

Pensoft Data Publishing Policies and Guidelines for Biodiversity Data

Lyubomir Penev^{1,7}, Daniel Mietchen², Vishwas Chavan³, Gregor Hagedorn⁴,
David Remsen³, Vincent Smith⁵, David Shotton⁶

1 *Institute of Biodiversity and Ecosystem Research, Sofia, Bulgaria* **2** *Science 3.0* **3** *Global Biodiversity Information Facility, Copenhagen, Denmark* **4** *Julius Kühn-Institute, Königin-Luise-Straße 19, 14195 Berlin, Germany* **5** *The Natural History Museum, Cromwell Road, London, UK* **6** *Image Bioinformatics Research Group, Department of Zoology, University of Oxford, UK* **7** *Pensoft Publishers, Sofia, Bulgaria*

Citation: Penev L, Mietchen D, Chavan V, Hagedorn G, Remsen D, Smith V, Shotton D (2011). Pensoft Data Publishing Policies and Guidelines for Biodiversity Data. Pensoft Publishers, http://www.pensoft.net/J_FILES/Pensoft_Data_Publishing_Policies_and_Guidelines.pdf

First published: 26th of May 2011. Implemented by Pensoft Publishers.

This document is published under [Creative Commons Attribution License CC-BY](https://creativecommons.org/licenses/by/4.0/).

TABLE OF CONTENT

DATA PUBLISHING POLICIES	2
Data Publishing in a Nutshell	2
What is a Dataset?	3
Why Publish Data.....	4
Data Publishing Licenses.....	5
How to Publish Data	7
Open Data Repositories	8
Publishing Data with GBIF through the Integrated Publishers Toolkit (IPT) and Darwin Core Archive (DwC-A) Format	8
Depositing Data in the Dryad Data Repository	10
Depositing Data on Phylogenies with TreeBASE.....	11
Depositing Genome Data with GenBank and Barcode of Life	12
Other Repositories	12
HOW TO CITE DATA IN PENSOFT JOURNAL ARTICLES.....	14
GUIDELINES FOR AUTHORS	17
Data Published within Supplementary Information Files	17
Data Published in Association with a Stand-Alone Scholarly Publication (a Data Paper)	19
What is a “Data Paper”	19
How to Write and Submit a Data Paper	20

Data Papers Describing Primary Biodiversity Data	21
Structure of Data Papers Describing Primary Biodiversity Data	23
Generation of Data Paper manuscripts using the GBIF Integrated Publishing Toolkit (IPT)	25
Data Papers Describing Ecological and Environmental Data.....	27
Data Papers Describing Genome Data	28
Barcode Data Release Papers	28
Data Papers Describing Software Tools.....	30
GUIDELINES FOR REVIEWERS OF DATA PAPERS	32
Quality of the Manuscript.....	33
Quality of the Data	33
Consistency between Manuscript and Data.....	33
ACKNOWLEDGEMENTS	34
REFERENCES	34

DATA PUBLISHING POLICIES

Data Publishing in a Nutshell

The current document describes the data publishing policies of Pensoft Publishers, with an emphasis on biodiversity and biodiversity-related ecological data. Publishing the data associated with research articles is strongly encouraged in all Pensoft's journals.

Data publishing in this digital age is the act of making data available on the Internet, so that they can be downloaded, analysed, re-used and cited by people and organisations other than the creators of the data. This can be achieved in various ways. In the broadest sense, any upload of a dataset onto a freely accessible website could be regarded as “data publishing”. There are, however, several issues to be considered during the process of data publication, including:

- Data hosting, long-term preservation and archiving
- Documentation and metadata
- Citation and credit to the data authors
- Licenses for publishing and re-use
- Data interoperability standards
- Format of published data
- Software used for creation and retrieval
- Dissemination of published data

This Data Publishing Policies document describes some general concepts, including a definition for datasets, incentives to publish data, and methods and licenses for data publishing. Further, it defines and compares the two main routes for data publishing, namely as [supplementary information files](#) to research articles, which may be made

available directly by the publisher, or published in a specialized data repository with a link to the research article or a [Data Paper](#), i.e. a specific, stand-alone publication describing a particular dataset or a collection of datasets.

More detailed instructions on how to prepare data for publication are listed below under the [Guidelines for Authors](#). The [Guidelines for Reviewers](#) section describes the main criteria for evaluation of data during the peer-review and editorial process. Special attention is given to existing standards, protocols and tools to facilitate data publishing, such as the GBIF [Integrated Publishing Toolkit \(IPT\)](#) and the [Darwin Core Archive \(DwC-A\)](#).

A separate section describes some of the leading data hosting/indexing [infrastructures and repositories](#) for biodiversity and ecological data.

What is a Dataset?

A dataset is understood here as a digital collection of logically connected facts (observations, descriptions or measurements), typically structured in tabular form as a set of records, with each record comprising a set of fields, and recorded in one or more computer data files that together comprise a data package. Certain types of research datasets, e.g. a video recording of animal behaviour, will not be in tabular form, although analyses of such recordings may be. Within the domain of biodiversity, a dataset can be any discrete collection of data underlying a paper – e.g., a list of all species occurrence data published in the paper, data tables from which a graph or map is produced, digital images or videos that are the basis for conclusions, an appendix with morphological measurements, or ecological observations.

More generally, with the development of XML-based publishing technologies, the research and publishing communities are coming to a much wider definition of data, proposed in the [BMC position statement](#) (2010): “the raw, non-copyrightable facts provided in an article or its associated additional files, which are potentially available for harvesting and re-use”.

As these examples illustrate, while the term „dataset“ is convenient and widely used, its definition is vague. Data repositories such as [Dryad](#), wishing for precision, do not use the term „dataset“. Instead, they describe **data packages** to which metadata and unique identifiers are assigned. Each data package comprises one or more related **data files**, these being data-containing computer files in defined formats, to which unique identifiers and metadata are also assigned. Nevertheless, the term “dataset” is used below except where a more specific distinction is required

For practical reasons, we propose a clear distinction between *static data* that represent specific completed compilations of data upon which the analyses and conclusions of a given scientific paper may be based, and *curated data* that belong to a large data collection (usually called a ‘database’) with ongoing goals and curation, for example the various bioinformatics databases that curate nucleotide sequences. Both forms are of strong potential scientific interest and application. Where a static dataset

is inextricably linked to a scientific paper, the data publisher must assure consistent and secure access to it on the same time-scale as the text content of the digital article. As a consequence, it is not permissible to upload a new version of such data in ways that would replace the original, unless strict versioning is undertaken and the reader of the published article has easy access to the original version of the data resource as well as to updated versions.

Curated data, on the other hand, are usually hosted on external servers or in data hosting centres. A primary goal of the data publishing process in this case is to guarantee that these data are properly described, up to date, available to others under appropriate licensing schemes, peer-reviewed, and where appropriate linked from a research article or a [Data Paper](#) at the time of publication. Especially in cases where the long-term viability of the curated project may be insecure (e.g. in the case of grant-funded projects), the publisher may in addition support the publication of a dated and versioned copy of such data (with the option to update these with another version later one, keeping access to all versions).

Why Publish Data

Data publishing becomes increasingly important and already affects the policies of the world's leading science funding frameworks and organizations (see for example the [NSF Data Management Plan Requirements](#), or [Riding the Wave \(How Europe Can Gain From the Rising Tide of Scientific Data\)](#) report submitted to the European Commission in October 2010. More generally, the concept of “open data” is described in the [Protocol for Implementing Open Access Data](#), the [Open Knowledge/Data Definition](#), the [Panton Principles for Open Data in Science](#), and the [Open Data Manual](#).

There are several incentives for authors and institutions to publish data:

- There is a widespread conviction that data produced using public funds should be regarded as a common good, and should be openly published and made available for inspection, interpretation and re-use by third parties.
- Open data increases transparency and the overall quality of science; published datasets can be re-analyzed and verified by others;
- Published data can be cited and re-used in the future, either alone or in association with other data;
- Open data can be integrated with other datasets across both space and time;
- Data integration increases recognition and opportunities for collaboration;
- Open data increases the potential for interdisciplinary research, and for re-use in new contexts not envisaged by the data creator;
- Duplication of data-collecting efforts and associated costs will be reduced;
- Published data can be indexed and made discoverable, browsable and searchable through internet services (e.g. Web search engines) or more specific infrastructures (e.g., [GBIF](#) for biodiversity data);

- Collection managers can trace usage and citations of digitized data from their collections;
- Data creators, and their institutions and funding agencies, can be credited for their work of data creation and publication through the conventional channels of scholarly citation; priority and authorship is achieved in the same way as with a publication of a research paper;
- Datasets and their metadata, and any related Data Papers, may be inter-linked into Research Objects, to expedite and mutually extend their dissemination, to the benefit of the authors, other scientists in their fields, and society at large;
- Published data may be structured as ‚Linked Data‘, by which term is meant data accessible using [RDF, the Resource Description Framework](#), one of the fundamentals of the semantic web. Since RDF descriptions are based on publicly available ontology terms, ideally derived from a limited number of complementary non-overlapping ontologies, this permits automated data integration, since data elements from different sources have built-in syntactic and semantic alignment. Additionally, since the underlying ontologies have a logical structure, unlike simple controlled vocabularies, which for example includes subsumption („is_a“) hierarchies (so that tigers and cheetahs are defined as members of the family Felidae, and tigers and elephants are known to be mammals), this has the potential to facilitate reasoning across diverse and distributed datasets.

