

Origins of data papers

Lyubomir Penev

CEO & Founder, Pensoft Publishers

l.penev@pensoft.net



Science Publisher & Technology Provider

- Based in Sofia & Brussels, ca. 50 permanent staff
- Publisher of 35 biodiversity and ecology journals (out of 60 on ARPHA) & hundreds of book titles
- Inventor of XML-based publishing in biodiversity & ecology (based on Plazi's TaxPub JATS XML standard)
- Inventor of publishing tools and workflows for biodiversity data, such as the ARPHA Writing Tool
- Inventor of the first large-scale conversion workflow of biodiversity full-text articles into Linked Open Data (LOD)
- Creator of the OpenBiodiv Biodiversity Knowledge Graph & OpenBiodiv-O ontology
- Coordinator of BiCIKL: Biodiversity Community Integrated Knowledge Library



XML-based full-featured publishing platform



AUTHORING - REVIEWING - PUBLISHING - HOSTING - ARCHIVING

FULL-FEATURED PUBLISHING PLATFORM AND SERVICES



Various data publishing models

- Standalone data publishing by aggregators (GenBank, GBIF, BOLD, etc.)
- Data underlying research article (deposited in a repository or published in supplementary files)
- Data described in data papers
- Integrated structured narrative & data
- FAIR Open Linked Data publishing



How to publish data?



Home Articles About

About Pensoft

Books

E-Books

Blog

Journals

My tasks



Lyubomir Penev



Guidelines

Research Ideas and Outcomes 3: e12431
<https://doi.org/10.3897/rio.3.e12431> (28 Feb 2017)

Reviewed v1



XML

PDF



0



Contents

Article info

Citation

Metrics

Reviews 2

Related

Figs

Tabs

Refs

Cited

Strategies and guidelines for scholarly publishing of biodiversity data

▼ Lyubomir Penev, Daniel Mietchen, Vishwas Shravan Chavan, Gregor Hagedorn, Vincent Stuart Smith, David Shotton, Éamonn Ó Tuama, Viktor Senderov, Teodor Georgiev, Pavel Stoev, Quentin John Groom, David Remsen, Scott C. Edmunds

Abstract ▲

The present paper describes policies and guidelines for scholarly publishing of biodiversity and biodiversity-related data, elaborated and updated during the Framework Program 7 EU BON project, on the basis of an earlier version published on Pensoft's website in 2011. The document discusses some general concepts, including a definition of datasets, incentives to publish data and licenses for data publishing. Further, it defines and compares several routes for data publishing, namely as (1) supplementary files to research articles, which may be made available directly by the publisher, or (2) published in a specialized open data repository with a link to it from the research article, or (3) as a data paper, i.e., a specific, stand-alone publication describing a particular dataset or a collection of datasets, or (4) integrated narrative and data publishing through online import/download of data into/from manuscripts, as provided by the Biodiversity Data Journal.

Article metadata

Data Publishing in a Nutshell

— Introduction

— What is a Dataset

— Why Publish Data

— How to Publish Data

— How to Cite Data

Data Publishing Policies

— General Policies for Biodiversity data

— Data Publishing Licenses

Data Deposition in Open Repositories

— General Information


— Taxonomy

— Species-by-Occurrence and Sample-Based data

Powered by arpha

DOI: [10.3897/rio.3.e12431](https://doi.org/10.3897/rio.3.e12431)

2011: The origin of the Data Paper for biodiversity

 **BMC** Part of Springer Nature [Explore Journals](#) [Get Published](#) [About BMC](#) [Search](#) [Login](#)

BMC Bioinformatics

[Home](#) [About](#) [Articles](#) [Submission Guidelines](#)


Abstract
Background
The data paper
Discussion
Conclusions
Declarations
References

Volume 12 Supplement 15
[Data publishing framework for primary biodiversity data](#)

[Download PDF](#)
[Export citations](#)

Research | [Open Access](#)






The data paper: a mechanism to incentivize data publishing in biodiversity science

[Vishwas Chavan](#) [†]  and [Lyubomir Penev](#) [†]

[†]Contributed equally

BMC Bioinformatics 2011 **12** (Suppl 15) :S2
<https://doi.org/10.1186/1471-2105-12-S15-S2> | © Chavan and Penev; licensee BioMed Central Ltd. 2011
Published: 15 December 2011

