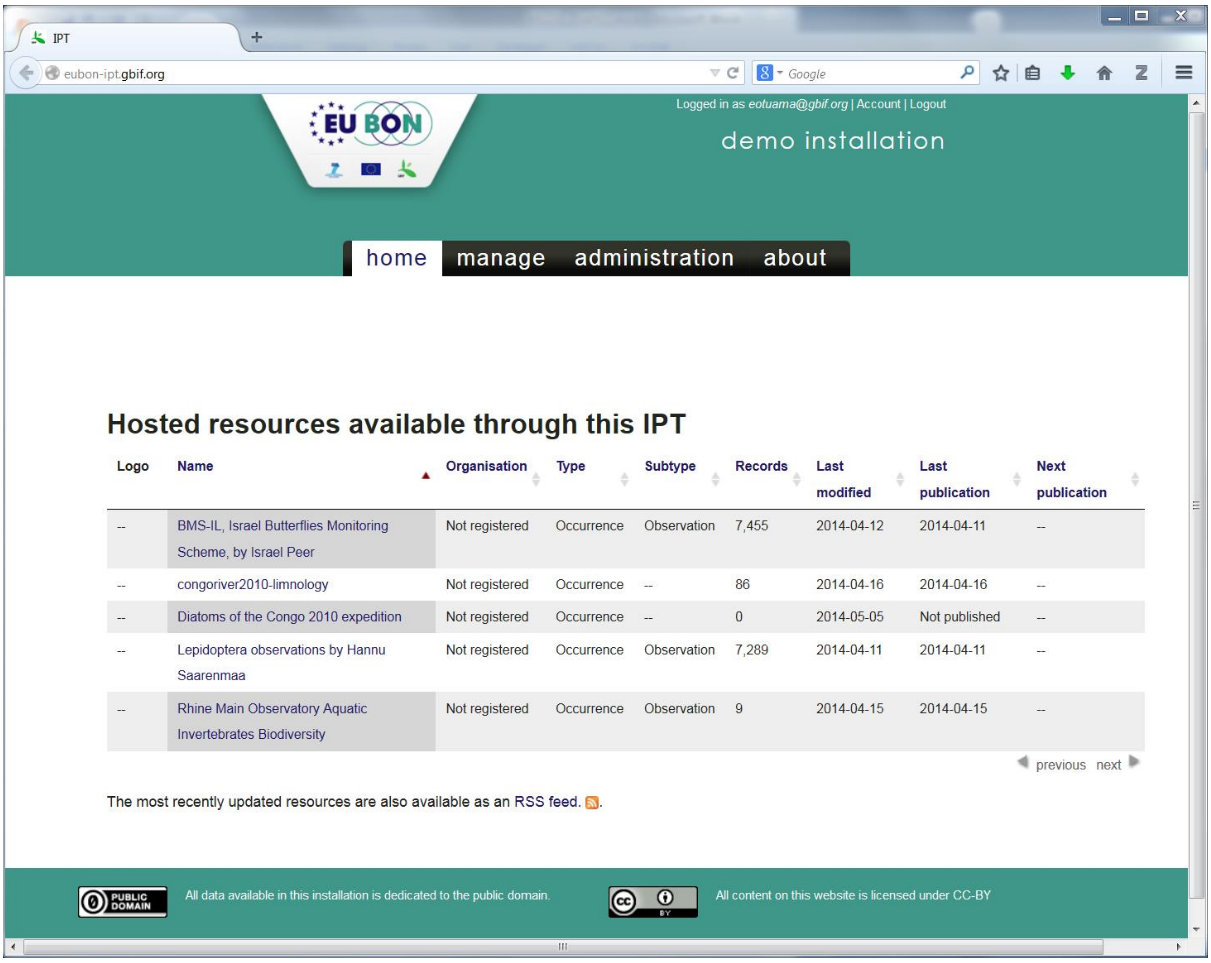
Suggested improvements: enhance IPT for sample-based data sets.

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

[1] https://code.google.com/p/gbif-providertoolkit/wiki/HowToContribute
[2] http://lists.gbif.org/mailman/listinfo/ipt
[3] https://code.google.com/p/gbif-providertoolkit/wiki/IPT2ManualNotes?tm=6
[4] https://code.google.com/p/gbif-providertoolkit/wiki/IPT2ManualNotes?wl=es

modified instance of the IPT that recognises the new core/extension. A prototype IPT (Figure 1) 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.



Logged in as eotuama@gbif.org | Account | Logout

demo installation

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	--

previous next

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 1. 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-Processor[1], 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 Figure 2, the user selects the appropriate

[1] http://tools.gbif.org/spreadsheet-processor/

spreadsheet template, 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.