Data Publishing Licenses

One of the basic postulates of the [Panton Principles](#) is that data publishers should define clearly the license or waiver under which the data are published, so re-use rights are clear to potential users. They recommend use of the most liberal licenses, or of public domain waivers, to prevent legal and operational barriers for data sharing and integration. For clarity, we list here the [short version of the Panton Principles](#):

1. When publishing data, make an explicit and robust statement of your wishes.
2. Use a recognized waiver or open publication license that is appropriate for data.
3. If you want your data to be effectively used and added to by others, it should be fully „open“ as defined by the [Open Knowledge/Data Definition](#) – in particular, non-commercial and other restrictive clauses should not be used.
4. Explicit dedication of data underlying published science into the public domain via PDDL or CC-Zero is strongly recommended and ensures compliance with both the [Science Commons Protocol for Implementing Open Access Data](#) and the [Open Knowledge/Data Definition](#).

A variety of waivers and licenses, that are specifically designed for and appropriate for the treatment of data, are described in Table 1.

Table 1. Data publishing licences recommended by Pensoft.

Data publishing license	URL
Open Data Commons Attribution License	http://www.opendatacommons.org/licenses/by/1.0/
Creative Commons CC-Zero Waiver	http://creativecommons.org/publicdomain/zero/1.0/
Open Data Commons Public Domain Dedication and Licence	http://www.opendatacommons.org/licenses/pddl/1-0/

The default data publishing license used by Pensoft is the [Open Data Commons Attribution License \(ODC-By\)](#), which is a license agreement intended to allow users to freely share, modify, and use the published data(base), provided that the data creators are attributed (cited or acknowledged). This ensures that those who publish their data receive the academic credit that is their due.

As an alternative, the other public domain licenses, namely the [Creative Commons CC0](#) (also cited as “CC-Zero” or “CC-zero”) and the [Open Data Commons Public Domain Dedication and Licence \(PDDL\)](#), are also **STRONGLY** encouraged for use in Pensoft journals. According to the [CC0 license](#), “the person who associated a work with this deed has dedicated the work to the public domain by waiving all of his or her rights to the work worldwide under copyright law, including all related and neighboring rights, to the extent allowed by law. You can copy, modify, distribute and perform the work, even for commercial purposes, all without asking permission.”

Publication of data under a non-attribution waiver such as CC0 avoids potential problems of „attribution stacking“ when data from several sources are aggregated for re-use, particularly if this re-use is undertaken automatically. In such cases, while there is no *legal* requirement to provide attribution to the data creators, the norms of academic citation best practice for fair use still apply, and those who re-use the data should reference the data source, as they would reference others’ research articles.

The Attribution-ShareAlike [Open Data Commons Open Open Database License \(OdbL\)](#) is **NOT** recommended for use in Pensoft’s journals, although it may be used as an exception in particular cases.

Many widely recognized open access licenses are intended for text-based publications on which copyright pertains, and are not intended for, and are not appropriate for, data or collections of data which do not carry copyright. Creative Commons licenses apart from CC-Zero (e.g., [CC-BY](#), [CC-BY-NC](#), [CC-BY-NC-SA](#), [CC-BY-SA](#), etc.) as well as [GFDL](#), [GPL](#), [BSD](#) and similar licenses widely used for open source software, are **NOT** appropriate for data, and their use for data associated with Pensoft journal articles is **strongly discouraged**.

Authors should explicitly inform the publisher if they want to publish data associated with a Pensoft journal article under a license that is different from the [Open Data Commons Attribution License \(ODC-By\)](#).

Any set of data published by Pensoft, or associated with a journal article published by Pensoft, must always clearly state its licensing terms in both a human-readable and a machine-readable manner.

Where data are published by a public data repository under a particular license, and subsequently associated with a Pensoft research article or [Data Paper](#), Pensoft journals will accept that repository license as the default for the published datasets.

Images, videos and similar 'artistic works' are automatically covered by copyright, unless specifically placed in the public domain by use of a public domain waiver such as CC0. Where copyright is retained by the creator, such multimedia entities can still be published under an open data attribution license, while their metadata can be published under a CC0 waiver.

Databases can contain a wide variety of types of content (images, audiovisual material, and sounds, for example, as well as tabular data, which might all be in the same database), and each may have a different license, which must be separately specified in the content metadata. Databases may also automatically accrue their own rights, such as the European Union Database Right, although no equivalent database right exists in the USA. In addition, the contents of a database, or the database itself, can be covered by other rights not addressed here (such as private contracts, trademark over the name, or privacy rights / data protection rights over information in the contents). Thus authors are advised to be aware of potential problems for data reuse from databases, and to clear other rights before engaging in activities not covered by the respective license.

How to Publish Data

At Pensoft, data can be published either (a) as supplementary files related to a research paper or (b) in association with a stand-alone description of the data resource (a [Data Paper](#)). Within these two main routes, Pensoft supports the following methods for data publishing:

- Supplementary data files (max. 20 MB each) related to and published with a research article, and stored on the publisher's website. Guidelines on the format are available [here](#).
- Primary biodiversity data (species-by-occurrence records) published through the GBIF [Integrated Publishing Toolkit \(IPT\)](#). A format that is strongly encouraged for the publication of biodiversity and species occurrence data, checklists and their associated metadata is the [Darwin Core Archive \(DwC-A\) format](#).
- Datasets other than primary biodiversity data (e.g., [ecological observations](#), [environmental data](#), [genome data](#) and other data types) preserved in certified institutional or international [data repositories](#) and linked permanently to a research article or a Data Paper.

Best practice recommendations:

- Deposition of data in an established international repository is **always to be preferred** to supplementary files published on a journal's website.

- Occurrence-by-species records should be deposited through GBIF IPT.
- Genomic data should be deposited at GenBank, either directly or via an affiliated repository, e.g. [Barcode of Life Data Systems \(BOLD\)](#).
- Phylogenetic data should be deposited at TreeBASE, either directly or through the Dryad Data Repository.
- All other biological data, including heterogeneous datasets, should be deposited in the Dryad Data Repository.
- Repositories not mentioned above, including institutional repositories, may be used at the discretion of the author.
- [Digital Object Identifiers \(DOIs\)](#) or other persistent links (URLs) to the data deposited in repositories, as well as the name of the repository, **should always be published** in the paper describing that data resource.

Open Data Repositories

Open data repositories (public databases, data warehouses, data hosting centres) are subject- or institution-oriented infrastructures, usually based at large national or international institutions. These provide data storage and preservation according to widely accepted standards, and provide free access to their data holdings for anyone to use and re-use under the minimum requirement of attribution, or under an open data waiver such as the CC-Zero waiver.

There are several directories of data repositories relevant to biodiversity and ecological data, such as those listed in the Open Access Directory (http://oad.simmons.edu/oadwiki/Data_repositories#Biology) and in Table 1 of Tessen and Patterson's (2011) [Data Issues in Life Sciences White Paper](#) (932 kB PDF document).

Such repositories could be used to host data associated with a published Data Paper, as explained below. For their own data, authors are advised to use an internationally recognised, trusted (normally ISO-certified), specialized repository (see [Krump 2011](#)). The following repositories and databases fall into that category.

Publishing Data with GBIF through the Integrated Publishers Toolkit (IPT) and Darwin Core Archive (DwC-A) format

The [Global Biodiversity Information Facility \(GBIF\)](#) was established in 2001, and is now the world's largest multilateral initiative for enabling free and open access to biodiversity data via the Internet. It comprises a network of 55 countries and 47 international organisations that contribute to its vision of 'a world in which biodiversity information is freely and universally available for science, society, and a sustainable future'. GBIFs' mission is 'to be the foremost global resource for biodiversity information, and engender smart solutions for environmental and

human well-being. The GBIF network facilitates access to over 276 million primary biodiversity data records contributed by 316 data publishers across the globe (as on May 2011).

GBIF is not a repository in the strict sense, but a distributed network of data publishers and local data hosting centers which discover and publish data employing tools based on community-agreed standards for exchange/sharing of primary biodiversity data. At global scale, discovery and access to data is facilitated through the GBIF data portal (<http://data.gbif.org/>). Pensoft facilitates discovery and publishing of data and metadata to the GBIF network through Pensoft's IPT Data Hosting Center (<http://ipt.pensoft.net/ipt/>), that is based on the GBIF Integrated Publishing Toolkit (IPT) (<http://code.google.com/p/gbif-providertoolkit/>). A list of IPT installations used by other data publishers can be accessed at: <http://tools.gbif.org/data-paper-authoring/>.

The *Darwin Core Archive* (DwC-A) is an international biodiversity informatics data standard and the preferred format for publishing data through the (GBIF) network. Each Darwin Core Archive consists of at least three files:

1. One or more data files keeping all records of the particular dataset in a tabular format such as a comma-separated or tab-separated list;
2. The archive descriptor (meta.xml) file describing the individual data file columns used, as well as their mapping to DwC terms; and
3. A metadata file describing the entire dataset which GBIF recommends to be based on EML (*Ecological Metadata Language 2.1.1*).

The format is defined in the *Darwin Core Text Guidelines*. Darwin Core is no longer restricted to occurrence data, and together with the more generic *Dublin Core metadata standard* (on which its ideas are based), it is used by GBIF and others to encode metadata about organism names, taxonomies and species information.