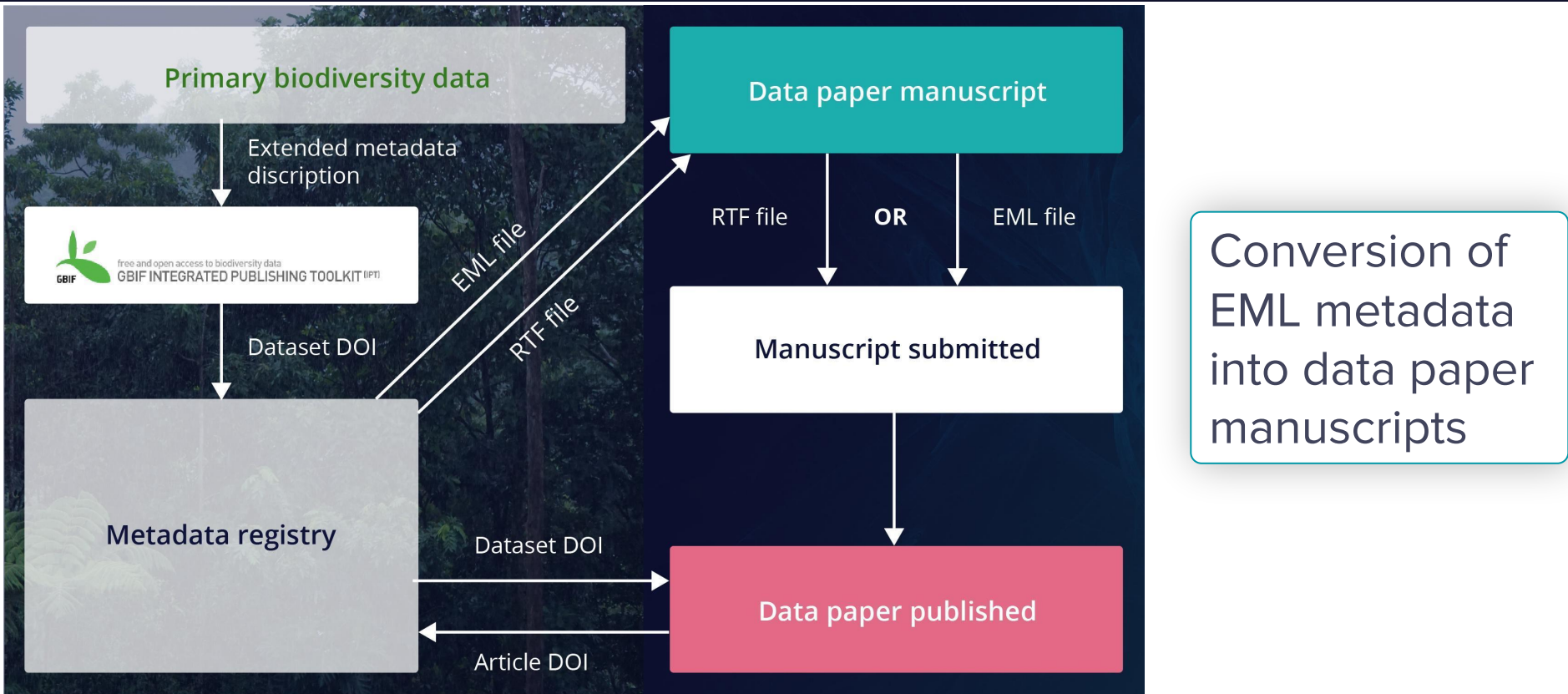
Metrics
Article accesses: 16537
Citations: 67 [more information](#)
Altmetric Attention Score: 76

Share This Article
    

[Get shareable link](#)

Abstract

Publishing data papers on GBIF <-> ARPHA



Conversion of EML metadata into data paper manuscripts

Start a manuscript in ARPHA Writing Tool

To start a manuscript select a journal and an article type

- Biodiversity Data Journal
- Research Ideas and Outcomes
- One Ecosystem
- BioDiscovery
- Biodiversity Information Science and Standards
- Food and Ecological Systems Modelling Journal
- ARPHA Conference Abstracts
- Viticulture Data Journal
- Biosystematics and Ecology

Research ideas


- Data Management Plan
- Grant Proposal
- PhD Project Plan
- PostDoc Project Plan
- Research Idea
- Software Management Plan

Early research outcomes

- Applied Study
- Case Study
- Clinical Case Studies
- Data analytics
- Data Paper (Biosciences)
- Data Paper (Generic)
- Emerging Technique
- Formal Model Article Format
- Forum Paper
- FSIX (Food Safety Knowledge)
- Methods
- Model validation
- OMICS Data Paper
- Project Report
- Questionnaire
- R Package
- Software Description

Brief research outcomes

- Commentary
- Conference Abstract
- Correspondence
- Ecosystem Accounting Table
- Ecosystem Inventory
- Ecosystem Service Mapping
- Ecosystem Service Models
- Institutional/Society announcement
- Interdisciplinary Perspectives
- Monitoring Schema
- Opinions
- Research Poster
- Research Presentation
- Short Communication
- Single-media Publication