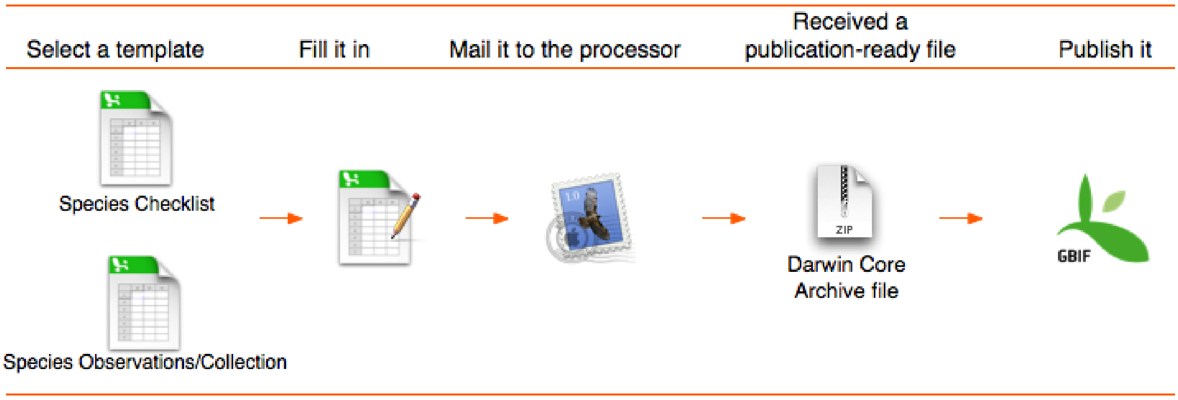


Figure 2. The web based processor ingests a spreadsheet and outputs a validated Darwin Core Archive.

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 Journal[1] and Pensoft Writing Tool[2]

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;
- ecological and biological observations of species and communities;

[1] http://biodiversitydatajournal.com
[2] http://pwt.pensoft.net

- 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 Core[1] and other data files into text and vice versa, from text to data;
- Automated import of data-structured manuscripts generated in various platforms (Scratchpads[2], GBIF Integrated Publishing Toolkit (IPT)[3], 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[4] and *Eupolybothrus cavernicolus* Komerički & Stoev sp. n. (Chilopoda: Lithobiomorpha: Lithobiidae): the first eukaryotic species description combining transcriptomic, DNA barcoding and micro-CT imaging data[5]. 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)[6].

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

Pros and Cons of the tool

Pros:

1. Integrated text (narrative) and data publication of two types of biodiversity data: (i) primary occurrence data (specimens, observations), (ii) Species checklists and taxonomies

[1] http://rs.tdwg.org/dwc/2009-02-20/terms/guides/text/index.htm
[2] http://scratchpads.eu/
[3] http://ipt.pensoft.net/ipt/
[4] http://biodiversitydatajournal.com/articles.php?id=995
[5] http://biodiversitydatajournal.com/articles.php?id=1013
[6] http://eol.org/

2. Occurrence data published in the different papers can be shared and collated together
3. Can be used to publish in the form of “data papers” of any kind of biodiversity-related data.
4. Data and content are archived in PubMedCentral after publication
5. Small datasets are downloadable straight from the article text
6. 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
7. Provides a publication venue for software and tools descriptions

Cons:

1. Currently, the BDJ and PWT are constrained to be used mostly in the biodiversity domain.
2. 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 Life[1] platform was developed within the EU FP7 project ViBRANT and consists of three integral tools, RefBank[2] and ReFindit[3] and Biosystematics Literature Repository based at ZENODO/CERN[4]. 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. 3).

[1] http://biblife.org
[2] http://refbank.org
[3] http://refindit.org
[4] http://zenodo.org

