PENSOFT pensoft.net

Origins of data papers

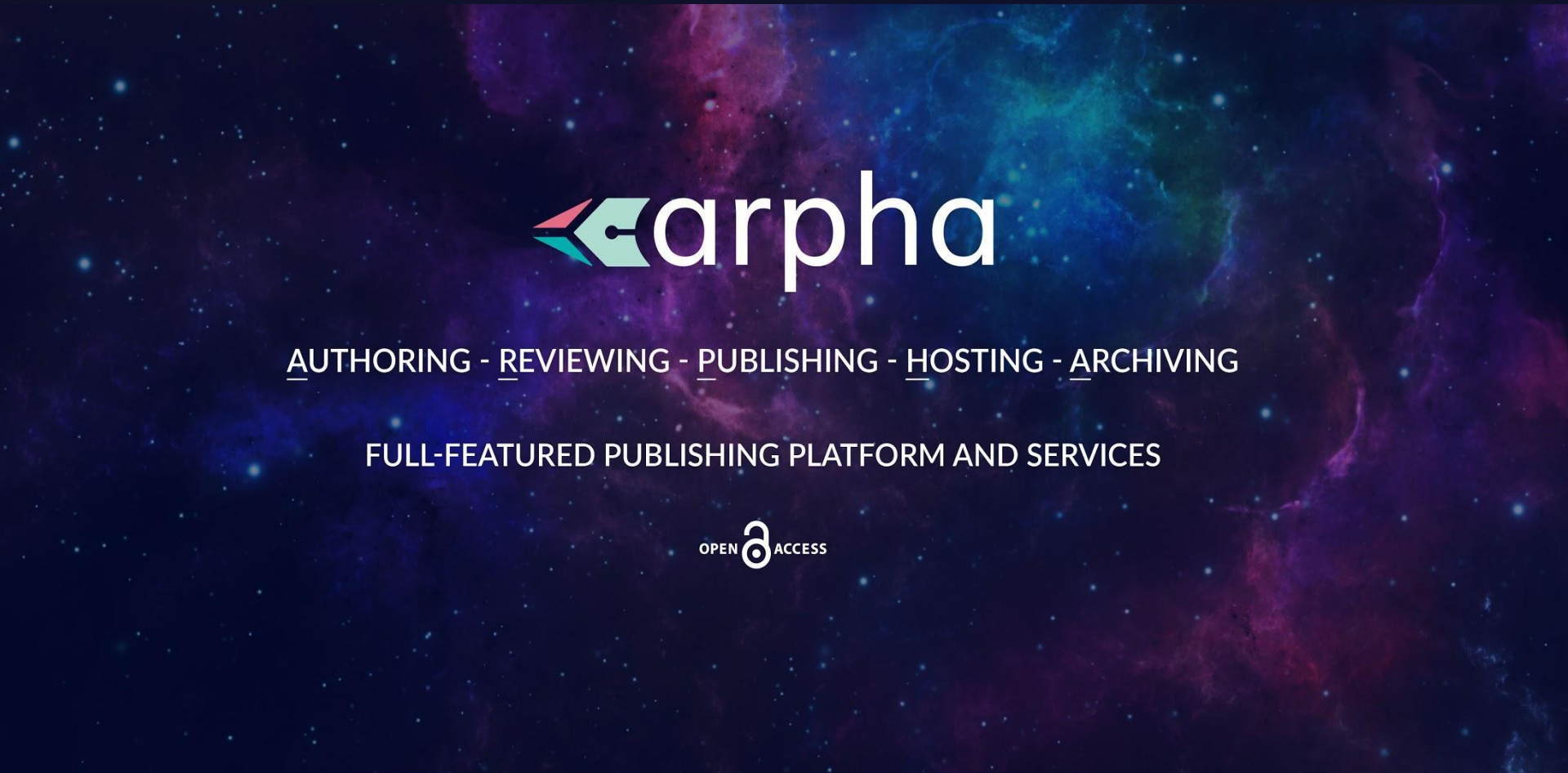
Lyubomir Penev
CEO & Founder, Pensoft Publishers
l.penev@pensoft.net

GBIF *webinar, 08/11/2022*

PENSOFT 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

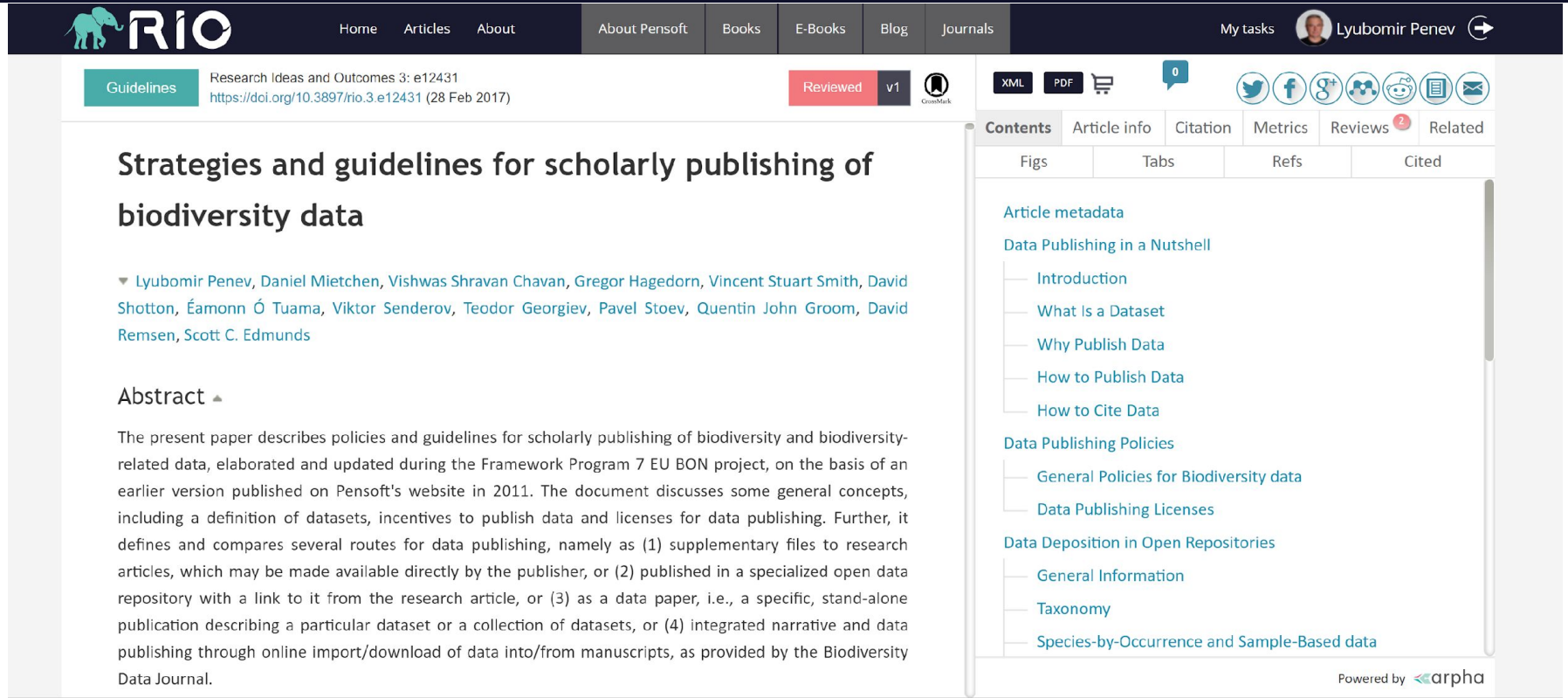
XML-based full-featured publishing platform