 Reset selection

Create manuscript

OR

Import a manuscript

Data set described in EML in GBIF

OCCURRENCE DATASET | REGISTERED NOVEMBER 28, 2018

Moss occurrences in Yugyd Va National Park, Subpolar and Northern Urals, European North-East Russia

Published by [Institute of Biology of Komi Scientific Centre of the Ural Branch of the Russian Academy of Sciences](#)

✉ Galina Zheleznova • Tatyana Shubina • Svetlana Degteva • Mikhail Rubtsov • Ivan Chadin

DATASET

PROJECT

METRICS

ACTIVITY

DOWNLOAD

4,120 OCCURRENCES

2 CITATIONS

This study produced a dataset containing information on moss occurrences in the territory of Yugyd Va National Park, located in the Subpolar and Northern Urals, European North-East Russia. The dataset summarizes occurrences noted by long-term bryological explorations in remote areas of the Subpolar and Northern Urals from 1943 to 2015, and from studies published since 1915. The dataset consists of 4,120 occurrence records. The occurrence data were extracted from herbarium specimen labels (3,833... [More](#))

Project ID: AAAA-A17-117112270073-0

Metadata last modified: December 7, 2018

Data last changed: December 7, 2018

Hosted by: [Institute of Biology of Komi Scientific Centre of the Ural Branch of the Russian Academy of Sciences](#)

License: CC BY 4.0

” How to cite [DOI](#) 10.15468/kfeugm

4,120
Occurrences

100%
With taxon match

99.6%
With coordinates

99.9%
With year

Download the EML metadata file

Description

Temporal scope

Geographic scope

Taxonomic scope

Methodology

Bibliography

Contacts

Data description

GBIF registration

Citation

Data description

Metadata language: English

Data language: English

GBIF registration

Registration date: November 28, 2018

Metadata last modified: December 7, 2018

Publication date: December 7, 2018

Hosted by: Institute of Biology of Komi Scientific Centre of the Ural Branch of the Russian Academy of Sciences

Installation: Institute of Biology (Syktyvkar, Russia) Integrated Publishing Toolkit (IPT) Installation

Installation contacts: Ivan Chadin

Endpoints: http://ib.komisc.ru:8088/ipt/archive.do?r=mosses_occurrence_yugyd_va (Darwin Core Archive) • http://ib.komisc.ru:8088/ipt/eml.do?r=mosses_occurrence_yugyd_va (EML)

Preferred identifier: DOI 10.15468/kfeugm

Alternative identifiers: http://ib.komisc.ru:8088/ipt/resource?r=mosses_occurrence_yugyd_va

Last successful ingestion: April 22, 2020 (Not modified)


Last ingestion with changes: December 7, 2018

Last ingestion with data change: December 7, 2018

Occurrences in last ingestion: 4.120

Import your manuscript from GBIF EML file

 Feedback

 Tips and Tricks

Import a manuscript

Import from EML metadata

Browse

Supported EML versions: 2.1.1, 2.1.0 (e.g. generated from GBIF IPT, DataONE and LTER)

Your manuscript in ARPHA Writing Tool

The screenshot displays the ARPHA Writing Tool interface. At the top, there is a navigation bar with icons and labels for 'View dashboard', 'Messages', 'Collections', 'Reviewers', 'Email contributors', 'Helpdesk', 'Tips and tricks', and 'Tutorial'. Below this, the main content area is titled 'Data Paper (Biosciences)'. On the left side, there is a sidebar menu with categories: 'Authors', 'Contributors', 'Article metadata' (with sub-items: Title, Abstract & Keywords, Classifications, Funder), 'Introduction', 'General description', 'Project description', 'Sampling methods', 'Geographic coverage', 'Taxonomic coverage', 'Traits coverage' (with sub-item: Data coverage of traits), 'Temporal coverage', 'Collection data', 'Usage rights', 'Data resources', 'Additional information', and 'Acknowledgements'. The main content area shows a draft of a data paper. The title is 'Moss occurrences in Yugyd Va National Park, Subpolar and Northern Urals, European North-East Russia'. The authors listed are Galina Zheleznova[†], Tatyana Shubina[†], Svetlana Degteva[†], Mikhail Rubtsov[†], and Ivan Chadin[†]. The affiliation is '† Institute of Biology of Komi Scientific Centre of the Ural Branch of the Russian Academy of Sciences, Syktyvkar, Russia'. There is a 'Corresponding author: Galina Zheleznova (zheleznova@ibi.komisc.ru)' section and a copyright notice '© Galina Zheleznova, Tatyana Shubina, Svetlana Degteva, Mikhail Rubtsov, Ivan Chadin'. A 'Citation:' field is also present. The article is marked as 'OPEN ACCESS'. The top of the main content area includes a rich text editor toolbar with various icons for text formatting and document management. The bottom of the main content area shows the 'Abstract' section with a 'Background' subsection. The background text reads: 'This study produced a dataset containing information on moss occurrences in the territory of Yugyd Va National Park, located in the Subpolar and Northern Urals, European North-East Russia. The dataset summarizes occurrences noted by long-term bryological explorations in remote areas of the Subpolar and Northern Urals from 1943 to 2015, and from studies published since 1915. The dataset consists of 4,120 occurrence records. The occurrence data were extracted from herbarium specimen labels (3,833 records) and data from the published literature (287 records). Most of the records (4,104) are georeferenced. A total of 302 moss taxa belonging to 112 genera and 36 families are reported herein to occur in Yugyd Va National Park, although currently the diversity of bryophytes in this National Park has not yet been fully explored.'