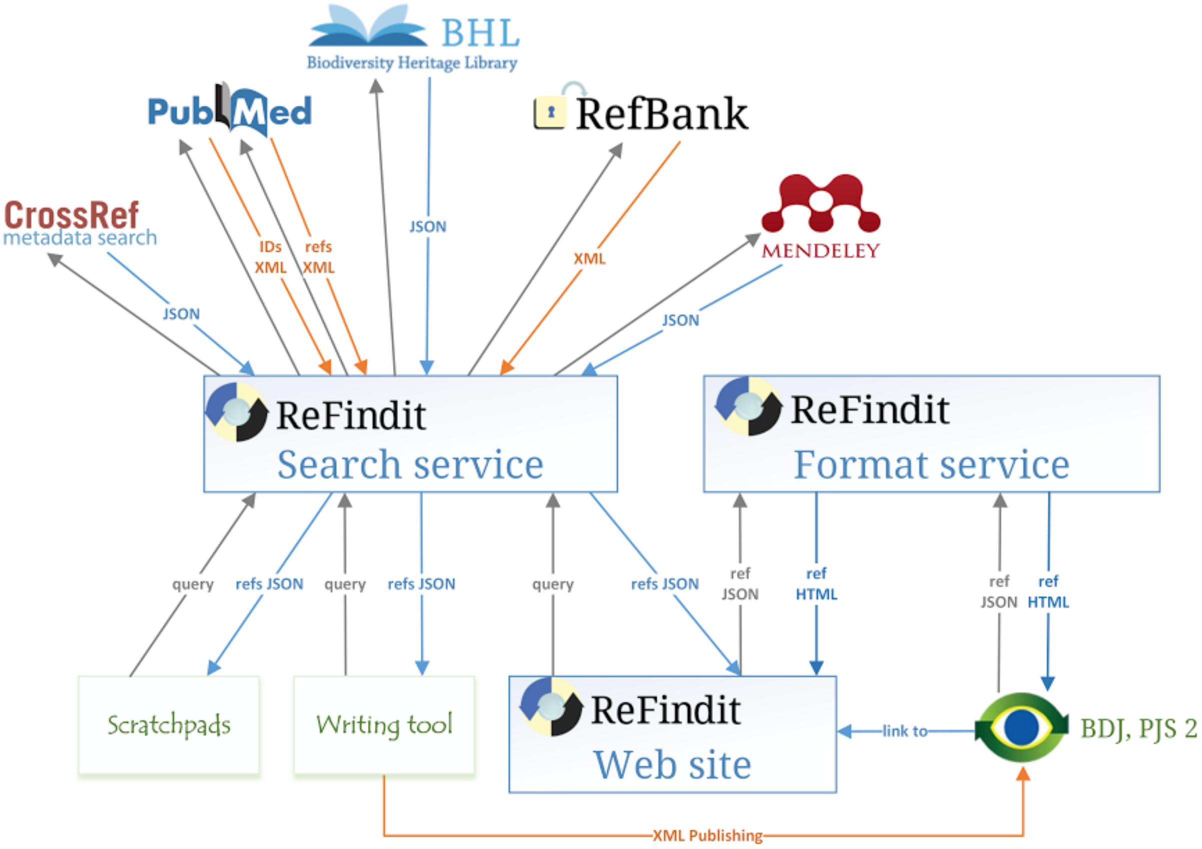


Figure 3. RefBank and ReFindit workflow .

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.

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)[1] and the Biodiversity Data Journal (BDJ)[2].

[1] http://pwt.pensoft.net
[2] http://biodiversitydatajournal.com

Pros and Cons of the tool

Pros

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

Cons

1. Currently, Biodiversity of Life is focusing mostly on the biodiversity domain, although technologically it is not constrained to that.
2. 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 ReFindiit tool ate fully operable. The Biosystematics Literature Repository is currently at beta testing stage.

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 ways[1].

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 is ever lost to technical failures), it provides a user-friendly interface for

[1] information provided by Metacat Administrator's Guide: http://knb.ecoinformatics.org/software/metacat

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 Web[1]. 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 Profile[1].

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 others[1]. 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 http://ppbio.inpa.gov.br/knb/style/skins/ppbio/. 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 itself with the data documentation. Information providers can also edit the metadata or delete it from the repository, again using Metacat's straightforward Web interface[1].

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 set[1]. 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

[1] information provided by Metacat Administrator's Guide: http://knb.ecoinformatics.org/software/metacat

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 interface[1].

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 repository[1]. 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.

[1] https://repository.dataone.org/software/cicore/trunk/mn/d1_mn_generic/

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

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, too[1].

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 is 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 viewed[1].

[1] information provided by Morpho User Guide: https://knb.ecoinformatics.org/software/dist/MorphoUserGuide.pdf

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 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

GeoNetwork[1] 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 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)[2] initiator of the project) and projects (OneGeology[3]).

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 (DAP[4]) is a REST web service based protocol designed for science data. There are multiple software packages which implement DAP, with OPeNDAP Hyrax[1]

[1] http://geonetwork-opensource.org/

[2] http://www.fao.org/home/en/

[3] http://www.onegeology.org/

[4] http://www.opendap.org/pdf/dap_2_data_model.pdf

and THREDDS[2] 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 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 modelers and are an efficient way to expose gridded data, with subsetting 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 modeling 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,

[1] http://www.opendap.org/
[2] https://www.unidata.ucar.edu/software/thredds/current/tds/

it was never standardised by TDWG. Resource metadata is 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 is 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 protocols[1].