Various data publishing models

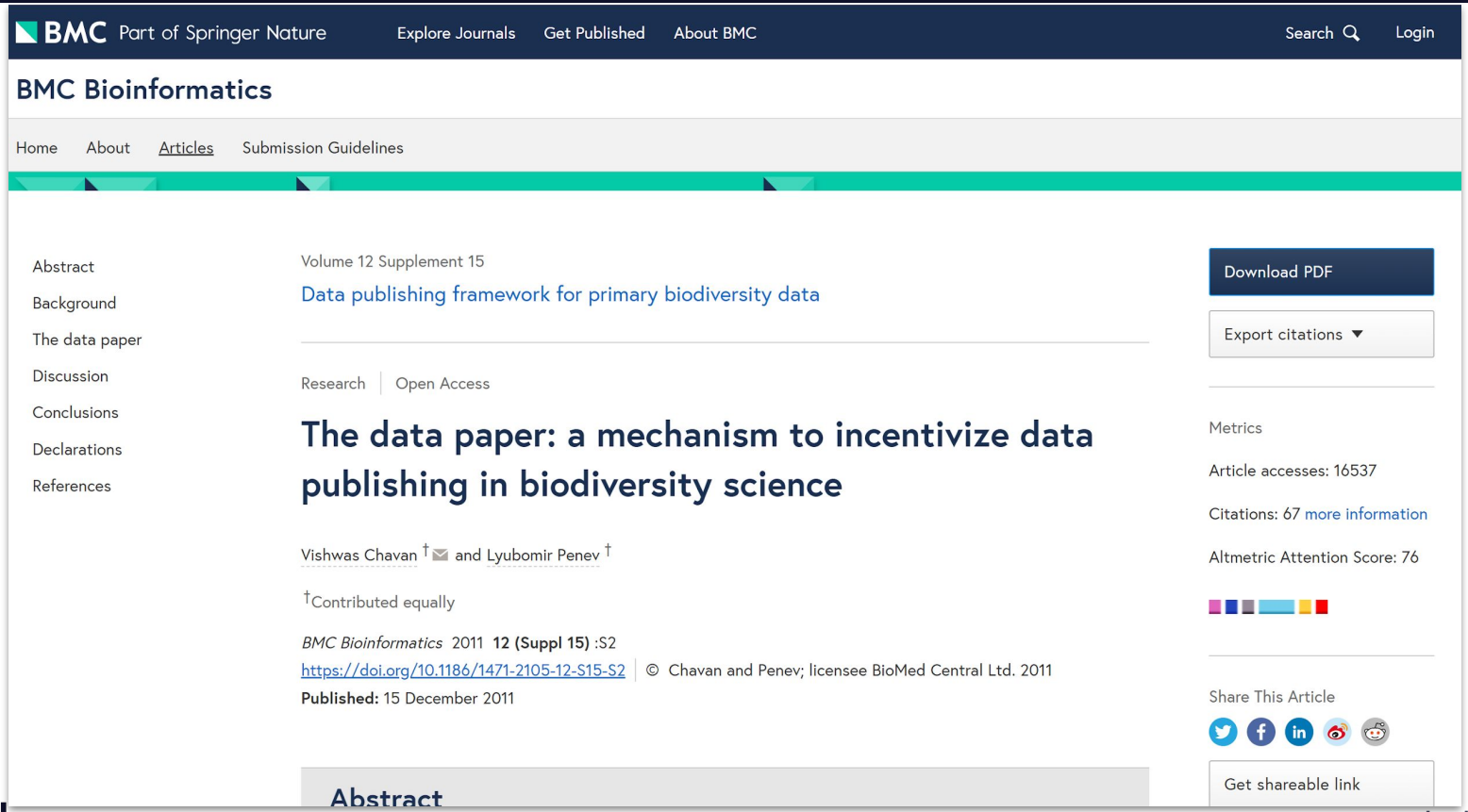
- Standalone data publishing by aggregators (GenBank, GBIF, BOLD, etc.)
- Data underlying research article (deposited in a repository or published in supplementary files)
- Data described in data papers
- Integrated structured narrative & data
- FAIR Open Linked Data publishing

How to publish data?



