Annex 2: List of tested and analyzed data sharing tools (non-exhaustive)

Below are the specifications of the tools surveyed, as to February 2015, with some updates from April 2016. The tools selected in the context of EU BON are available in the main text of the publication and are described in more details. This list is also available on the EU BON Helpdesk website, where it will be regularly updated as needed. Additional lists are available through the GBIF resources page, the DataONE software tools catalogue, the BioVel BiodiversityCatalogue and the BDTracker

A.1 GBIF Integrated Publishing Toolkit (IPT)

Main usage, purpose, selected examples

The Integrated Publishing Toolkit is a free open source software tool written in Java which 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 for submission to a peer-reviewed journal. Currently, the IPT supports three core types of data: checklists, occurrence datasets and sample based data (plus datasets at metadata level only).

The IPT is a community-driven tool. Core development happens at the GBIF Secretariat but the coding, documentation, and internationalization 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 (Robertson et al, 2014). New translations into other languages are welcomed.

The IPT is available for download in both compiled and source code 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 totaling 117.5 million records.

Examples of use of IPT

Darwin Core Archives are required for data harvest to the new VertNet 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) and Canadensys 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 Colombia uses the IPT as a central part of their data publishing model in which it has facilitated publication of primary data.

Pros and Cons of the tool

Pros

1. Publication of two types of biodiversity data: i) primary occurrence data (specimens, observations), ii) species checklists and taxonomies.
2. Integrated metadata editor for publishing data set level metadata.
3. Internationalisation: user interface available in six different languages: English, French, Spanish, Traditional Chinese, Brazilian Portuguese, Japanese; instructions are available for translating the interface.
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 (EML) standard.
8. The tool is supported by good documentation and mailing list; the User Manual is also available in both English and Spanish.

Cons

1. Currently [February 2015], the IPT can only be used for occurrence data sets and checklists. A new version has been released meanwhile which accommodates sample based data.
2. The IPT lacks built-in data validation. Since the IPT is designed to run effectively on a common computer, validating extremely large data sets (+100 million records) becomes an impractical operation. GBIF has been working with its partners, however, to provide pluggable remote validation services on performant data architecture to fill this gap.
3. The IPT depends on server administrators to backup its data. There are plans to address this problem by adding long-term data storage and redundancy to the IPT this year.

Recommendations

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

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

Tool status

The IPT is currently used to publish occurrence data sets and checklists and associated metadata (or metadata documents alone). Work is underway to enhance it for publication of sample-based data. This requires developing a data model for sample-based data that is compatible with the DwC-A model. This will include a new core and extension and a modified instance of the IPT that recognises the new core/extension. A prototype IPT 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.

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, Encyclopedia of Life EOL, and The Data Conservancy (DataONE project), collaborated to develop the GBIF Spreadsheet-Processor, a web application that supports publication of biodiversity data to the GBIF network using pre-configured Microsoft Excel spreadsheet templates. Two main data types are supported: i) occurrence data as represented in natural history collections or species observational data and ii) simple species checklists.

The tool provides a simplified publishing solution, particularly in areas where web-based publication is hampered by low-bandwidth, irregular uptime, and inconsistent access. It enables the user to convert local files to a well-known international standard using an asynchronous web-based process. The user selects the appropriate spreadsheet template, 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.

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 and Pensoft Writing Tool

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

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:

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

Cons:

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

Recommendations

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

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

Tool status

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

A.4 Bibliography of Life

Main usage, purpose, selected examples

The Bibliography of Life platform was developed within the EU FP7 project ViBRANT and consists of three integral tools, RefBank and ReFindit and Biosystematics Literature Repository based at ZENODO/CERN. 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.

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) and the Biodiversity Data Journal (BDJ).

Pros and Cons of the tool

Pros

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

Cons

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

Recommendations

Standards used: MODS (Metadata Object Description Schema), 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.

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

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. Metacat was adopted by the Brazilian Research Program in Biodiversity – PPBio in 2010 and currently stores data collected in 24 different field stations in Brazil. Currently there are more than 400 data packaged available to users in https://ppbiodata.inpa.gov.br/metacatui/#data/page/0. All the data from PPBio is curated and validated by a data manager.

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

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. A data center using Metacat can become DataONE member node with a relatively simple configuration.

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

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. 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 EU BON activities on modelling and sample based sites, particularly for data that is not suitable yet for inclusion in GBIF.

Tool status

This tool is ready to be used.

A.7 DataONE "Slender Node"

Main usage, purpose, selected examples

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

Pros and Cons of the tool

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

Recommendations

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.