GBIF has produced a series of documents and supporting tools that focus primarily on Darwin Core publishing. They are divided into three profiles, each of which represents a series of documents based on the different content types on which GBIF focuses:

- Primary biodiversity data: <http://www.gbif.org/informatics/primary-data/publishing/>
- Checklists: <http://www.gbif.org/informatics/name-services/publishing/>
- Resource metadata: <http://www.gbif.org/informatics/discoverymetadata/publishing/>

In addition to the GBIF Integrated Publishing Toolkit, there are two additional tools developed for producing Darwin Core Archives:

1. A suite of MS Excel Templates (<http://tools.gbif.org/spreadsheet-processor/>), which are coupled with a web service that processes completed files and returns a validated Darwin Core Archive. Templates exist for primary biodi-

versity data, simple checklists, and EML metadata. See <http://tools.gbif.org/spreadsheet-processor/> for further details.

2. Darwin Core Archive Assistant (<http://tools.gbif.org/dwca-assistant/>) is a tool that composes an XML metafile, the only XML component of a Darwin Core Archive. It displays a drop-down list of Darwin Core and extension terms, accessed dynamically from the GBIF registry, and displays these to the user who describes the data files. This allows Darwin Core Archives to be created for sharing without the need to install any software. See <http://tools.gbif.org/dwca-assistant/> for details.

The Darwin Core Archive (DwC-A) files can be used to publish data underlying any taxonomic revision or checklist through IPT or as supplementary files (see a [sample paper](#) by Talamas et al. (2011)). It also can be used to publish species occurrence data. The publication of large datasets in the form of [Data Papers](#) is also supported.

Darwin Core Archive files can also be generated from data uploaded on IPT and then published as a zipped supplementary file, associated with a research article.

Depositing Data in the Dryad Data Repository

Pensoft encourages authors to deposit data underlying biological research articles in the [Dryad data repository](#) which is being used for this purpose by many other journals and societies, in cases where no suitable more specialized public data repository (e.g. GBIF for species-by-occurrence data and taxon checklists, or GenBank for genome data) exists.

Data can be deposited with Dryad either before or at the time of submission of the manuscript to the journal, or after the manuscript acceptance but before submission of the finally revised, ready-for-layout version for publication. One or more individual data files are aggregated into a data package for submission.

Once you deposit your data package, it receives a unique and stable identifier, namely a DataCite DOI. Individual data files within this package are given their own DOIs, based on the package DOI, as do subsequent versions of these data files, as explained at https://www.nescent.org/wg_dryad/DOI_Usage. You should include appropriate Dryad DOIs in the final text of the manuscript, both in the in-text citation statement and in the formal data reference in your paper's reference list, as explained and exemplified above. This is very important, since if the data DOI does not appear in the final published article, that greatly weakens its connection to the underlying data.

More information about depositing data in Dryad can be found at <http://www.datadryad.org/repo/depositing>, and in a 2-minute video on [SciVee](#) (<http://www.scivee.tv/node/26563>; doi: [10.4016/26563.01](https://doi.org/10.4016/26563.01)).

Advantages of depositing data in Dryad include:

- **Visibility:** Making your data available online (and linking it to the publication) provides a independent way for others to discover your work.
- **Citability:** all data you deposit will receive a persistent, resolvable identifier that can be used in a citation, as well as listed on your CV.
- **Workload reduction:** if you receive individual requests for data, you can simply direct them to the files in Dryad.
- **Preservation:** your data files will be permanently and safely archived in perpetuity.
- **Impact:** other researchers have more opportunities to use and cite your work.

Pensoft supports Dryad and its goal of enabling authors to publicly archive sufficient data to support the findings described in their journal article. Dryad is a safe, sustainable location for data storage and there are no restrictions on data format. Note that data deposited to Dryad is made available for reuse using the [Creative Commons CC0](#) waiver, detailed above.

You may wish to take a look at some example data packages in Dryad to see how data packages related to published articles are displayed, such as [doi:10.5061/dryad/13.10](https://doi.org/10.5061/dryad/13.10), [doi:10.5061/dryad.7994](https://doi.org/10.5061/dryad.7994) and [doi:10.5061/dryad.8682](https://doi.org/10.5061/dryad.8682).

If some of the data files in a data package are more suitable for other more specialized repositories, for example TreeBase for phylogenetic trees (see below), they can be submitted to Dryad, which will then forward them to TreeBase for hosting, while retaining within Dryad a copy of the metadata and a link to the original data file.

Data deposited to Dryad in association with Pensoft journal articles will be made public immediately on publication of the article, but can be made available to publication editors and reviewers in advance of publication using a special access code.

Depositing Data on Phylogenies with TreeBASE

[TreeBASE](#) is a repository of phylogenetic information, specifically user-submitted phylogenetic trees and the data used to generate them. TreeBASE accepts all types of phylogenetic data (e.g., trees of species, trees of populations, trees of genes) representing all biotic taxa. Data in TreeBASE are exposed to the public if they are used in a publication that is in press or published in a peer-reviewed scientific journal, book, conference proceedings, or thesis. Data used in publications that are in preparation or in review can be submitted to TreeBASE but are embargoed until after publication, and only available before publication to the publication editors or reviewers using a special access code.

Depositing Genome Data with GenBank and Barcode of Life

The [NCBI GenBank](#) is the National Institutes of Health (NIH) genetic sequence database, an annotated collection of all publicly available DNA sequences. GenBank is part of the [International Nucleotide Sequence Database Collaboration](#), which comprises the DNA DataBank of Japan (DDBJ), the European Bioinformatics Institute (EBI) which is part of the European Molecular Biology Laboratory (EMBL), and GenBank at NCBI. These three organizations exchange data on a daily basis. An example of a GenBank record for a *Saccharomyces cerevisiae* gene may be viewed [here](#). There are several options for [submitting data to GenBank](#).

The [Barcode Submission Tool](#) is a web-based tool for the submission of GenBank sequences for [Barcode of Life](#) projects; currently, only mitochondrial cytochrome coxidase subunit I (COI) genes are being accepted with this tool.

The [Consortium for the Barcode of Life](#) (CBOL) urges participants in major DNA barcoding initiatives to consider submitting “**BARCODE data release papers**” for possible publication in academic journals. A BARCODE data release paper is a short manuscript that announces and documents the public deposit to a member of the International Nucleotide Sequence Data Collaboration (INSDC, which includes GenBank, EMBL, and DDBJ) of a significant body of data records that meet the [BARCODE data standards](#).

Other Repositories

- [PANGAEA](#). The information system PANGAEA is operated as an Open Access library aimed at archiving, publishing and distributing georeferenced data from earth system research. The description of each dataset is always visible and includes the name of the principle investigator (PI), who may be contacted to request data access. Each dataset can be identified and cited by using a [DOI](#). Data are archived as supplements to publications or as citable data collections. Citations are available through the portal of the German National Library of Science and Technology ([GetInfo](#)). Archiving follows the [Recommendations of the Commission on Professional Self Regulation in Science](#) (200 kB PDF document) for safeguarding good scientific practice. The policy of data management and archiving follows the [Principles and Responsibilities of ICSU World Data Centers](#) and the [OECD Principles and Guidelines for Access to Research Data from Public Funding](#). PANGAEA is open to any project or individual scientist to archive and publish data. [Data submission can be started here](#).
- The [Knowledge Network for Biocomplexity \(KNB\)](#) is a national network intended to facilitate ecological and environmental research on biocomplexity. For scientists, the KNB is an efficient way to discover, access, interpret, integrate and analyze complex ecological data from a highly-distributed set of field stations, laboratories, research sites, and individual researchers.

- The [National Biological Information Infrastructure](#) is a broad, collaborative program to provide increased access to data and information on United States biological resources. The NBII links diverse, high-quality biological databases, information products, and analytical tools maintained by [NBII partners](#) and other contributors in government agencies, academic institutions, non-government organizations, and private industry.
- [DataBasin](#) is a free system that connects you with spatial datasets, non-technical tools, and a network of scientists and practitioners. One can explore and download a vast library of datasets, connect to external data sources, upload and publish your own datasets, connect to experts, create working groups, and produce customized maps that can be easily shared.
- [DataONE](#) is currently an NSF-funded project that aims at providing the distributed framework, sound management, and robust technologies that enable long-term preservation of diverse multi-scale, multi-discipline, and multi-national observational data. DataONE initially emphasizes observational data collected by biological (genome to ecosystem) and environmental (atmospheric, ecological, hydrological, and oceanographic) scientists, research networks, and environmental observatories.
- The [PaleoBiology Database](#) is a public resource for the global scientific community whose purpose is to provide global, collection-based occurrence and taxonomic data for marine and terrestrial animals and plants of any geological age, as well as web-based software for statistical analysis of the data. The project's wider, long-term goal is to encourage collaborative efforts to answer large-scale paleobiological questions by developing a useful database infrastructure and bringing together large data sets. There is an option to protect data for private use only.
- The Research Collaboratory for Structural Bioinformatics (RCSB)'s [Protein Data Bank](#) (PDB) contains information about experimentally-determined structures of proteins, nucleic acids, and complex assemblies. As a member of the Worldwide Protein Data Bank ([wwPDB](#)), the RCSB PDB curates and annotates PDB data according to agreed standards. See its [data deposition policies and services](#).
- The [Universal Protein Resource \(UniProt\)](#) is a comprehensive resource for protein sequence and annotation data. The UniProt databases are the UniProt Knowledgebase (UniProtKB), the UniProt Reference Clusters (UniRef), and the UniProt Archive (UniParc). The UniProt Metagenomic and Environmental Sequences (UniMES) database is a repository specifically developed for metagenomic and environmental data.
- [INSPIRE](#). Although currently at a prototype stage, the [INSPIRE Geportal](#) is an important spatial data infrastructure that will enable the sharing of environmental spatial information among public sector organisations and better facilitate public access to spatial information across Europe. INSPIRE is not yet open for hosting of data.