DOI: 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

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
Published: 15 December 2011

Abstract

Download PDF

Export citations

Metrics

Article accesses: 16537

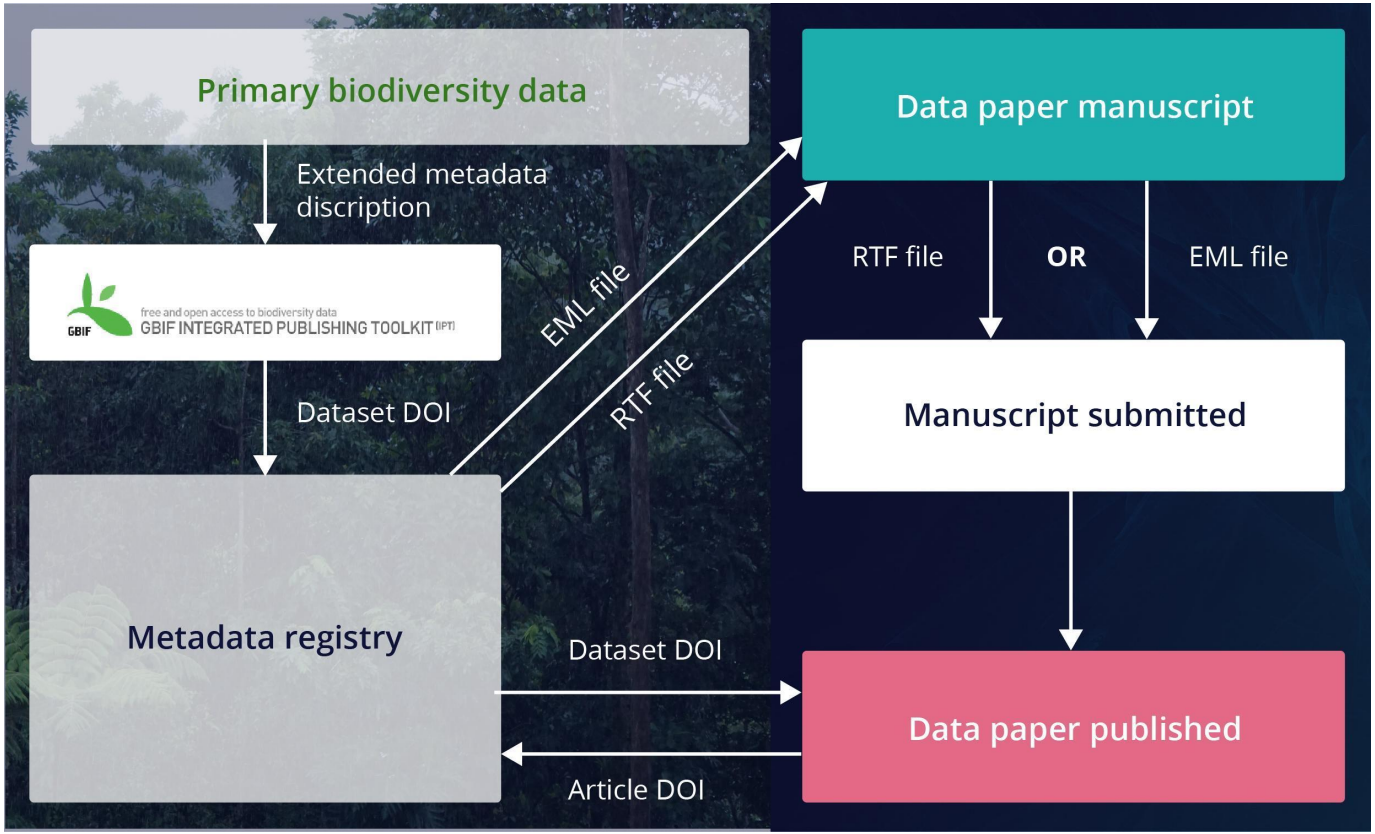
Citations: 67 more information