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 information systems. The BPS is agnostic to the kind of data being exchanged and any

[1] http://www.biocase.org/whats_biocase/unit_net.shtml

conceptual schema, such as ABCD (Access to Biological Collection Data)[1] for the BioCASE network[2], 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)[3], the Australian Virtual Herbarium (AVH)[4], and GeoCASE[5].

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 standardised 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-Dahlem[6]. 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

Scratchpads[7] 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 (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

[1] http://www.tdwg.org/standards/115/
[2] http://www.biocase.org/whats_biocase/unit_net.shtml
[3] http://ggbn.org/
[4] http://avh.chah.org.au/
[5] http://www.geocase.eu/
[6] http://www.bgbm.org/en/biodiversity-informatics
[7] http://scratchpads.eu

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 cloud[1] 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 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 Europeana, 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

[1] http://plutof.ut.ee

from GenBank[1]. 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 portal[2]. Another example is eBiodiversity portal[3] 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.

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.

[1] http://www.ncbi.nlm.nih.gov/genbank/
[2] http://unite.ut.ee
[3] http://elurikkus.ut.ee

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 organisations[1].

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 advantages[2]:

- 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 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 added[1].

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 is associated with peer-reviewed journal articles, but data associated with non-peer reviewed

[1] http://datadryad.org/pages/organization
[2] http://datadryad.org/pages/faq

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 charges[2].

Tool status

This tool is ready to be used.

A.19 Species Observation System

Main usage, purpose, selected examples

Species Observation System[1], 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 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.

[1] www.artportalen.se and www.artsobservasjoner.no

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 in launched.

A.20 DEIMS: Drupal Ecological Information Management System

Main usage, purpose, selected examples

The International Ecological Information Management System (DEIMS)[1] 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 research (LTER)[2] 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 EnvEurope[3] (LTER-Europe[4])/ 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, ExpeER[5], 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;

[1] https://drupal.org/project/deims
[2] http://www.lternet.edu/
[3] http://www.enveurope.eu/
[4] http://www.lter-europe.net/
[5] http://www.expeeronline.eu/

(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

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 Server[1] 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

[1] http://plazi.org

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 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. 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 material[4].

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 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 Life[5], and observation records to GBIF[6] using Darwin Core Archives. The latter will also be a base to harvest data for EU BON's modeling activities.

Pros and Cons of the tool

Pros

1. The Plazi Treatment Server is a one of its kind. With the US ETF[7] 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-iBiosphere[8], domain specific web sites)

[1] http://plazi.org/?q=GoldenGATE
[2] http://plazi.org/?q=taxonx
[3] https://github.com/tcatapano/TaxPub/releases
[4] 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)
[5] http://eol.org
[6] http://gbif.org
[7] http://biowikifarm.net/v-botknow-test/web/About_BKP
[8] http://www.pro-ibiosphere.eu/

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 RefBank [1] as a reference system for bibliographic references and is working in close collaboration with Zenodo (Biosystematics Literature Community, BLC)[2] to build a repository for articles that are not accessible in digital form. To discover bibliographic references, Refindit [3] 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.
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

[1] http://refbank.org/
[2] https://zenodo.org/collection/user-biosyslit
[3] http://refindit.org/

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 ranges[1]. 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 processor[2] 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 DataUP[3] 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

[1] http://en.wikipedia.org/wiki/Numeric_precision_in_Microsoft_Excel
[2] http://tools.gbif.org/spreadsheet-processor/
[3] http://dataup.cdlib.org/

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.

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) [1]– 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)[2] and exchanges data between the collaboration partners (NCBI[3], DDBJ[4]). INSDC forms the most comprehensive database for all molecular data types and linked metadata.

The Barcode of Life Data Systems (BOLD)[5] - 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/PlutoF[6] – 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.

SILVA[7] – 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)[8] - provides access to the 16S rRNA gene sequence alignment for browsing, blasting, probing, and downloading.

NGS sequences:

Sequence Read Archive (SRA)[9] – 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)[10] – standardising the description, exchange and integration of molecular/genomic data.

[1] http://www.ebi.ac.uk/ena/
[2] http://www.insdc.org/
[3] http://www.ncbi.nlm.nih.gov/
[4] http://www.ddbj.nig.ac.jp/
[5] http://www.boldsystems.org/
[6] http://unite.ut.ee/
[7] http://www.arb-silva.de/
[8] http://greengenes.secondgenome.com/downloads
[9] http://www.ncbi.nlm.nih.gov/sra
[10] http://gensc.org/

Recommendations

1. 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.
2. 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.