HOW TO CITE DATA IN PENSOFT JOURNAL ARTICLES

This section of the Pensoft Data Publishing Policies and Guidelines is based on a draft set of Data Citation Best Practice Guidelines currently being developed for publication by David Shotton, with assistance from colleagues at Dryad and elsewhere (<http://bit.ly/1E2E21>), and on earlier papers concerning data citation mechanisms in biodiversity science published in ZooKeys (Penev et al. 2009a, 2009b).

The well-established norm for citing genetic data is that one simply cited the Genbank identifier (accession number) in the text. Similar usage is also commonplace for items in other bioinformatics databases. Pensoft is not recommending a change in that practice. The following guidelines apply to more heterogeneous research data published in other institutional or subject-specific data repositories, frequently described in related journal articles or Data Papers (see below). They are intended to permit data citations to be treated as ‘first class’ citation objects, on a par with bibliographic citations, and to enable them to be more easily harvested from reference lists, so that those who have made the effort to publish their research data might more easily be ascribed academic credit for their work through the normal mechanisms of citation recognition.

For such data in data repositories, each published data package and each published data file should always be associated with a persistent unique identifier. A Digital Object Identifier (DOI) issued by DataCite should be used wherever possible. If this is not possible, the identifier should be one issued by the data repository or database, and should be in the form of a persistent and resolvable URL. As an example, the use of DOIs in the Dryad Data Repository is explained at https://www.nescent.org/wg_dryad/DOI_Usage.

Data citations may relate either to the author’s own data, or to data created and published by others (“third-party data”). In the former case, the dataset may have been previously published, or may be published for the first time in association with the article that is now citing it. All these types of data should, for consistency, be cited in the same manner.

Generic recommendations

As is the norm when citing another research article, any citation of a data publication, including a citation of one’s own data, should always have two components:

- An **in-text citation statement** containing an **in-text reference pointer** that directs the reader to a formal data reference in the paper’s reference list.
- A formal **data reference** within the article’s reference list.

We recommend that the in-text citation statement *also* contains a separate citation of the research article in which the data were firstly described, if such an article exists, with its own in-text reference pointer to a formal article reference in the paper’s reference list, unless the paper being authored is the one providing that first description of

the data. If the in-text citation statement includes the DOI for the data (a desirable practice), this DOI should always be presented as a dereferenceable URI, as shown below.

The data reference in the article’s reference list should contain the minimal components recommended in the DataCite Metadata Kernel v2.0 specification. In DataCite terms: Creator PublicationYear Title Publisher Identifier; alternatively (but meaning the same thing): Author PublicationYear Title DataRepositoryName DOI. These components should be presented in whatever format and punctuation style the journal specifies for its references.

The following example demonstrates in general terms what is required.

In-text citation:

“This paper uses data from the [*name*] data repository at http://dx.doi.org/***** (Jones *et al.* 2008a), first described in Jones *et al.* 2008b. “

Data reference and article reference in reference list:

“Jones A, Bloggs B, Smith C (2008a). Title of data package. **Repository name. doi:*****.**

Jones A, Saul D, Smith C (2008b). Title of journal article. Journal Volume: Pages. **doi:#####.**”

Note that the authorship and the title of the data package may, for valid academic reasons, differ from those of the author’s paper describing the data – indeed, to avoid confusion of what is being referenced, it is highly desirable that the titles of the data package and of the associated journal article are clearly different.

Pensoft journals require the following formats for data citation:

1. When referring to the author’s own **newly published data**, cited from within the paper in which these data are first described, the citation statement and the data reference should take the following form.
 - The citation statement of data deposition should be included in the body of the paper, in a *separate section* named **Data Resources**, situated after the Material and Methods section.
 - In addition, the formal data reference should be included in the paper’s reference list, using the recommended journal’s reference format

The following dummy example demonstrates what is required.

In-text citation:

“The data underpinning the analysis reported in this paper were deposited on [*date*] in the Dryad Data Repository at <http://dx.doi.org/10.5061/dryad.#####> (Miller *et al.* 2009) and at GBIF, the Global Biodiversity Information Facility, <http://ipt.pensoft.net/ipt/resource.do?r=xxxxxx>.

Data reference in reference list:

“Miller JA, Griswold CE, Yin CM (2009) Locality data for all specimens of the spider families *Theridiosomatidae*, *Mysmenidae*, *Anapidae* and *Symphytognathidae* collected during an inventory of the Gaoligongshan, Yunnan, China, 1998-2007. Data file 1: Microsoft Excel (1997-2003); Data file 2: KML (Keyhole Markup Language) version 2.1 for GoogleEarth. *Dryad Data Repository*. doi:10.5061/dryad.####; and *GBIF, the Global Biodiversity Information Facility*, <http://ipt.pensoft.net/ipt/resource.do?r=xxxxxx>.

Note: If at the time of authoring the article, the DOI for the data package is not known, the author should enter a dummy place-holder for the missing value (as shown), to be replaced with the real value later in the article production process.

2. When acknowledging re-use in the paper of **previously published data** (including the author’s own data) **that is associated with another published peer-reviewed journal article**, the citation and reference should take the same form, except that the full correct DOI should be employed, and that the journal article first describing the data should also be cited.
 - A statement of usage of the previously published data, with citation of the data source(s) and of the related journal article(s), should be placed in a *separate section* named **Data Resources**, situated after the Material and Methods section.
 - In addition, the formal data reference and a formal reference to the related journal article should be included in the paper’s reference list, using the recommended journal’s reference format.

The following real example demonstrates what is required.

In-text citation:

“The data underpinning this analysis were obtained from the Dryad Data Repository at <http://dx.doi.org/10.5061/dryad.829> (Miller 2003a), and were first described by Miller (2003b).”

Data reference and article reference in reference list:

“Miller JA (2003a) Data from: Assessing progress in systematics with continuous jackknife function analysis. *Dryad Data Repository*. doi:10.5061/dryad.829.

Miller JA (2003b) Assessing Progress in Systematics with Continuous Jackknife Function Analysis. *Systematic Biology*, **52**(1), 55-65. doi:10.1080/10635150390132731

3. **When acknowledging re-use of previously published data** (including the author’s own data) **that has NO association with a published research article**, the same general format should be adopted, although a reference to a related journal article clearly cannot be included.

- A statement of usage of previously published data, with citation of the data source(s), should be placed in a *separate section* named **Data Resources**, situated after the Material and Methods section.
- In addition, the formal data reference should be included in the paper's reference list, using the recommended journal's reference format.

The following real example demonstrates what is required.

In-text citation:

“The global seismic monitor data underpinning this analysis were obtained from the GeoForschungsZentrum Potsdam (GFZ). <http://dx.doi.org/10.1594/GFG.GEOFON.gfz2009kciu>. (Geofon operator 2009).”

Data reference in reference list:

“Geofon operator (2009): GEFON event gfz2009kciu (NW Balkan Region). GeoForschungsZentrum Potsdam (GFZ). [doi:10.1594/GFG.GEOFON.gfz2009kciu](https://doi.org/10.1594/GFG.GEOFON.gfz2009kciu).”

GUIDELINES FOR AUTHORS

Data Published within Supplementary Information Files

Online publishing allows an author to provide data sets, tables, video files, or other information as supplementary information files associated with papers, or to deposit such files in one of the repositories described above, greatly increasing the impact of the submission. For larger biodiversity datasets, authors should consider the alternative of submitting a separate [Data Paper](#) (see description below).

Submission of data to a recognised data repository is encouraged as a superior and more sustainable method of data publication than submission as a supplementary information file with an article. Nevertheless, Pensoft will accept supplementary information files if authors wish to submit them with their articles. Details for uploading such files is given in Step 4 of the [Pensoft submission process](#) (available through the “Submit a Manuscript” button after registration on any of the [Pensoft journal websites](#)).

By default, the maximum file size for each supplementary information file that can be uploaded onto the Pensoft web site is 20 MB. If you need more than that, or wish to submit a file type not listed below, please contact us before uploading.

The supplementary information files will not be displayed in the printed version of the article, but will exist as linkable downloadable files in the online version.

When submitting a supplementary information file, the following information should be completed:

- File format (including name and a URL of an appropriate viewer if the format is unusual)
- Title of the supplementary information file. (The authorship will be assumed to be the same as for the paper itself.)
- Description of the data, software listings, protocols or other information contained within the supplementary information file.

All supplementary information files should be referenced explicitly by file name within the body of the article, e.g. “See Supplementary File 1: Movie 1 recording the original data used to perform this analysis”.