Altmetric Attention Score: 76

Share This Article

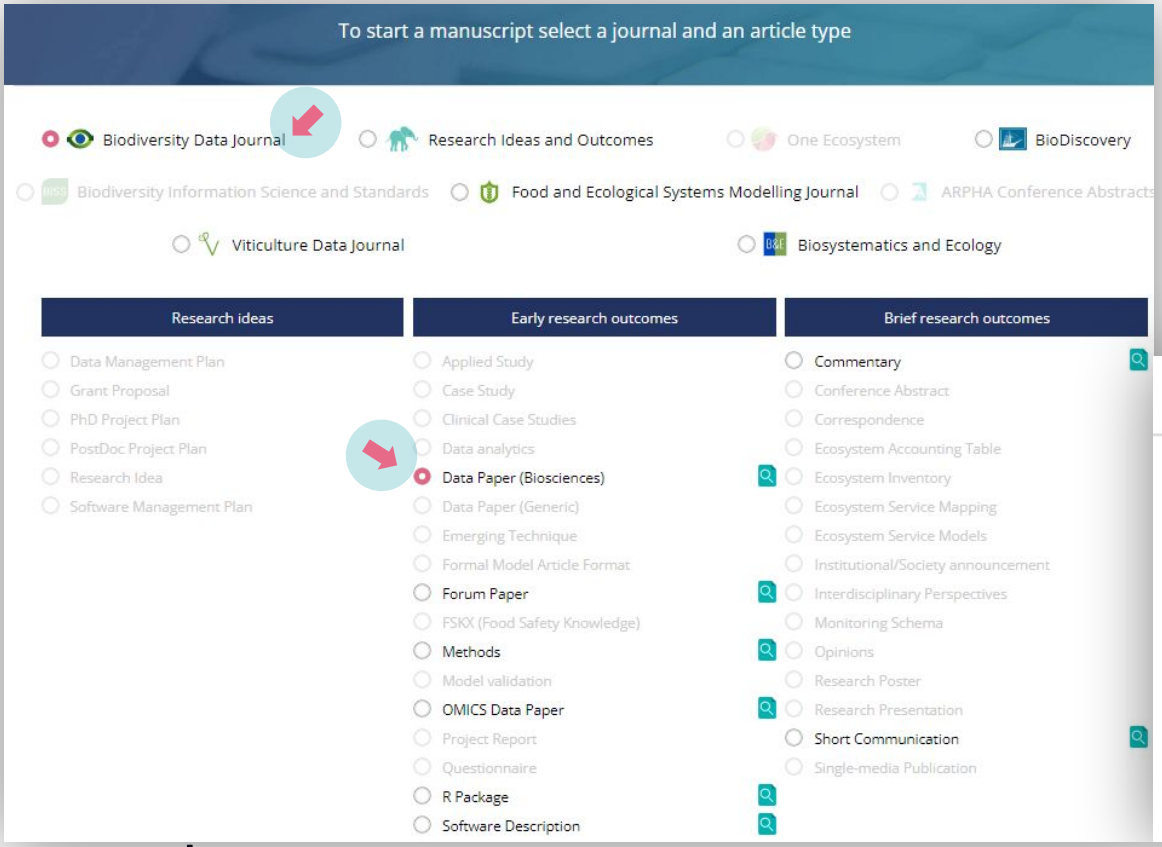
Get shareable link

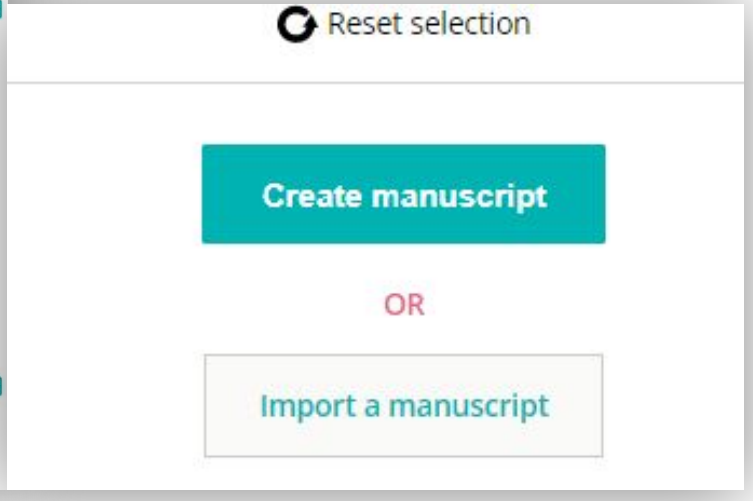
Publishing data papers on GBIF <-> ARPHA



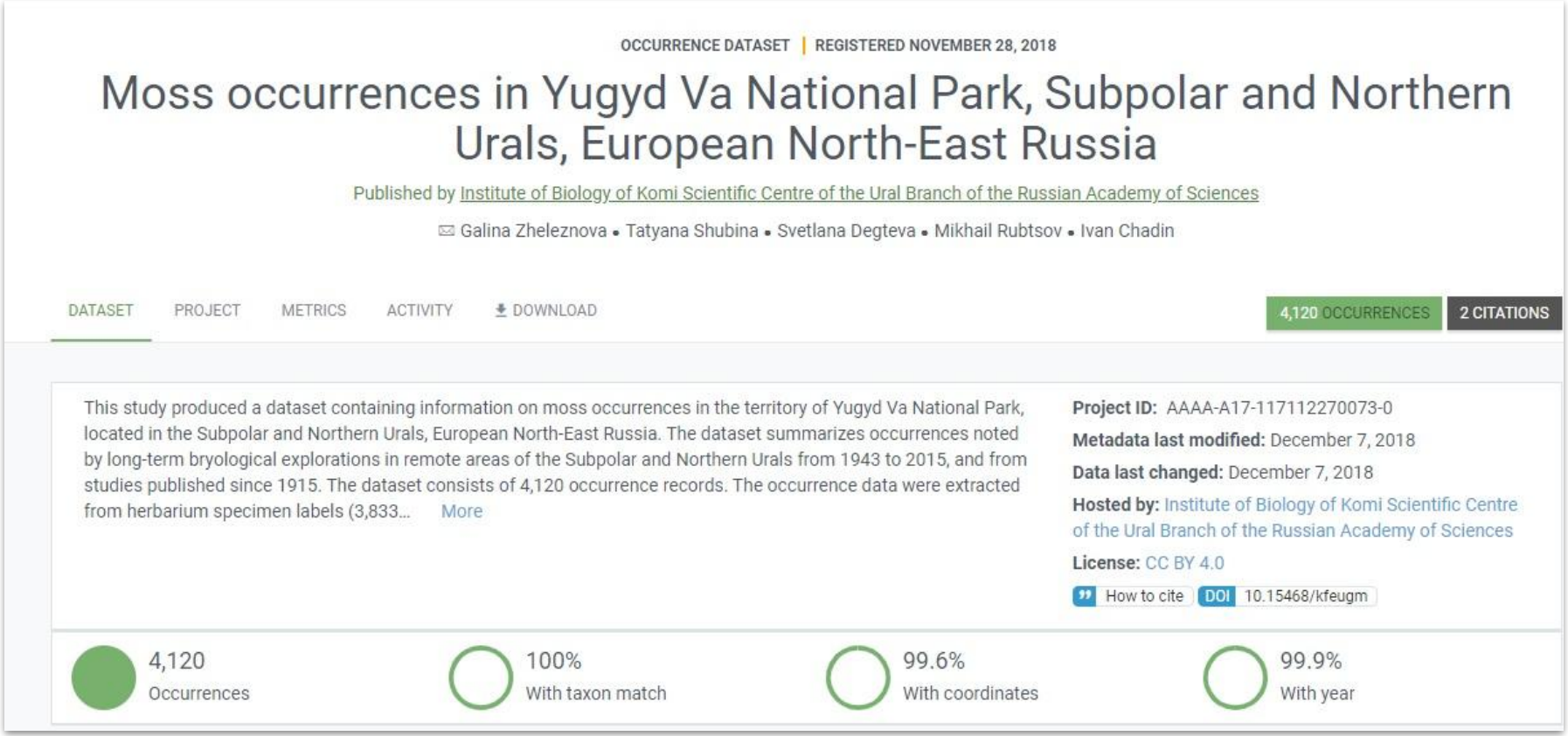
Conversion of EML metadata into data paper manuscripts