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

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 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) initiator of the project) and projects (OneGeology).

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 and 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) is a REST web service based protocol designed for science data. There are multiple software packages which implement DAP, with OPeNDAP Hyrax and THREDDS 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 sub setting and time-slicing capabilities. There is current development to make Hyrax and THREDDS DataONE-enabled.

Recommendations

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

Tool status

These tools are available and ready for use.

A.12 DiGIR

Main usage, purpose, selected examples

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

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

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

Pros and Cons of the tool

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

Recommendations

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

Tool status

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

A.13 TAPIRlink

Main usage, purpose, selected examples

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

Pros and Cons of the tool

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

Recommendations

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

Tool status

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

A.14 BioCASE

Main usage, purpose, selected examples

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

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 conceptual schema, such as ABCD (Access to Biological Collection Data) for the BioCASE network, 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), the Australian Virtual Herbarium (AVH), and GeoCASE.

Pros and Cons of the tool

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

Recommendations

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

Tool status

The BPS is actively maintained and developed by the Informatics research Group of the Botanic Garden and Botanical Museum Berlin-Dahlem. 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 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 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 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 Europea, Index Fungorum) and may be developed collectively further by the users. Currently there are more than 1500 users who develop their private and institutional databases or use analytical tools for biodiversity data. PlutoF cloud also provides data curation, possibilities, including third party annotations to the data from external resources, such as genetic data from GenBank. 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. Another example is eBiodiversity portal 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.

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.

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:

- 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.

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

Recommendations

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

Tool status

This tool is ready to be used.

A.19 Species Observation System

Main usage, purpose, selected examples

Species Observation System, 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.

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) 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) 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 (LTER-Europe)/ 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, and GEO BON networks; (iii) personnel database metadata editor for the creation or editing of the information, relevant to the scientists' contact details and research expertise;

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

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

Pros and Cons of the tool

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

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

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

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

Recommendations

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 provides access to the treatments of taxa. Each taxonomic usage is accompanied minimally by a text that describes the taxon or at least offers some further references, and thus defines the concept in a scientist's mind. There are millions of treatments in the scientific literature, which form an extremely valuable source of information. These treatments are increasingly linked to their underlying data, such as observation data, keys for identifications or other digital objects. There are two bottlenecks to providing semantically useful modern internet access. The first is that a huge number are not even digitally available, or at most

are parts of semantically unstructured PDF-formatted documents. The second is that a substantial amount of the literature is only accessible through a paywall or comes with restrictions on their use. With the increasing wealth of digitised observation records, upon which most of the publications are based, it becomes imperative to provide access to the treatments, to link to them, and to enhance them with links to the material referenced in them.

The treatment repository fulfills this niche. It offers with Golden Gate and respective XML schemas (TaxonX, TaxPub) 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.

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 (EOL), and observation records to GBIF using Darwin Core Archives. The latter will also be a base to harvest data for EU BON's modelling activities.

Pros and Cons of the tool

Pros

1. The Plazi Treatment Server is a one of its kind. With the US ETF 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, domain specific web sites)
3. Currently 34000 treatments from 2700 documents are available.
4. New technical requests can be met quickly, and Plazi has in recent years been on the forefront to build interfaces to import data into GBIF and EOL.
5. Plazi uses RefBank as a reference system for bibliographic references and is working in close collaboration with Zenodo (Biosystematics Literature Community, BLC) to build a

repository for articles that are not accessible in digital form. To discover bibliographic references, Refindit 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 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. 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 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 which allows Excel users to document data in Excel (including at least populating standard Dublin Core metadata fields and checking Excel documents for compliance with best practices). DataUP works as an ActiveX add-in for Excel on Windows and is available as a web site for all Excel users. DataUP can also upload data to the ONEShare member node of DataONE. In principle, a version of DataUP can be created which enables upload to another data repository which implements the DataONE Tier 3 (authenticated write) member node API.

Recommendations

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

Tool status

The tool is available and ready for use.

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

The Barcode of Life Data Systems (BOLD) - 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 – 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 – 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) - provides access to the 16S rRNA gene sequence alignment for browsing, blasting, probing, and downloading.

NGS sequences:

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.

Recommendations

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

Bibliographic References

Robertson T, Döring M, Guralnick R, Bloom D, Wieczorek J, Braak K, et al. (2014) The GBIF Integrated Publishing Toolkit: Facilitating the Efficient Publishing of Biodiversity Data on the Internet. PLoS ONE 9(8): e102623. doi:10.1371/journal.pone.0102623