Ideally, the supplementary information file formats should not be platform-specific, and should be viewable using free or widely available tools. Suitable file formats are:

For supplementary documentation:

- PDF (Adobe Acrobat; ISO 32000-1)
- HTML (Hypertext Markup Language)
- XML (Extensible Markup Language)

For animations:

- SWF (Shockwave Flash)
- DHTML (Dynamic HTML)

For images:

- GIF (Graphics Interchange Format)
- JPEG/JFIF (JPEG File Interchange Format)
- PNG (Portable Network Graphics)
- SVG (Scalable Vector Graphics)
- TIFF (Tagged Image File Format)

For movies:

- MOV (QuickTime)
- MPG (MPEG)
- OGG (an open and free multimedia container format)

For datasets:

- XLS (Excel spreadsheet)
- CSV (Comma separated values)
- ODS (OpenOffice spreadsheets)

The file names should use the standard file extensions (as in “Supplementary-Figure-1.png”). Please also make sure that each supplementary information file is of a single table, figure, image or movie.

Data Published in Association with a Stand-Alone Scholarly Publication (a Data Paper)

What is a “Data Paper”

A Data Paper is a scholarly journal publication whose primary purpose is to describe a dataset or a group of datasets, rather than to report a research investigation (Chavan and Ingwersen 2009, Chavan and Penev, 2011). As such, it contains facts about data, not hypotheses and arguments in support of those hypotheses based upon data, as found in a conventional research article. Its purposes are three-fold:

- to provide a citable journal publication that brings scholarly credit to data publishers;
- to describe the data in a structured human-readable form, and
- to bring the existence of the data to the attention of the scholarly community.

The description should include several important elements (usually called *metadata*, or “description of data”) that document, for example, how the dataset was collected, which taxa it covers, the spatial and temporal ranges and regional coverage of the data records, provenance information concerning who collected and who owns the data, details of which software was used to create the data, or could be used to view the data, and so on.

Most Pensoft journals have data sections in which Data Papers are published. By 2012, Pensoft will also open a new Biodiversity Data Journal based on an innovative editorial and peer-review platform. Pensoft thus welcomes the submission of Data Papers, that can be indexed and cited like any other research article, thus bringing registration of priority, a permanent publication record, recognition and academic credit to the data creators. In other words, the Data Paper is a mechanism to acknowledge efforts in authoring ‘fit-for-use’ and enriched metadata describing a data resource. The general objective of Data Papers in biodiversity science is to describe all types of biodiversity data resources, including environmental data resources.

An important feature of Data Papers is that they should always be linked to the published datasets they describe, and that link (a URL, ideally resolving a DOI) should be published within the paper itself. Conversely, the metadata describing the dataset held within data archives should include the bibliographic details of the Data Paper once that is published, including a resolvable DOI.

How to Write and Submit a Data Paper

In principle, any valuable dataset hosted in a trusted data repository could be described in a Data Paper, and published following these Guidelines. Each Data Paper consists of a set of elements (sections), some of which are mandatory and some not. An example

for such a list of elements needed to describe primary biodiversity data is available in the section [Data Papers Describing Primary Biodiversity Data](#) below.

A sample Data Paper which can be used as illustration of the concept can be downloaded [here](#).

All claims in a Data Paper should be substantiated by the associated data. If the methodology is standard, please explain in what respects your data are unique and merit a publication in the form of Data Paper.

(Alternatively, if the methodology used to acquire the data differs significantly from established approaches, please consider submitting your data as [supplementary file\(s\)](#) associated with a standard paper (see above) in which these methodologies can be more fully explained.)

At the time of submission of the Data Paper manuscript, the data described should be freely available online in a [public repository](#) under a suitable data license, so that they can be retrieved anonymously for reuse, resampling and redistribution by anyone for any purpose, subject to one condition at most - that of proper attribution using scholarly norms (see the [Data Publishing Licenses](#) and [How to Cite Data](#) sections, above). The repository, or at least one public mirror thereof, should not be under the control of the submitting authors. The relevant data package DOIs or accession numbers, as well as any special instructions for acquiring and re-publishing the data, should be included in the submitted Data Paper manuscript.

The procedures for data retrieval should be described, along with the mechanisms for updating and correcting information. This can be achieved by referencing an existing description if that is up to date, citable in its exact version, and publicly accessible on the web.

All methodological details necessary to replicate the original acquisition of the raw data have to be included in the Data Paper, along with a description of all data processing steps undertaken to transform the raw data into the form in which the data have been deposited in the repository and presented in the paper. Authors should discuss any relevant sources of error and how these have been addressed.

In addition to Data Papers describing new data resources, Data Papers describing existing resources are welcome, as long as these are up to date and their current version is publicly accessible and can be cited. If possible, authors should outline possible re-use cases, taking into account that future uses of the data might involve researchers from different backgrounds, or be undertaken automatically. We encourage the provision of tools to facilitate visualization and reuse of the data.

For primary biodiversity (species-by-occurrence) data, authors are **strongly encouraged** to use the data publishing workflow of the Integrated Publishing Toolkit (IPT) developed by the GBIF, described above. From IPT, data manuscripts can be generated in RTF format directly from the metadata through the “Manage Resources” menu, provided that the respective dataset has already been indexed and properly described by metadata in IPT. The process of data indexing through IPT is described in the [IPT Manual](#). The editorial workflow for Data Papers generated through IPT is described in the next section, [Data Papers Describing Primary Biodiversity Data](#).

For datasets that are not indexed through the IPT - e.g., [genome](#), [ecological or biodiversity-related environmental data](#) - the authors should submit a manuscript written in any text editor, e.g., MS Word or Open Office. The structures of such Data Papers are described below.

Data Papers Describing Primary Biodiversity Data

Primary biodiversity data are the digital text or multimedia data records that detail the instance of an organism – the ‘what, where, when, how and by whom’ of the organism’s occurrence and recording (see <http://www.gbif.org/informatics/primary-data/>).

Currently, the majority of primary biodiversity data consists of species-by-occurrence data records available from published sources and/or [natural history collections](#). Other types of primary biodiversity data that merit publication are [observational data](#) and [multimedia resources in biodiversity](#).

Authoring Metadata through the Integrated Publishing Toolkit (IPT)

The GBIF Integrated Publishing Toolkit (IPT) facilitate authoring of metadata. For this purpose, the [GBIF Metadata Profile \(GMP\)](#) was developed to standardise how biodiversity data resources are indexed and described through the GBIF network. The definitions of the metadata elements are taken from:

- (a) ISO 19139: ISO 19139: North American Profile of ISO19115:2003 – Geographic information – Metadata,
- (b) EML: Ecological Metadata Language (EML) Specification, and
- (c) NCD: Natural Collections Descriptions (NCD): A data standard for exchanging data describing natural history collections.

The GMP elements, together with their descriptions, are listed in the [Structure of Data Papers Describing Primary Biodiversity Data](#) section, below.

The GBIF [Integrated Publishing Toolkit \(IPT\)](#) makes it easy to share three types of biodiversity-related information: primary taxon occurrence data (also known as primary biodiversity data), taxon checklists, and general metadata about data sources. An IPT instance, as well as the data and metadata registered through the IPT, is connected to the Global Biodiversity Resource Discovery System (GBRDS - also known as the GBIF Registry), indexed for consultation via the GBIF network and portal, and made accessible for public use. More information about the GBIF IPT can be found at <http://www.gbif.org/informatics/infrastructure/publishing/>.

The IPT is a server-side software tool that allows users to author metadata, map databases or upload text files that conform to the Darwin Core standard, to install extensions and vocabularies to allow for richer content and, ultimately, to register datasets for publication and sharing through GBIF. IPT operators undertake the responsibility of running an Internet server which should be maintained, namely that is it should remain online

and be addressable. Any set of metadata can be downloaded from any IPT (version 2.0.2+) into RTF format in the form of a Data Paper manuscript, and can then be submitted for publication through the normal journal submission and peer review process.

Therefore data authors have the following options:

- Install and run an IPT instance, registering it with GBIF.
- Use an account on the Pensoft IPT Data Hosting Centre at <http://ipt.pensoft.net/ipt/>; please ask the Journal's Editorial Office to open an account for you.
- Approach any other existing IPT operator and seek to host data through them.

A list of existing IPT installations supporting the authoring of Data Papers is available at <http://tools.gbif.org/data-paper-authoring/>, and the IPT user manual is available at: <http://code.google.com/p/gbif-providertoolkit/wiki/IPT2ManualNotes>.

Once you have decided to publish your data and generate a Data Paper manuscript through the GBIF IPT, please consider the following simple rules:

1. The metadata within one IPT resource must describe **only one core set** of biodiversity data (e.g., either occurrence data or a taxon checklist), that is uploaded through the IPT, indexed in the GBIF Data Portal, and published in Darwin Core Archive Format. The IPT will generate for you an RTF manuscript that will describe the core dataset. The link to the core dataset will appear in your manuscript under the heading "Data published through GBIF".
2. Additional datasets that relate to the core one, e.g., ecological or environmental data, can also be briefly described within the same resource and linked through the "External links" field of the IPT. Those datasets will appear in the section "External datasets" of your manuscript.
3. It is possible to open a resource and enter the respective metadata for it without upload of a core dataset. This option should be used to describe a dataset that has been already uploaded on a repository (e.g., data previously indexed through GBIF for which you have a GBIF link). In this case, you will need to insert the link(s) to the dataset(s) in the "External links" field of the IPT.
4. The option explained in point 3 above could also be used to describe non-digitized natural history collections.
5. We **strongly recommend** uploading a **core set of biodiversity data** through the IPT Darwin Core Archive format, which facilitates not only publication of your data but also its easy sharing and integration with other data, hence its re-use and dissemination.