Start a manuscript in ARPHA Writing Tool





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

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

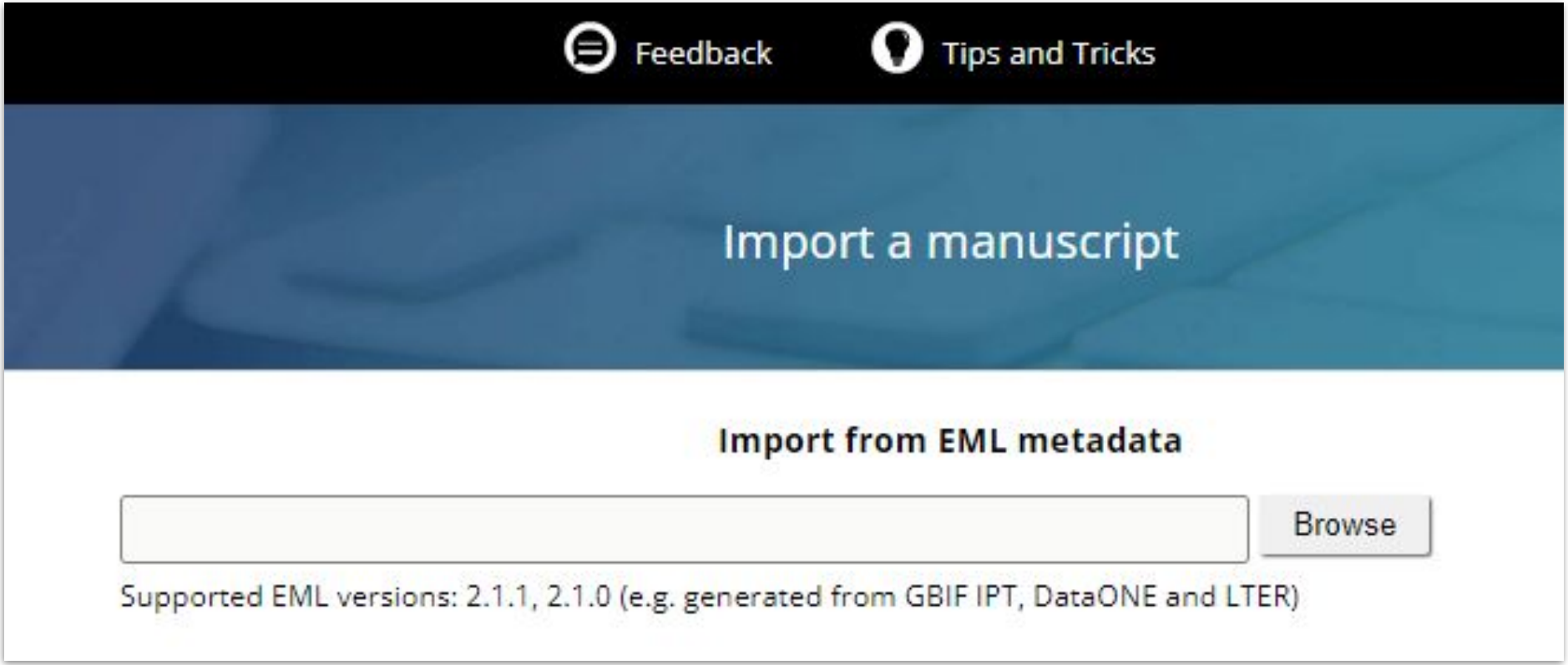
Last successful ingestion: April 22, 2020 (Not modified)

Last ingestion with changes: December 7, 2018

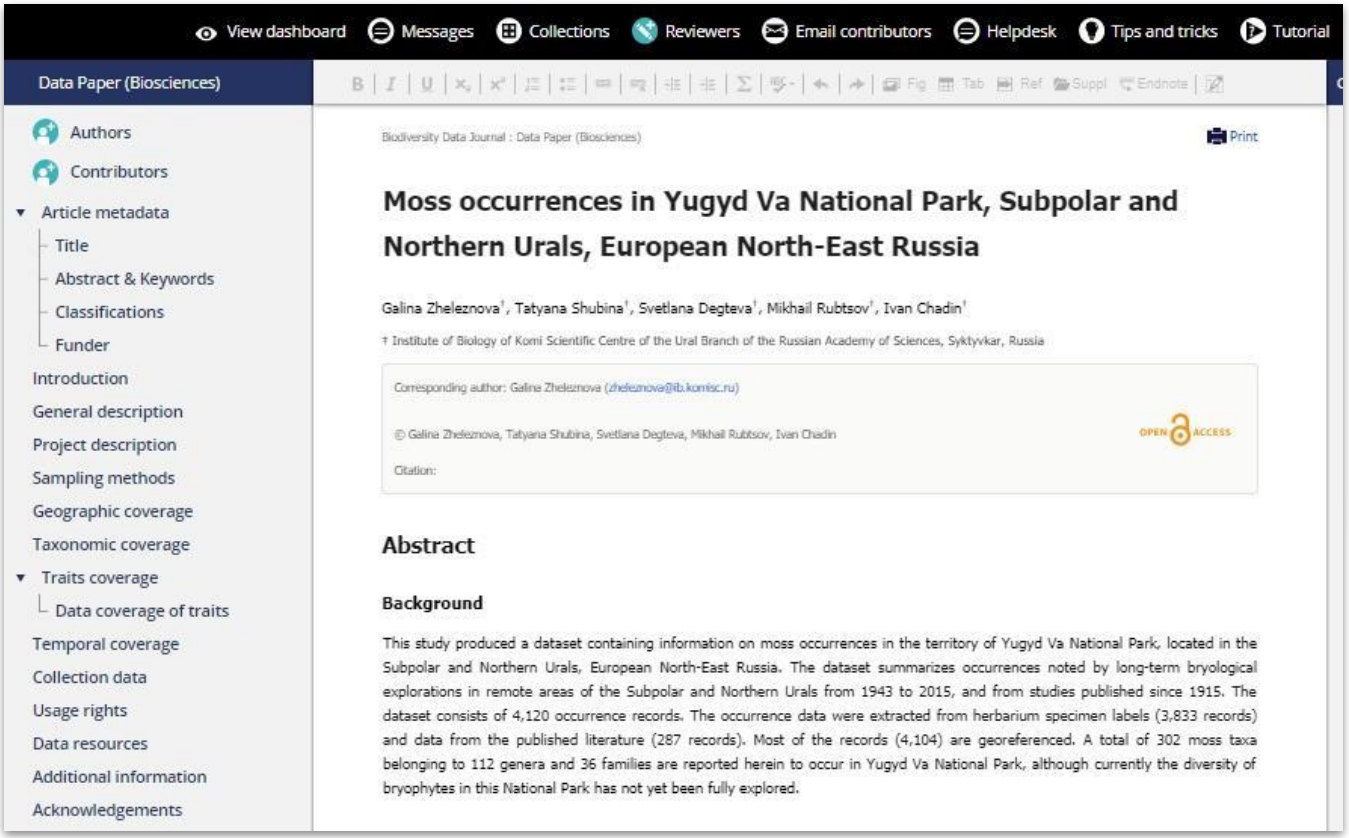
Last ingestion with data change: December 7, 2018