EML metadata
converted into
data paper
manuscript

Author-performed data check before submission



About

Focus and Scope

How It Works

Guidelines for Authors

Criteria for Publication

Open Access Policy

Indexing and Archiving Policy

Globally Unique Innovations

Peer Review



Data Publishing Guidelines

Data Quality Checklist and Recommendations

Data Review Guidelines

Linked data table for primary biodiversity data

CHECKLIST

Characters

- The dataset is UTF-8 encoded
- The only characters used that are not numbers, letters or standard punctuation, are tabs and whitespaces
- Each character has only one encoding in the dataset
- No line breaks within data items
- No field-separating character within data items (tab-separated data preferred)
- No "?" or replacement characters in place of valid characters
- No Windows carriage returns
- No leading, trailing, duplicated or unnecessary whitespaces in individual data items

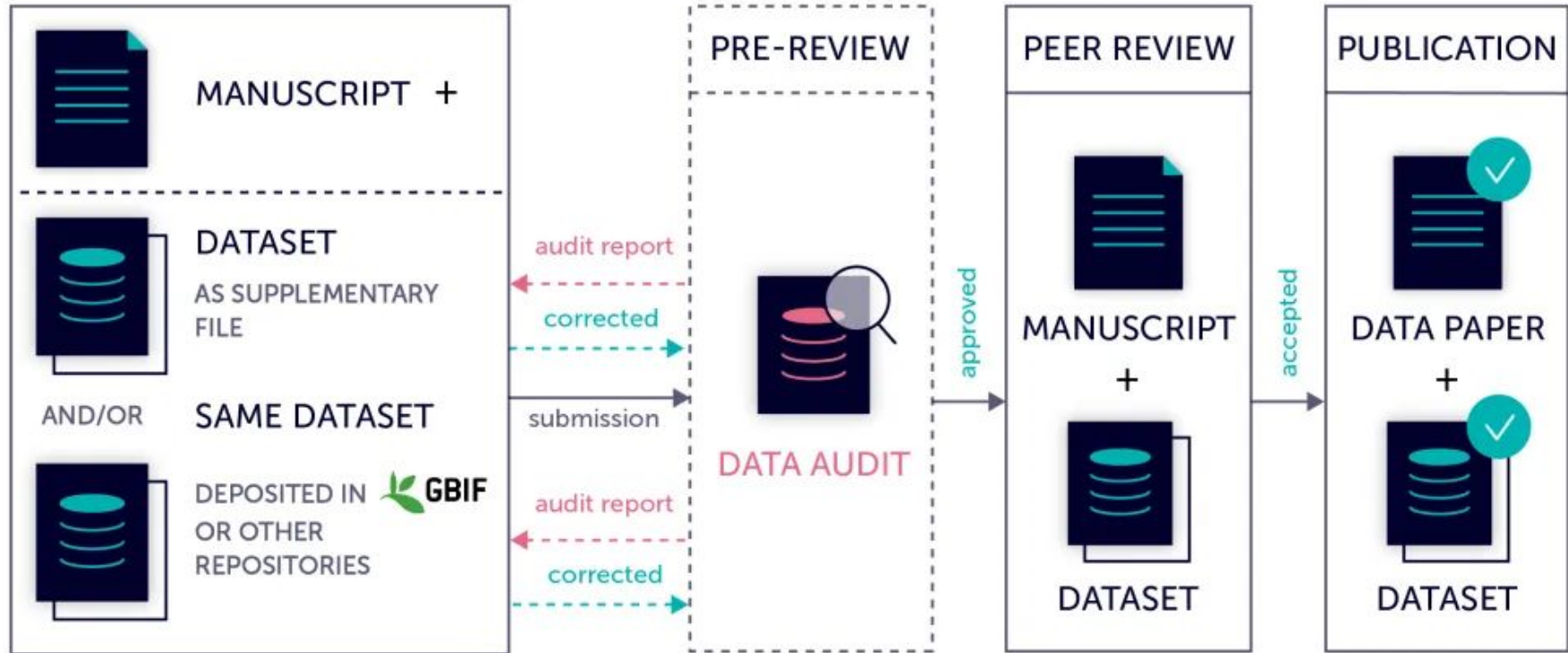
Records

- No broken records, i.e. records with too few or too many fields
- No blank records
- No duplicate records (as defined by context)