Structure of Data Papers Describing Primary Biodiversity Data

The structure of a Data Paper largely resembles that of a standard research paper. However, it must contain several specific elements. These elements are listed in Ta-

ble 2 below, which describes the general structure of the Data Paper (left column) mapped to the metadata elements (right column), and is intended to serve as a human-readable model for any Data Paper manuscript, whether generated through the IPT or written independently in a text editor. A sample Data Paper which can be used as an illustration of the concept can be downloaded from http://www.pensoft.net/J_FILES/temp/Pensoft_Data_Publishing_Policies_and_Guidelines.pdf

Table 2. Structure of Data Paper and its mapping from GBIF IPT Metadata Profile elements.

Section/Sub-Section headings of the Data Paper describing primary biodiversity data	Mapping from GBIF IPT Metadata Profile elements, and formatting instructions
<TITLE>	Derived from 'title' element. Format: a centred sentence without a full stop (.) at the end.
<Authors>	Derived from the 'creator', 'metadataProvider' and 'AssociatedParty' elements. From these elements, combinations of 'first name' and 'last name' are derived, separated by commas(.). Corresponding affiliations of the authors are denoted with numbers (1, 2, 3,...) superscripted at the end of each last name. If two or more authors share the same affiliation, it will be denoted by use of the same superscript number. Format: centred.
<Affiliations>	Derived from the 'creator', 'metadtaProvider' and 'AssociatedParty' elements. From these elements, combinations of 'Organisation Name', 'Address', 'Postal Code', 'City', 'Country' constitute the affiliation.
<Corresponding authors>	Derived from the 'creator' and 'metadataProvider' elements. From these elements 'first name', 'last name' and 'email' are derived. Email addresses are written in parentheses (). In a case of more than one corresponding author, these are separated by commas. If both creator and metadataProvider is the same, the creator is denoted as the corresponding author. Format: indented from both sides.
<Received, Revised, Accepted, and Published dates>	These will be inserted manually by the Publisher of the Data Paper, to indicate the dates of original manuscript submission, revised manuscript submission, acceptance of manuscript and publication of the manuscript as a Data Paper in the journal.
<Citation>	This will be inserted manually by the Publisher of the Data Paper. It will be combination of Authors, Year of Data Paper publication (in parentheses), Title, Journal Name, Volume, Issue number (in parentheses), and DOI of the Data Paper, in both native and resolvable HTTP format.
<Abstract>	Derived from the 'abstract' element. Format: indented from both sides.
<Keywords>	Derived from 'keyword' element. Keywords are separated by commas (,).
<Introduction>	Free text.
<Taxonomic Coverage>	Derived from the Taxonomic Coverage elements. These elements are 'general taxonomic coverage description', 'taxonomicRankName', 'taxonomicRankValue' and 'commonName'. 'TaxonomicRankName' and 'taxonomicRankValues'.
<Spatial Coverage>	Derived from the Spatial Coverage elements. These elements are 'general geographic description', 'westBoundingCoordinate', 'eastBoundingCoordinate', 'northBoundingCoordinate', 'southBoundingCoordinate'.

Section/Sub-Section headings of the Data Paper describing primary biodiversity data	Mapping from GBIF IPT Metadata Profile elements, and formatting instructions
<Temporal Coverage>	Derived from the Temporal Coverage elements namely, 'beginDate' and 'endDate'.
<Project Description>	Derived from project elements as described in the GBIF Metadata Profile. These elements are 'title' of the project, 'personnel' involved in the project, 'funding' sources, 'StudyAreaDescription/descriptor', and 'designDescription'.
<Natural Collections Description>	Derived from project NCD elements as described in the GBIF Metadata profile. These elements are 'parentCollectionIdentifier', 'collectionName', 'collectionIdentifier', 'formationPeriod', 'livingTimePeriod', 'specimenPreservationMethod', and 'curatorialUnit'.
<Methods>	Derived from methods elements as described in the GBIF Metadata Profile. These elements are 'methodStep/description', 'Sampling/StudyExtent/description', 'sampling/samplingDescription', and 'qualityControl/description'.
<Dataset descriptions>	Derived from physical and other elements as described in the GBIF Metadata Profile. These elements are 'objectName', 'characterEncoding', 'formatName', 'formatVersion', 'distribution/online/URL', 'pubDate', 'language', and 'intellectualRights'.
<Additional Information>	Derived from 'additionalInfo' element.
<References>	Derived from 'citation' element. This element assumes a reference to a research article or a web link, cited in the metadata description.

For information on the meaning and use of each metadata element implemented in the IPT to describe primary (species-by-occurrence) biodiversity data, please see [Detailed descriptions of the GBIF Metadata Profile \(GMP\) elements](#). Descriptions of the elements are also available in the form of pop-up help in the metadata editor of IPT.

Generation of Data Paper Manuscripts using the GBIF Integrated Publishing Toolkit (IPT)

As described in the previous section, data creators will be able to author Data Paper manuscripts in various ways. However, to lower the technical barrier and make the process easy-to-adapt, a conversion tool to automatically export metadata to a manuscript available in IPT 2.0.2+ at the click of a button. The step-by-step process in generation of a Data Paper manuscript from the metadata is depicted in Figure 1, and is described below.

1. The Data Creator completes the metadata for a biodiversity resource dataset using the metadata editor in IPT 2.0.2+. IPT assigns the Persistent Identifier to the authored metadata.

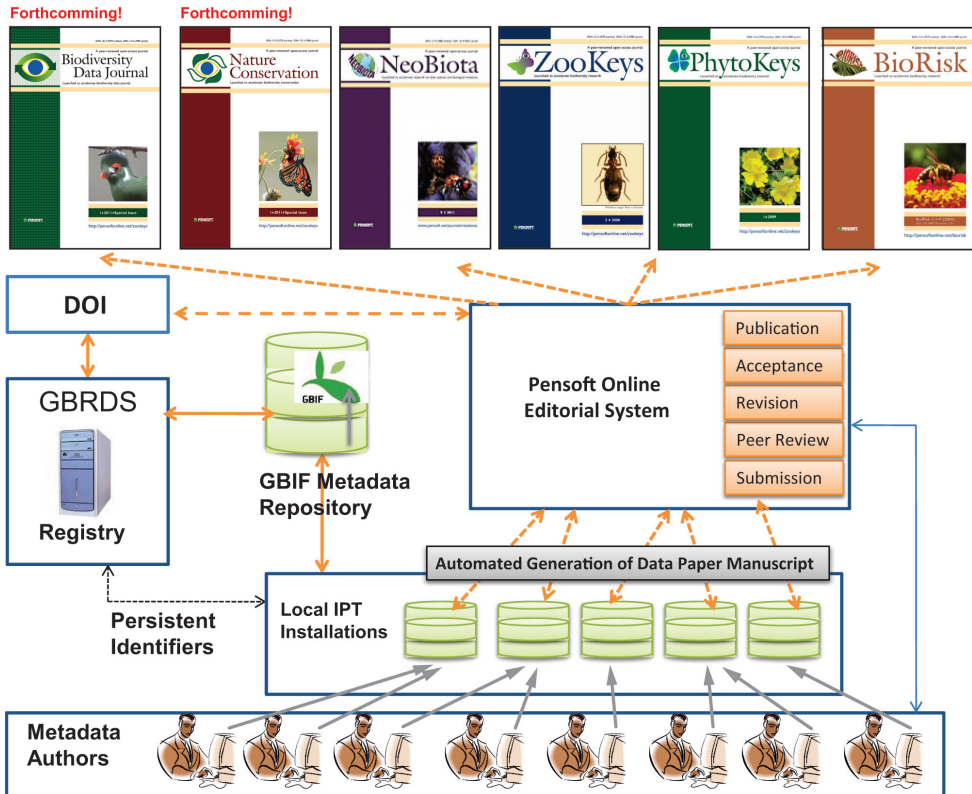


Figure 1. The GBIF/Pensoft workflow of data publishing and automated generation of Data Paper manuscripts

- A list of IPT installations supporting authoring of the Data Paper is accessible at <http://tools.gbif.org/data-paper-authoring/>.
- Pensoft's IPT-based Data Hosting Center (<http://ipt.pensoft.net/ipt/>) can also be used for authoring metadata. Please ask the Journal's Editorial Office to open an account for you.

2. Once the metadata are complete to the best of the author's ability, a Data Paper manuscript may be generated automatically from these metadata using the automated tool available within IPT 2.0.2+ (menu: Manage Resources - RTF download).
3. The author checks the created manuscript, completing the textual Introduction, and then submits it for publication in the Data Paper section of an appropriate Pensoft journal through the online submission system.
4. The manuscript undergoes peer review according to the journal's policies and the [Guidelines for Reviewers of Data Papers](#) (below). After review, and in case of acceptance, the manuscript is returned to the author by the editor alongside the reviewers' and editorial comments, for any required pre-publication modifications.