Occurrences in last ingestion: 4.120

Import your manuscript from GBIF EML file

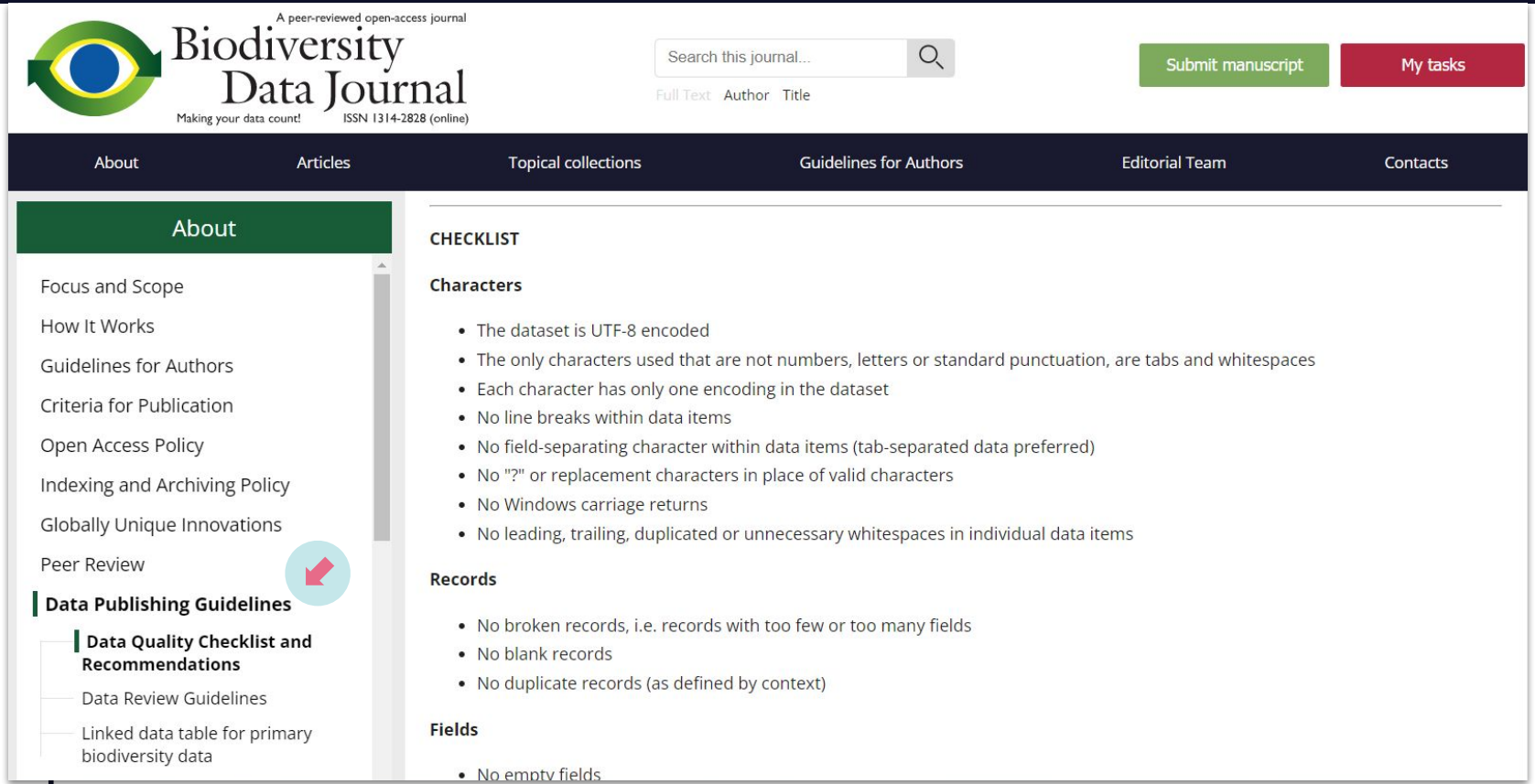


Your manuscript in ARPHA Writing Tool

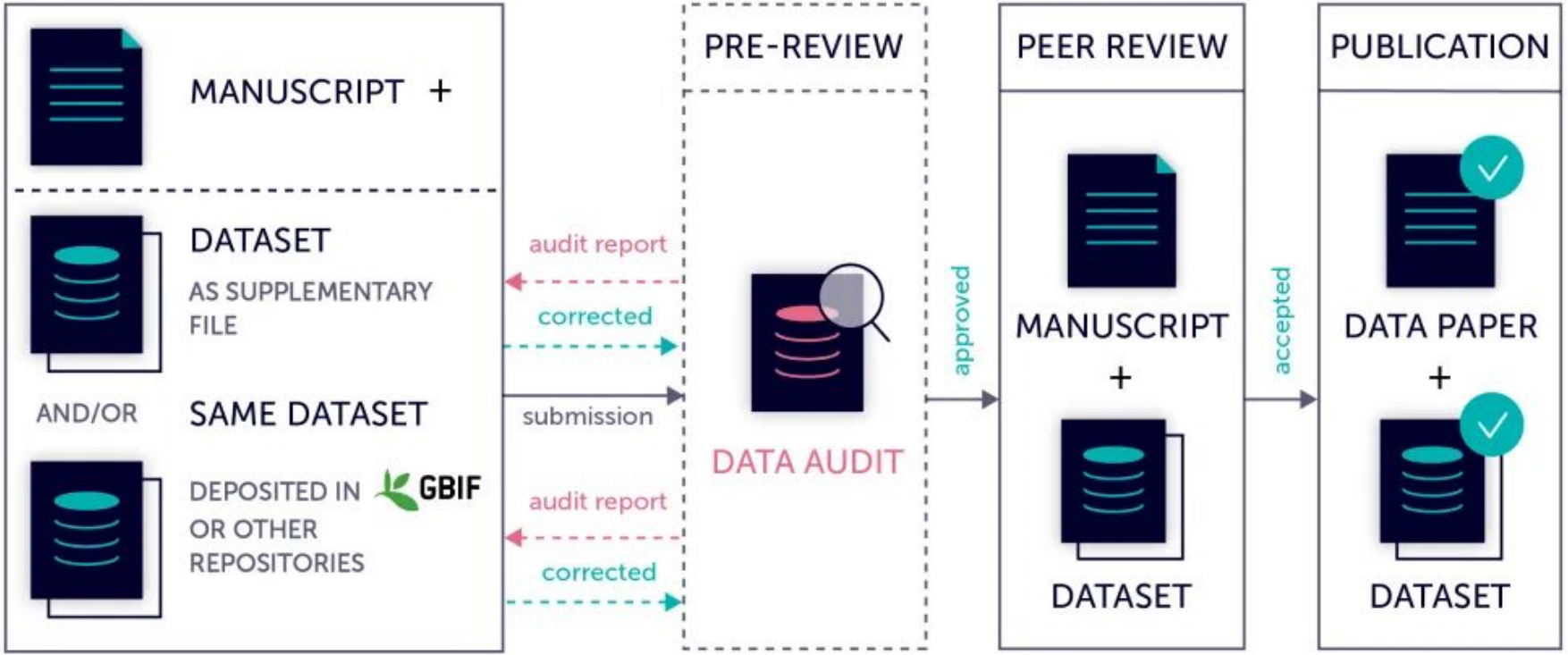


EML metadata converted into data paper manuscript

Author-performed data check before submission



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	37,91560º
1	40.9449º
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)

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 (MCMCMC) 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 (PP) ≥ 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. Download as CSV

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

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

XML PDF 0