Fields

- No empty fields

Data audit after submission but prior to peer review



Data audit report

Data audit for technical evaluation of

Vascular plants dataset of the COFC herbarium (University of Cordoba, Spain) **(associated GBIF dataset)**

Downloaded on 2019-06-19 from <https://www.gbif.org/dataset/837c0162-f762-11e1-a439-00145eb45e9a>

Dr Robert Mesibov (robert.mesibov@gmail.com; <https://www.datafix.com.au>)
2019-06-20

About this evaluation

Pensoft does a technical evaluation of the dataset (or datasets) referred to in the data paper. If the dataset passes or has only minor problems, the data paper manuscript is referred to reviewers. If the dataset has major problems, a review of the paper is postponed until the dataset has been corrected.

To see what features of a dataset are checked in a technical evaluation, please go to

<https://zookeys.pensoft.net/about#DataQualityChecklistandRecommendations>

Please note that Pensoft does not check the details of the *content* of a dataset, for example whether the correct author is given for a scientific name, or whether the correct latitude/longitude is given for a locality.

Recommendation. The dataset associated with the manuscript has been processed by GBIF and the data paper could go on to review. However, there are many data problems in the GBIF upload, and I recommend to the authors that these problems be fixed and the data re-uploaded to GBIF for processing. The problems are detailed below by Darwin Core field in the field order in the dataset.

Many of the problems are not trivial and are causing data loss. For example, the decimalLatitude in FF92A873-601C-4360-86C7-5C9D483D6DAE is "30S266977.44". GBIF has rejected the location as "Coordinate invalid" (<https://www.gbif.org/occurrence/2235670578>).

Sources:

<https://blog.pensoft.net/2019/10/17/could-biodiversity-data-be-finally-here-to-last/>

<https://blog.pensoft.net/2019/10/17/case-study-data-audit-for-the-vascular-plants-dataset-of-the-cofc-herbarium-university-of-cordoba-spain-a-data-paper-in-phytokeys/>

Data audit: List or errors

(7) *municipality* has "_" for EC42F49A-68D5-4504-8F9E-0010859712A1.

(8) *locality* needs cleaning for the many pseudo-duplicates, e.g.

- 2 casco urbano, avda. del Brillante, nº 187, carril de la Huerta de los Arcos
- 7 casco urbano, avda. del Brillante, nº 187, carril Huerta de los Arcos
- 7 casco urbano, avda. del Brillante, nº 187, Carril Huerta los Arcos

and the many unnecessarily quoted entries, e.g.

"casa ""Rompealbardas""
""Villa Carmen"", ""El Calvario""

Also, *locality* is "_" for CC465E40-9868-4B01-8D2B-5CB9AC747674 and 8547AA0D-682B-4848-B31F-0399427D51FA

(9) *decimalLatitude* errors:

- 1 30S266977.44
- 1 37,91560°
- 1 40.9449°
- 1 41.9425N

Also, several entries have too many significant figures and should be rounded off, e.g. "37.0233172796695"

Sources:

<https://blog.pensoft.net/2019/10/17/could-biodiversity-data-be-finally-here-to-last/>

<https://blog.pensoft.net/2019/10/17/case-study-data-audit-for-the-vascular-plants-dataset-of-the-cofc-herbarium-university-of-cordoba-spain-a-data-paper-in-phytokeys/>

Data downloadable from the published article

Research Article

MycoKeys 45: 75-92

<https://doi.org/10.3897/mycokeys.45.30813> (29 Jan 2019)

XML

PDF



0



inference algorithm was estimated by jModeltest v. 2.1.10 (Darriba et al. 2012) using Akaike information criterion. Bayesian phylogenetic analyses were carried out using the Metropolis-coupled Markov chain Monte Carlo (MCMC) method in MrBayes v. 3.2 (Ronquist et al. 2012), under a GRT+I+G model. Markov chains were run for one million generations, with six chains and random starting trees. The chains were sampled every 100 generations. Among these, the first 2000 trees were discarded as the burn-in phase of each analysis and the resulting trees were used to calculate Bayesian posterior probabilities. Bayesian posterior probabilities (P.P.) ≥ 0.95 were considered as a significant support (Alfaro et al. 2003). Pairwise genetic distances (proportions of variable sites) within and between five *Apophysomyces* species were computed using MEGA v. 6 (Tamura et al. 2013), with pairwise deletion of gaps and missing data.

Table 1.

Sequences used for phylogenetic analysis. Type species of *Apophysomyces* are in bold.