5. The corresponding author inserts all accepted corrections or additions recommended by the reviewers and the editor **in the metadata (not the manuscript of the paper)**, thereby improving the metadata for the data resource itself. Once the metadata have been improved, these can be made available on the IPT by pressing Publish button in the Manage Resources menu).
6. The final revised version of the Data Paper manuscript can then be created using the same automated metadata-to-manuscript conversion tool within IPT 2.0.2+ (menu: Manage Resources - RTF download) as was used to create the initially submitted draft. It is really important to remember that all changes in the metadata will be exported in the revised manuscript only **after** they have been recorded as metadata by pressing the button “Publish” in the “Manage Resources” menu.
7. After manual re-insertion of the text of the Introduction, the revised Data Paper manuscript can then be submitted to the journal for final review and subsequent acceptance decision.
8. Once the manuscript is accepted, it goes to a proofing stage, at which point submission, revision, acceptance and publication dates are added by the publisher, and a Digital Object Identifier (DOI) is assigned to the Data Paper. This facilitates persistent accessibility of the online scholarly publication.
9. Once the final proofs are approved by the author, the Data Paper is published in four different formats: (a) print format, (b) PDF format, identical to print version, (c) semantically enhanced HTML to provide interactive readings and links to external resources, and (d) final published XML to be archived in PubMedCentral and other archives to facilitate future data mining.
10. After publication, the DOI of the Data Paper is linked with the Persistent Identifier of the metadata document registered in the GBIF Registry which is given in the Data Paper. This provides multiple cross-linking between the data resource, its corresponding metadata and the corresponding Data Paper.
11. Depending on the journal’s policies and scope, the published Data Paper will be actively disseminated through the world’s leading indexers and archives, including Web of Knowledge (ISI), PubMedCentral, Scopus, Zoological Record, Google Scholar, CAB Abstracts, Directory of Open Access Journal (DOAJ) and EBSCO.

Data Papers Describing Ecological and Environmental Data

Metadata descriptions of primary biodiversity data used in the GBIF Metadata Profile (GMP) and the Integrated Publishing Toolkit (IPT) are based on standards created by ecologists and geographers, in particular [ISO 19139: North American Profile of ISO19115:2003 – Geographic information – Metadata](#) and the [Ecological Metadata Language \(EML\) Specification](#). Therefore, the same basic elements and the overall

Data Paper structure explained in the previous section can also be used to describe ecological and environmental data.

As a result, Data Papers for ecological and environmental data will have a basic structure similar to that of papers on primary biodiversity data (see [sample data paper](#)). Authors are encouraged to include additional elements (sections) in the manuscripts if they expect this to improve the description of the specifics of their environmental and ecological data. The **main difference is that ecological and environmental data cannot be processed through the GBIF IPT**, and hence they should be deposited in another public data hosting centre listed in the section [Open Data Repositories](#), such as [PANGAEA](#) or [DRYAD](#). Additionally, the Data Paper cannot be created automatically using the GBIF IPT, and thus must be authored manually.

Authors intending to publish Data Papers describing ecological and environmental data are advised to use the following steps:

1. Deposit your data in a certified public (international or institutional) repository.
2. Write a Data Paper manuscript following the structure of the [sample data paper](#), adding additional elements/sections to the manuscript if these are necessary to describe the specifics of your dataset(s).
3. Add the permanent link(s) to the particular dataset(s) hosted in the repository you have chosen.
4. Submit the Data Paper to an appropriate Pensoft journal.
5. Once the paper is accepted and published, enter the bibliographic reference and the DOI of the Data Paper in the relevant metadata field of your data package in the repository that hosts your data.

Data Papers Describing Genome Data

Pensoft journals require, as a condition for publication, that genome data supporting the results in the paper should be archived in an appropriate public archive, and accession numbers must be included in the final version of the paper. Sufficient additional metadata (such as sample locations, individual identities, etc.) should also be provided to allow easy repetition of analyses presented in the paper. It is possible that a single investigation may result in data in more than one archive.

DNA sequence data should be archived in [GenBank](#) or another public database. Expression data should be submitted to the [Gene Expression Omnibus](#) or an equivalent database, whereas phylogenetic trees should be submitted to [TreeBASE](#). More idiosyncratic data, such as microsatellite allele frequency data, can be archived in a more flexible digital data repository such as [Dryad](#) or [Knowledge Network for Biocomplexity \(KNB\)](#).

Barcode Data Release Papers

Barcode-of-Life COI (mitochondrial encoded *cytochrome oxidase 1*) genome data can be published in a form of a Data Paper, as has been recently announced by the [Consortium for the Barcode of Life \(CBoL\)](#) and illustrated by some sample papers published in [PLoS ONE](#)¹. CBoL urges participants in major DNA barcoding initiatives to consider submitting “[BARCODE Data Release Papers](#)” for publication in academic journals. A [BARCODE Data Release Paper](#) is a short manuscript that announces and documents the public release to a member of the [International Nucleotide Sequence Data Collaboration \(INSDC\)](#), which includes [GenBank](#), [EMBL](#), and [DDBJ](#)) of a significant body of data records that meet the [BARCODE data standards](#). The instructions below are incorporated from the [Guidelines to Authors of BARCODE Data Release Papers](#) and adjusted for the specifics of Pensoft journals, which would welcome such papers.

Definition: A [BARCODE Data Release Paper](#) is a short manuscript that announces and documents the public release to a member of the [International Nucleotide Sequence Data Collaboration \(INSDC\)](#), which includes [GenBank](#), [ENA](#), and [DDBJ](#)) of a significant body of data records that meet the [BARCODE data standards](#).

Contents: [BARCODE Data Release Papers](#) are meant to announce and document the public availability of a significant body of new DNA barcodes. The barcode records should therefore be a coherent set of records that provides noteworthy new research capabilities for a taxonomic group, ecological assemblage or specified geographic region. Authors should explain the rationale for creating a comprehensive library of [BARCODE](#) data for that taxonomic group, ecological habitat, and/or geographic region. If the data have been collected as part of a larger, longer-term research project, the manuscript should explain the wider project and its planned use of the data for taxonomic, biogeographic, evolutionary, and/or applied research, or for other purposes.

The [BARCODE Data Release Paper](#) manuscript should describe:

- The scope of taxonomic, ecological, and geographic coverage;
- The sources of voucher specimens;
- The sampling and laboratory protocols used;
- The processes used to identify the species to which voucher specimens belong.

The manuscript should provide summaries of data density and quality such as those shown in [Table 3](#):

¹ Two [BARCODE](#) data release papers have been published in [PLoS ONE](#): Hubert N, Hanner R, Holm E, Mandrak NE, Taylor E, et al. (2008) [Identifying Canadian Freshwater Fishes through DNA Barcodes](#). [PLoS ONE](#) 3(6): e2490; Kerr KCR, Lijtmaer DA, Barreira AS, Hebert PDN, Tubaro PL. (2009) [Probing Evolutionary Patterns in Neotropical Birds through DNA Barcodes](#). [PLoS ONE](#) 4(2): e4379

Table 3. Suggested data fields for a BARCODE Data Release Paper

Average number of records per species	
Range of records per species	Min-Max
Average sequence length (and Min/Max)	
Range of intraspecific variation*	Min-Max
Median variation within species*	X%
Range of divergence between closest species-pairs**	Min-Max
Median divergence between closest species-pairs**	

* Calculated as the arithmetic average of all K2P distances between specimens in each species.

** Closest species pairs refers to each species and the other species with which it has the least divergent barcode sequence. The true phylogenetic sister-species may not be included in the dataset, and could have a lower interspecies divergence.

Manuscripts should also include an Appendix with a table that presents:

1. The taxonomic identification (a formal species name or a provisional species label in a public database);
2. The collecting locality to a reasonable level of precision;
3. The voucher specimen identifier in the format required in the BARCODE data standard;
4. The accession number in GenBank, EMBL or DDBJ; and
5. The [Barcode of Life Data Systems \(BOLD\)](#) record number (optional).

Review Criteria: In addition to the general [Guidelines for Reviewers](#) listed in next section, CBOL recommends that reviewers use the following evaluation criteria for BARCODE Data Release Papers, and suggests that authors anticipate such evaluation:

1. **Data quality:** All data records should meet the BARCODE data standards agreed to by CBOL and INSDC. The manuscript should demonstrate the effectiveness of the BARCODE records in distinguishing species, as well as pointing out limitations of the BARCODE data for species identification.
2. **Significance of the data records:** The data records being released should represent a significant addition to the public knowledge base. The manuscript should demonstrate the significance of the new records relative to the previously released BARCODE data for that combination of taxonomic group, ecological habitat, and geographic region. Manuscripts that announce the release of the first BARCODE records representing a higher proportion species in a taxonomic group will have higher significance.
3. **Relevance to other research programs and societal applications:** BARCODE data release manuscripts will be considered more relevant if they treat taxonomic groups, ecological habitats, and geographic regions that are connected with other basic research programs in evolutionary biology or ecology, or are components of applied research for socioeconomic reasons

(e.g., agriculture, food safety, conservation, environmental monitoring, public health).

4. **Documentation and accessibility:** The voucher specimens and their associated data and metadata will be valuable resources for the research community. The data table in the Appendix must provide links to the voucher specimens and taxonomic identification, as well as the INSDC Accession Numbers. Reviewers will evaluate the degree to which voucher specimens are available in permanent repositories (as opposed to private research collections) and degree to which taxonomic identifications are documented in published or other resources. Provisional non-Linnean taxonomic labels may be used, but they should be linked to online databases that document the author's concept of the taxonomic unit.

Data Papers Describing Software Tools

An increasing number of software tools also merit description in scholarly publications. The structure of the Data Paper proposed below for such software tools is largely based on the [Description of a Project \(DOAP\) RDF schema](#) and XML vocabulary developed by Edd Dumbill to describe software projects, in particular those that are [open-source](#). The main difference, however, is that the Data Paper aims at the description of the *software product and not of the software source code*; Data Papers of this kind are addressed mainly to *end users of the software and less to developers and software engineers*.