Contents | Article info | Citation | Metrics | Comment | Related

Figs | Tabs | Map | Taxa | Refs | Cited

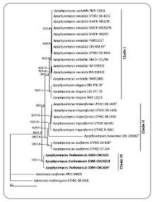


Figure 1. doi Download

Phylogenetic tree derived from maximum likelihood analysis of a combined ITS, LSU, and H3 genes of 28 sequences. *Saksenaea 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 ≥ 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.

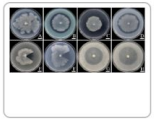


Figure 2. doi Download

Solubilization of non-soluble minerals in agar media by *Apophysomyces thailandensis* SDBR-CMUS26 (holotype). **A** $Ca_3(PO_4)_2$ **B** $CuCO_3 \cdot Cu(OH)_2$ **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.

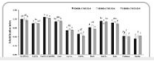
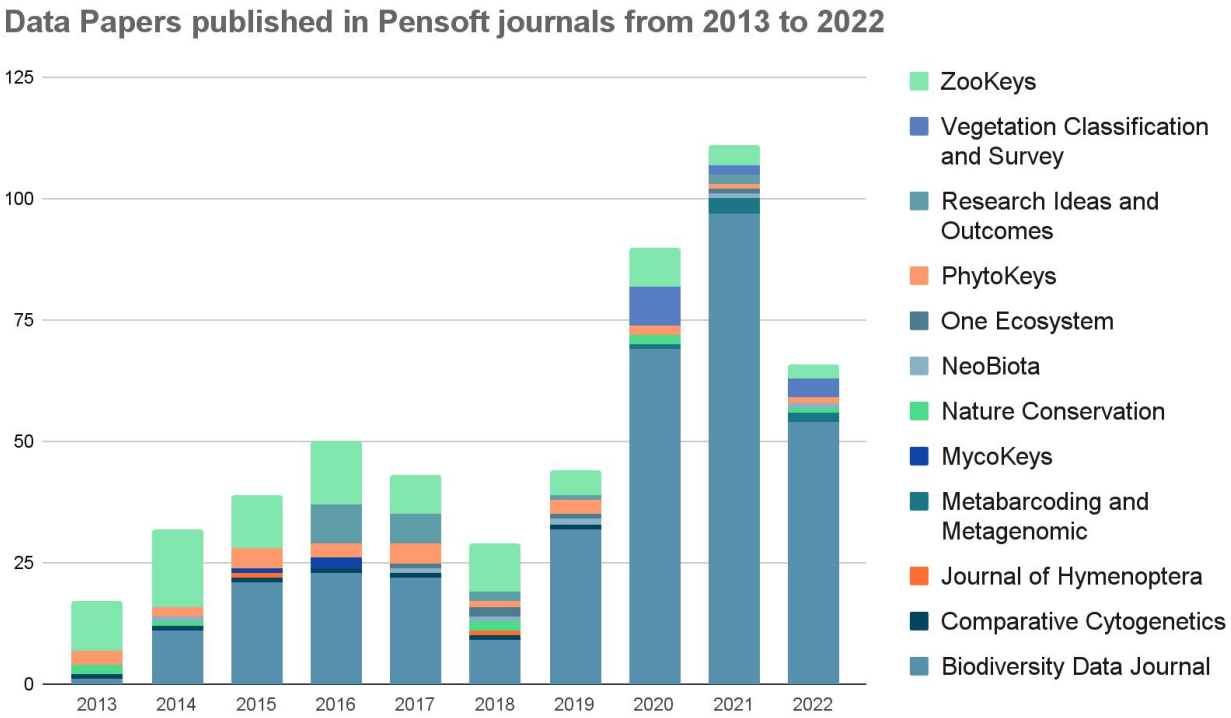