Download as CSV

Taxa	Strain/isolate	GenBank accession number			References
		<i>ITS</i>	D1/D2 domain	<i>H3</i>	
<i>Apophysomyces elegans</i>	CBS 476.78	FN556440	FN554249	FN555155	Alvarez et al. 2010
<i>Apophysomyces elegans</i>	CBS 477.78	FN556437	FN554250	FN555154	Alvarez et al. 2010
<i>Apophysomyces elegans</i>	FMR 12015	HE664070	–	–	Da Cunha et al. 2012
<i>Apophysomyces variabilis</i>	CBS 658.93	FN556436	FN554258	FN555161	Alvarez et al. 2010
<i>Apophysomyces variabilis</i>	UITHSC 06 4222	FN556438	FN554255	FN555162	Alvarez et al. 2010

Contents Article info Citation Metrics Comment Related
Figs Tabs Map Taxa Refs Cited

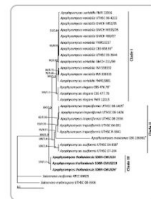


Figure 1.

doi Download

Phylogenetic tree derived from maximum likelihood analysis of a combined *ITS*, *LSU*, and *H3* genes of 28 sequences. *Saksenaia vasiformis* and *S. erythrospora* were used as outgroup. Numbers above branches are the bootstrap statistics percentages (left) and Bayesian posterior probabilities (right). Branches with bootstrap values $\geq 50\%$ are shown at each branch and the bar represents 0.1 substitutions per nucleotide position. The fungal isolates from this study are in bold. Superscript T = type species.

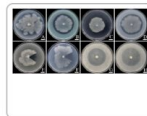


Figure 2.

doi Download

Solubilization of non-soluble minerals in agar media by *Apophysomyces thailandensis* SDBR-CMUS26 (holotype). A $\text{Ca}_3(\text{PO}_4)_2$ B $\text{CuCO}_3 \cdot \text{Cu}(\text{OH})_2$ C CuO D ZnCO_3 E FePO_4 F MnO G Feldspar H Kaolin. Scale bars: 10 mm. Fungal colonies in E and F were cut for the solubilization area (halo zone) observation.

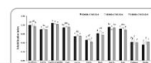


Figure 3.

doi Download

Solubilization index of the ability to solubilize minerals by *Apophysomyces thailandensis* SDBR-CMUS26 (holotype).