According to DOAP, major properties of a software tool description include elements such as homepage, developer, programming language and operational system. Other properties include: Implements specification, anonymous root, platform, browse, mailing list, category, description, helper, tester, short description, audience, screenshots, translator, module, documenter, wiki, repository, name, repository location, language, service endpoint, created, download mirror, vendor, old homepage, revision, download page, license, bug database, maintainer, blog, file-release, and release.

A basic version of a DOAP description can be generated using an online tool called [doapamatic](#).

A sample structure of a Data Paper describing a software tool is listed in Table 4 below. Please note that this structure is provided in order to recommend a more or less unified character of this kind of Data Papers. The sections and sub-sections listed in the left column are *mandatory* for a Data Paper, while their content, listed in the right column in a form of elements or recommendations, needs to be defined by the authors to describe the software tool in the best possible way.

Table 4. Metadata elements (based on EML and DOAP) to be included in a Data Paper describing a software tool

Section/Sub-Section headings of the Data Paper describing a software tool	Mapping from the available EML and DOAP metadata elements; a few other elements have been added to provide a better mapping to the Data Paper structure, with formatting guidelines
<TITLE>	Derived from the 'name' element. This must be extended to a concise description of the software tool and its implementation, e.g.: "BioDiv, a web-based tool for calculation of biodiversity indexes". Format: This is a centred sentence without full stop (.) at end.
<Authors>	Derived from the 'developer', 'maintainer' and eventually 'helper', 'tester', and 'documenter'. From these elements, combinations of 'first name' and 'last name' are derived, separated by (,) commas. Corresponding affiliations of the authors are denoted with numbers (1, 2, 3,...) in superscript at the end of each last name. If two or more authors share same affiliation, it will be denoted by use of the same superscript number. Format: centred.
<Affiliations>	Derived from the 'developer', 'maintainer' and 'helper'. From these elements, combinations of 'Organisation Name', 'Address', 'Postal Code', 'City', and 'Country' will constitute the affiliation.
<Corresponding authors>	Derived from any of the 'developer', 'maintainer', 'helper', 'tester', and 'documenter' elements. From these elements 'first name', 'last name' and 'email' are derived. Email addresses are written in parentheses (). In case of more than one corresponding author, these are separated by commas. Format: indented from both sides.
<Received, Revised, Accepted, and Published dates>	These will be inserted manually by the Publisher of the Data Paper to indicate the dates of original manuscript submission, revised manuscript submission, acceptance of manuscript and publishing of the manuscript as Data Paper in the journal.
<Citation>	This will be inserted manually by the Publisher of the Data Paper. It will be combination of Authors, Year of Data Paper publication (in parentheses), Title, Journal Name, Volume, Issue number (in parentheses), and DOI of the Data Paper in both native and resolvable HTTP format.
<Abstract>	Derived from the 'short description' element. Format: indented from both sides.
<Keywords>	Keywords should reflect most important features of the tool and areas of implementation, and should be separated by commas (,).
<Introduction>	Free text.
<Project Description>	Derived from 'description' element; if applicable, it should also include sub-elements such as "title" of the project, 'personnel' involved in the project, 'funding' sources', and other appropriate information.
<Web Location (URIs)>	Derived from the elements 'homepage', 'wiki', 'download page', 'download mirror', 'bug database', 'mailing list', 'blog', 'vendor'
<Technical specification>	Derived from the elements 'platform', 'programming language', 'operational system' (if OS-specific), 'language' ', 'service endpoint'
<Repository>	Derived from the elements 'repository type' (CVS, SVN, Arch, BK), 'repository browse uri' (CVS, SVN, BK), 'repository location') SVN, BK, Arch), 'repository module' (CVS, Arch), 'repository anonymous root' (CVS)

Section/Sub-Section headings of the Data Paper describing a software tool	Mapping from the available EML and DOAP metadata elements; a few other elements have been added to provide a better mapping to the Data Paper structure, with formatting guidelines
<License>	Derived from the 'license' element
<Implementation>	Derived from 'Implements specification' and 'audience' elements; please remember that this section is of primary interest to end users, and should be written in detail, if possible including use cases, citations and links.
<Additional Information>	Any kind of helpful additional information may be included.
Acknowledgement	Lists all acknowledgments at the author's discretion
<References>	Includes literature references and web links cited in the text.

GUIDELINES FOR REVIEWERS OF DATA PAPERS

Data Papers describing data resources submitted to Pensoft journals will be subjected to conventional pre-publication anonymous and private peer review, as a routine method to enhance the completeness, truthfulness and accuracy of the descriptions of the relevant data resources, thereby improving their use and uptake. In the near future, we will add the possibility of public peer review for Data Papers during the editorial process, as well as of post-publication review in the form of comments about and recommendations for the Data Paper in question.

Peer review of Data Papers is expected to evaluate the completeness and quality of the dataset(s) description (metadata), as well as the publication value of data. This may include the appropriateness and validity of the methods used, compliance with applicable standards during collection, management and curation of data, and compliance with appropriate metadata standards in the description of the data resources. In order to allow for accuracy and usefulness, metadata needs to be as complete and descriptive as possible.

Reviewers will consider the following aspects of (a) the quality of the manuscript, (b) the quality of the data, and (c) the consistency between the description within the Data Paper and the repository-held metadata relating the data resource itself.

Quality of the Manuscript

- Do the title, abstract and keywords accurately reflect the contents of the Data Paper?
- Is the Data Paper internally consistent, suitably organized and written in proper English?
- Are relevant non-textual media (e.g. tables, figures, audio, video) used to an appropriate extent and in a suitable manner?
- Have abbreviations and symbols been properly defined?
- Does the Data Paper put the data resource being described properly into the context of prior research, citing pertinent articles and datasets?

- Are conflicts of interest, relevant permissions and other ethical issues addressed in an appropriate manner?

Quality of the Data

- Are the data completely and consistently recorded within the dataset(s)?
- Does the data resource cover scientifically important and sufficiently large region(s), time period(s) and/or group(s) of taxa to be worthy of a separate publication?
- Are the data consistent internally and described using applicable standards (e.g. in terms of file formats, file names, file size, units and metadata)?
- Are the methods used to process and analyses the raw data, thereby creating processed data or analytical results, sufficiently well documented that they could be repeated by third parties?
- Are the data plausible, given the protocols? Authors are encouraged to report any tests undertaken to address this point.
- Is the repository to which the data are submitted appropriate for the nature of the data?

Consistency between Manuscript and Data

- Does the Data Paper provide an accurate description of the data?
- Does the manuscript properly describe how to access the data?
- Are the methods used to generate the data (including calibration, code and suitable controls) described in sufficient detail?
- Is the dataset sufficiently unique to merit publication as a Data Paper?
- Are the use cases described in the Data Paper consistent with the data presented? Would other possible use cases merit comment in the paper?
- Have possible sources of error been appropriately addressed in the protocols and/ or the paper?
- Is anything missing in the manuscript or the data resource itself that would prevent replication of the measurements, or reproduction of the figures or other representations of the data?
- Are all claims made in the Data Paper substantiated by the underlying data?

ACKNOWLEDGEMENTS

This document was developed in part within the European Union FP7 project “ViBRANT - Virtual Biodiversity Research and Access Network for Taxonomy”. We thank several colleagues who commented or contributed to the draft: Eamonn

O'Tuama, Tim Robertson and Kyle Braak (GBIF); Teodor Georgiev, Pavel Stoev and Ivailo Stoyanov (Pensoft Publishers); and Todd Vision and Peggy Schaeffer (Dryad).

REFERENCES

- BMC position statement on open data. Archived in: <http://www.webcitation.org/5yHxhniPL>
- Chavan V, Penev L (in press) Data Paper: Mechanism to incentivise discovery of biodiversity data resources. *BMC Bioinformatics*.
- Chavan V, Ingwersen P (2009) Towards a data publishing framework for primary biodiversity data: challenges and potentials for the biodiversity informatics community. *BMC Bioinformatics*, 10 (Suppl 14): S2. doi:10.1186/1471-2105-10-S14-S2
- Krump J (2011) Criteria for the Trustworthiness of Data Centres. *D-Lib Magazine* 17 (1/2). doi:10.1045/january2011-klump
- Penev L, Erwin T, Miller J, Chavan V, Moritz T, Griswold C (2009a) Publication and dissemination of datasets in taxonomy: ZooKeys working example. *ZooKeys* 11: 1-8. doi: 10.3897/zookeys.11.210
- Penev L, Sharkey M, Erwin T, van Noort S, Buffington M, Seltmann K, Johnson N, Taylor M, Thompson FC, Dallwitz MJ (2009b) Data publication and dissemination of interactive keys under the open access model: ZooKeys working example. *ZooKeys* 21: 1–17. doi: 10.3897/zookeys.21.274
- Shotton D (2011) Data Citation Best Practice Discussion Document, Version 1, 11-05-2011. <http://bit.ly/1E2E21>.
- Talamas E, Masner L, Johnson N (2011) Revision of the Malagasy genus *Trichoteleia* Kieffer (Hymenoptera, Platygastridae, Platygastridae). *ZooKeys* 80: 1-126. doi: 10.3897/zookeys.80.907
- Thessen AE, Patterson DJ (2011) Data Issues in the Life Sciences. White paper. Marine Biological Laboratory, 55 pp. http://dataconservancy.org/sites/default/files/Data_Issues_in_the_Life_Sciences_White_Paper.pdf