Figure 3. doi Download

Powered by arpha

Instantly growing popularity of Data Papers

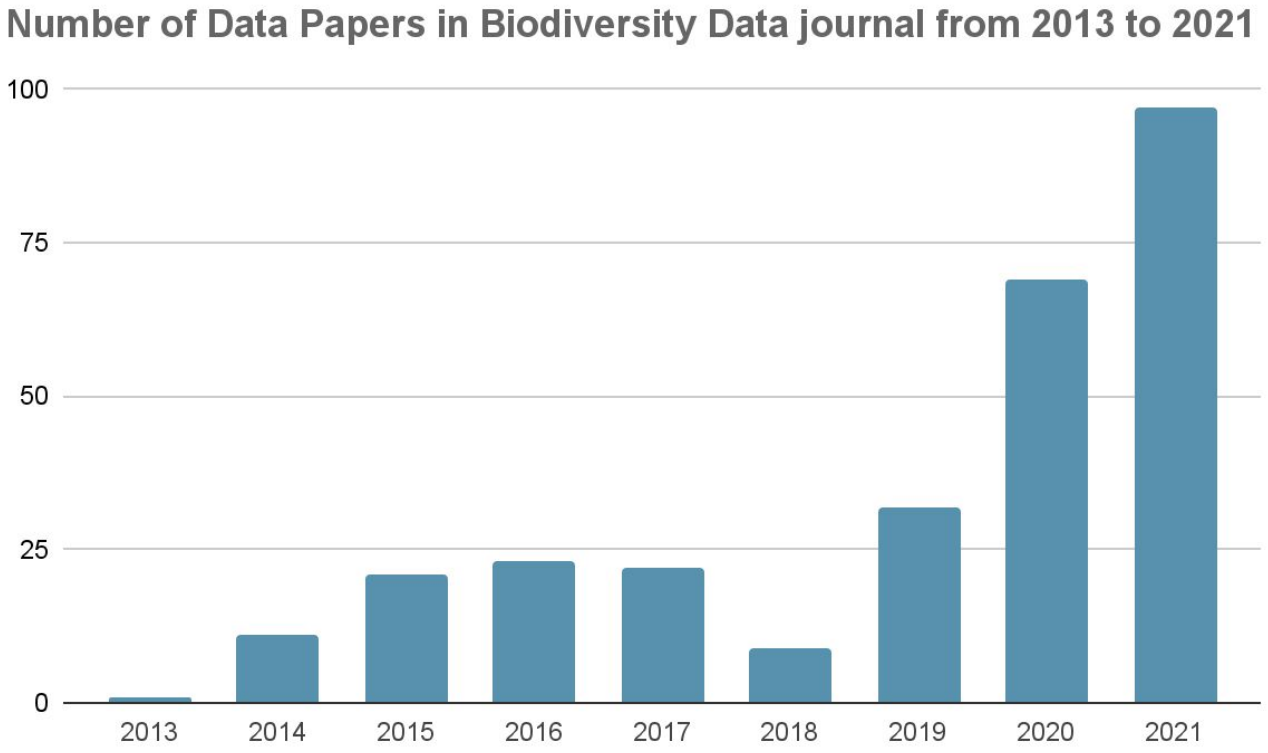
Data Paper statistics across Pensoft journals



Instantly growing popularity of Data Papers

Data Paper statistics in the Biodiversity Data Journal

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



Previous joint projects of GBIF & BDJ

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

Search this journal... Full Text Author Title

Submit manuscript

About | Articles | Topical collections | Guidelines for Authors | Editorial Team | Contacts

In this collection

Papers published: **59**

Documents added: **59**

Printed version: **Paperback**

OPEN ACCESS

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

Previous joint projects of GBIF & BDJ

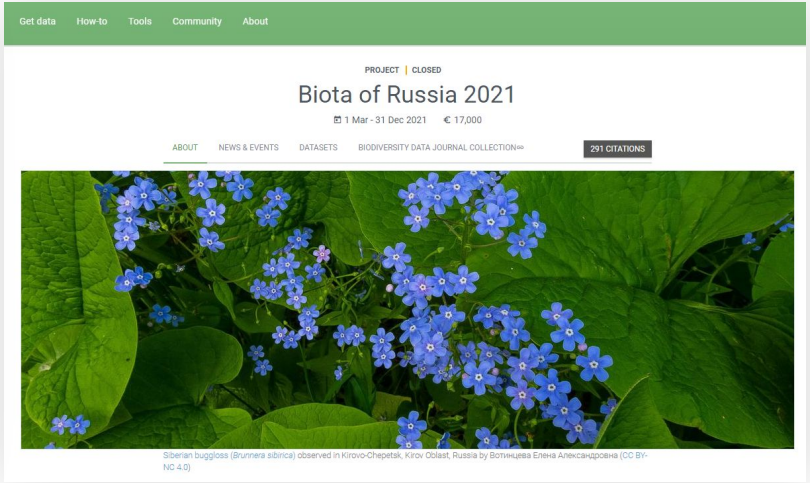
- 59 data papers published in 2020-2021 in the Biodiversity Data Journal's special collection "Biota of Russia"
- 58,720 total unique views (93,521 total views)

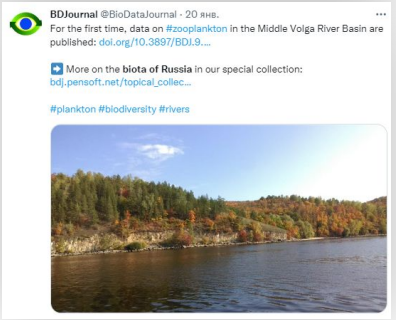
TOP-3 most popular papers according to the number of unique views:

- "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

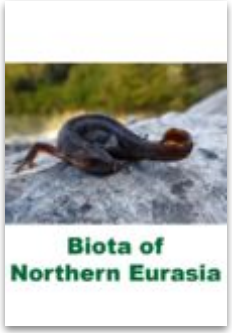
Two-years results

- Each article is promoted on Twitter and Facebook
- Joint Pensoft-GBIF promotion across social media and blogs





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!

PENSOFT

GBIF