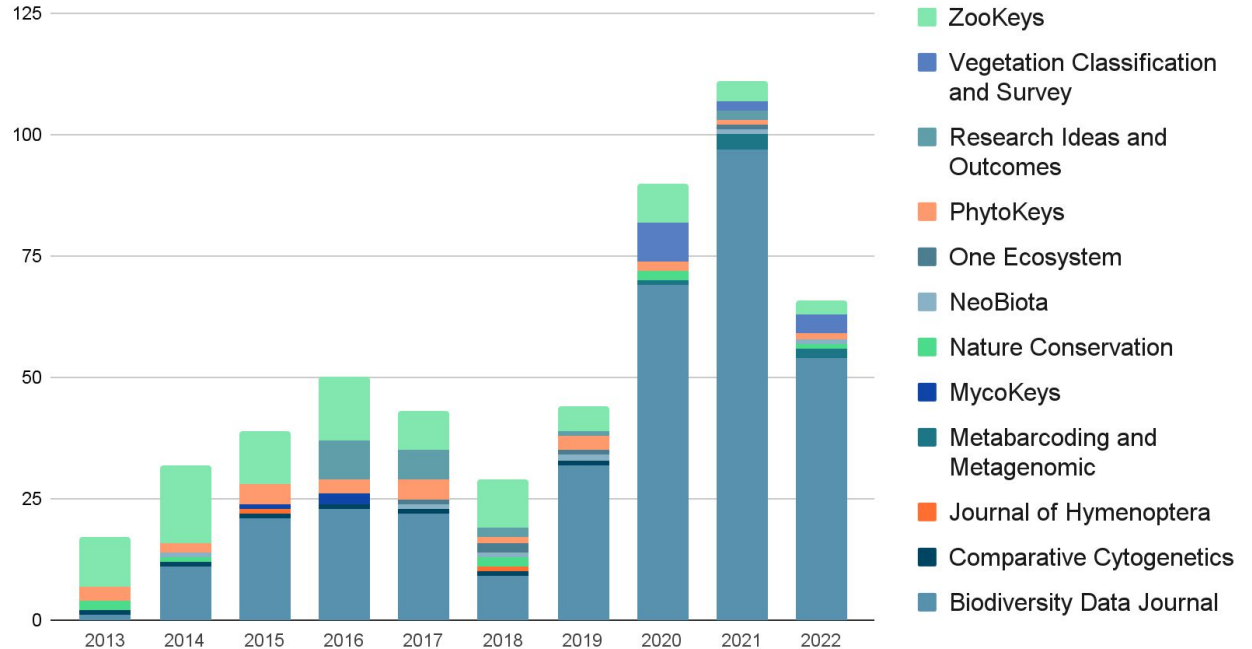
Powered by

arphahub.com

Instantly growing popularity of Data Papers

Data Paper
statistics across
Pensoft journals

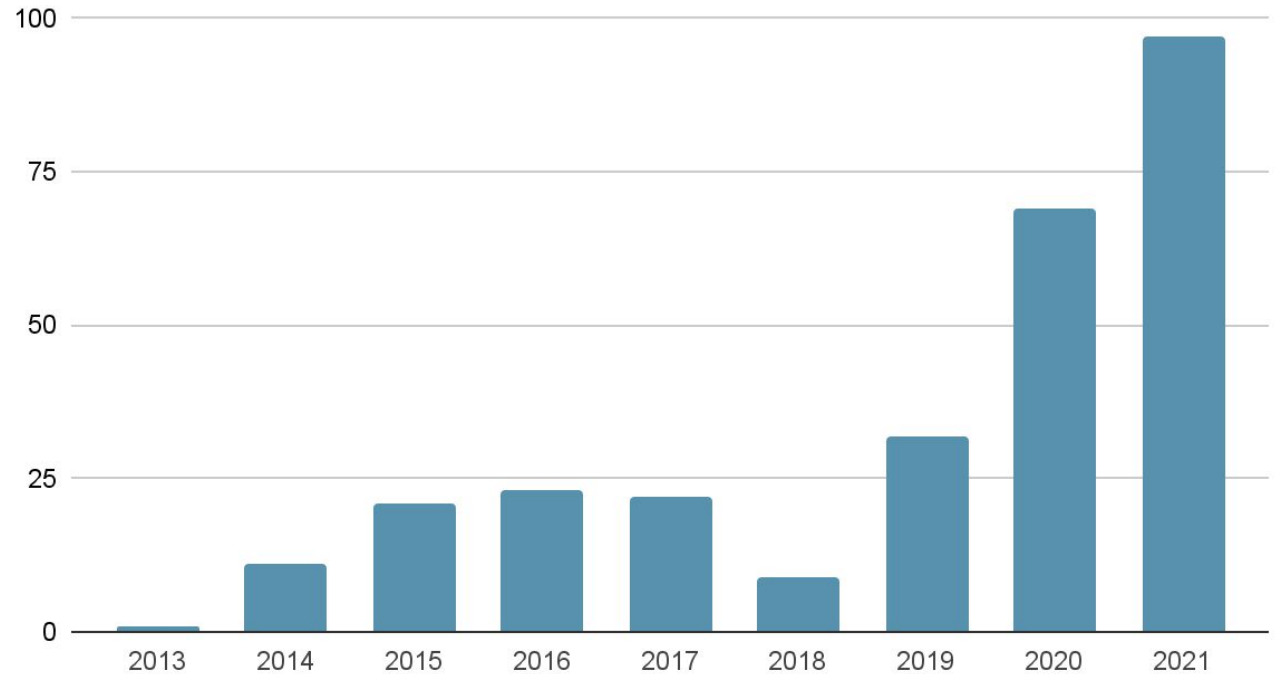
Data Papers published in Pensoft journals from 2013 to 2022



Instantly growing popularity of Data Papers

Data Paper
statistics in
the Biodiversity
Data Journal

Number of Data Papers in Biodiversity Data journal from 2013 to 2021



Previous joint projects of GBIF & BDJ



A peer-reviewed open-access journal
**Biodiversity
Data Journal**
Making your data count! ISSN 1314-2828 (online)

[Submit manuscript](#)

[Full Text](#) [Author](#) [Title](#)

[About](#)[Articles](#)[Topical collections](#)[Guidelines for Authors](#)[Editorial Team](#)[Contacts](#)

In this collection



Biota of Russia



Papers published: **59**

Documents added: **59**

Printed version: **Paperback**



Biota of Russia

Sort by: [Publication date newest](#) ▼

Edited by Vince Smith, Dmitry Schigel, Ivan Chadin, Alexey Seregin, Nina Filippova, Pedro Cardoso, Vladimir Blagoderov, Alexander Sennikov

Editors of the 2021 submissions: *Dmitry Schigel, Ivan Chadin, Alexey Seregin, Nina Filippova, Pedro Cardoso, Vladimir Blagoderov, Alexander Sennikov.*

A special collection of [data papers](#) on Russia in the Biodiversity Data Journal (BDJ) by GBIF - the Global Biodiversity Information Facility - in collaboration with the [Finnish Biodiversity Information Facility \(FinBIF\)](#) and [Pensoft Publishers](#).

In correspondence with the funding priorities of this programme, at least 80% of the records in a dataset should have coordinates that fall within Russia. However, authors of the paper may be affiliated with institutions anywhere in the world. Each of the data papers is a descriptor of more than 5,000 occurrence records from the target region that are new to GBIF.org. Datasets may contain additional records from other regions, and can be published as occurrence or sampling-event datasets, as well as checklists.

The collection is closed for submissions.

See [full description](#) of the 2021 call and additional resources.


Editors of the West of Urals 2020 call: *Vince Smith, Dmitry Schigel, Ivan Chadin, Alexey Seregin*

West of Urals 2020 call (archived): www.gbif.org/news/1VHfuSBGwSzDBxqRHuCAHY

doi [10.3897/bdj.coll.59](https://doi.org/10.3897/bdj.coll.59)

Previous joint projects of GBIF & BDJ




-  59 data papers published in 2020-2021 in the Biodiversity Data Journal's special collection "Biota of Russia"
-  58,720 total unique views (93,521 total views)

TOP-3 most popular papers according to the number of unique views:

- ["Flora of Russia" on iNaturalist: a dataset](#) - 3,759 unique views; 6,437 total views
- [MHA Herbarium: Eastern European collections of vascular plants](#) - 1,551 unique views; 2,599 total views
- [Reptile occurrences data in the Volga River basin \(Russia\)](#) - 1,509 unique views; 2,228 total views

Two-years results

 Each article is promoted on Twitter and Facebook

 Joint Pensoft-GBIF promotion across social media and blogs

Pensoft @Pensoft · 23 mar. 2021 r.

🇷🇺 Pleased to announce a new call for submissions for the "Biota of Russia" article collection in @BioDataJournal (bdj.pensoft.net/topical_collect...). The first 36 #DataPapers submitted until 15 September will be published FOR FREE!


@GBIF @iajtieto #Biodiversity #OpenData

GBIF @GBIF · 23 mar. 2021 r.

Got data from Russia? Between now and 15 Sep 2021, @BioDataJournal APCs waived for up to 36 data papers

- with more than 5k records new to GBIF.org in 2021
- with high-quality data and metadata
- with geographic coverage in Russia

gbif.org/news/HdDwty9j...




Get data How-to Tools Community About

PROJECT | CLOSED

Biota of Russia 2021

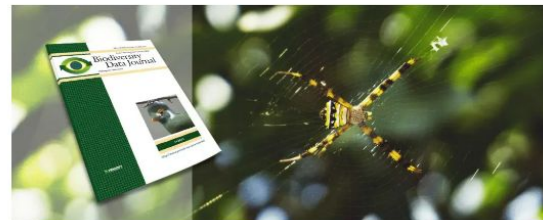
1 Mar - 31 Dec 2021 € 17,000

ABOUT NEWS & EVENTS DATASETS BIODIVERSITY DATA JOURNAL COLLECTION 291 CITATIONS



Siberian bugloss (*Brunnera sibirica*) observed in Kirovo-Chepetsk, Kirov Oblast, Russia by Вотничева Елена Александровна (CC BY-NC 4.0)

Call for data papers describing datasets from Russia to be published in Biodiversity Data Journal



March 24, 2021

Article collections, Biodiversity data, Biodiversity Data Journal, Data Publishing

article collection, biodiversity, biodiversity data, Biodiversity Data

GBIF PARTNERS WITH FINBIF AND PENSOFT TO SUPPORT PUBLICATION OF NEW DATASETS ABOUT BIODIVERSITY FROM ACROSS RUSSIA

BDJournal @BioDataJournal · 20 янв.

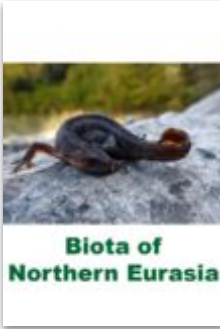
For the first time, data on #zooplankton in the Middle Volga River Basin are published: doi.org/10.3897/BDJ.9...

More on the biota of Russia in our special collection: bdj.pensoft.net/topical_collect...

#plankton #biodiversity #rivers



New special collection “Biota of Northern Eurasia”



-  Submission deadline: **1 December 2022**
-  **50** data papers to be published
-  Criteria for describing a dataset:
 - with more than 7,000 presence records new to GBIF.org in 2022
 - with high-quality data and metadata
 - with geographic data coverage in Northern Eurasia
 - authors of the paper must be affiliated with Northern Eurasian institutions in Ukraine, Kazakhstan, Kyrgyzstan, Uzbekistan, Tajikistan, Turkmenistan, Moldova, Georgia, Armenia, and Azerbaijan—or outside Northern Eurasia



**Thanks and
happy data
publishing!**