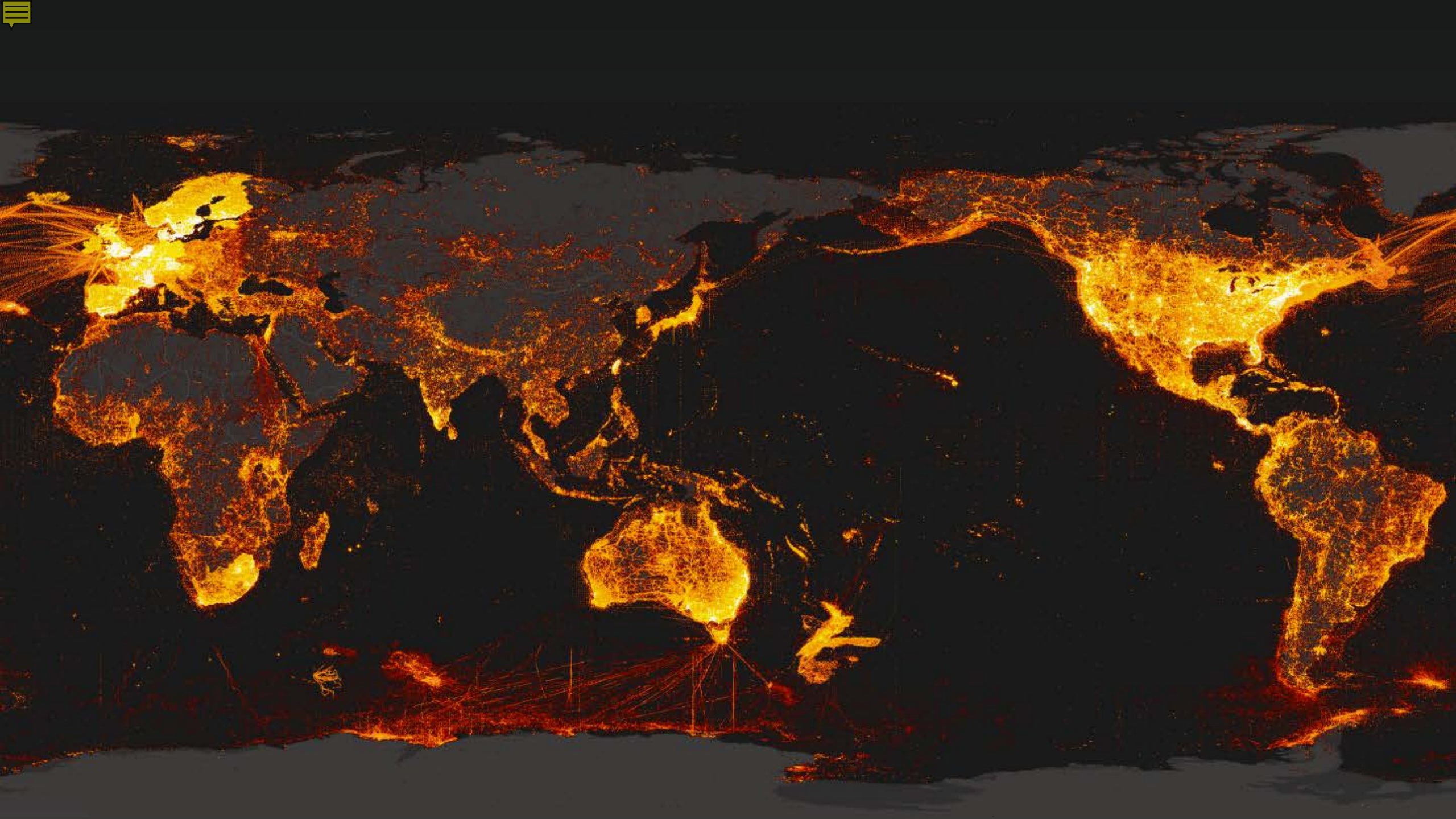


GBIF Global Biodiversity Information Facility

# How to Publish a Data Paper

**Dmitry Schigel**
GBIF Scientific officer

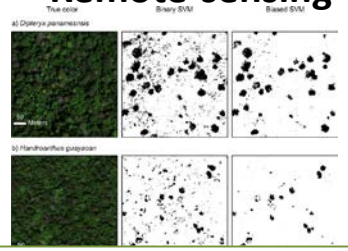
Photo: *Meripilus giganteus*, 18 Sep 2016
*https://flic.kr/p/LV1yEu*

# STREAMS OF SPECIES OCCURRENCE DATA

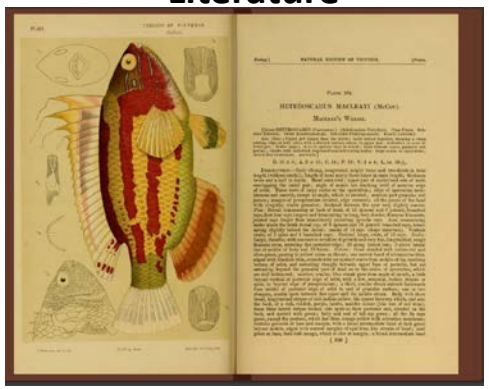
GBIF Global Biodiversity Information Facility

## Sequences

## Remote-sensing

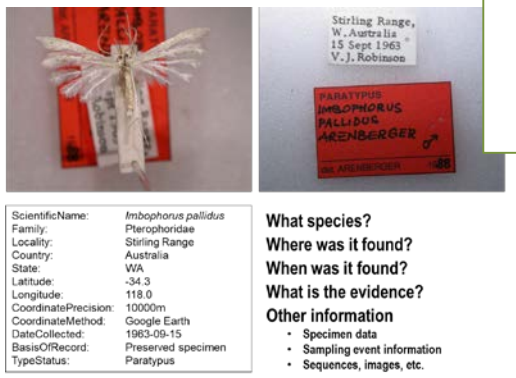


## Literature



What species?
Where was it found?
When was it found?
What is the evidence?
Other information

- Specimen data
- Sampling event information
- Sequences, images, etc.

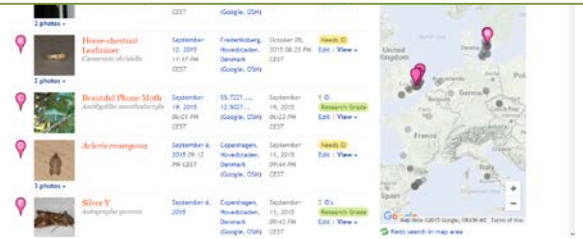



| | |
|---|---|
| ScientificName: | *Imbophorus pallidus* |
| Family: | Pterophoridae |
| Locality: | Stirling Range |
| Country: | Australia |
| State: | WA |
| Latitude: | -34.3 |
| Longitude: | 118.0 |
| CoordinatePrecision: | 10000m |
| CoordinateMethod: | Google Earth |
| DateCollected: | 1983-09-15 |
| BasisOfRecord: | Preserved specimen |
| TypeStatus: | Paratypus |

What species?
Where was it found?
When was it found?
What is the evidence?
Other information

- Specimen data
- Sampling event information
- Sequences, images, etc.

## Specimens



## Citizen science

## Animal tracking

http://e360.yale.edu/features/dna_technology_discovering_new_species - SUZE Bateson, University of Guelph
doi: 10.1371/journal.pone.0118403 / http://www.biodiversitylibrary.org/bibliography/4821

GBIF

# Types of data shared through GBIF

what

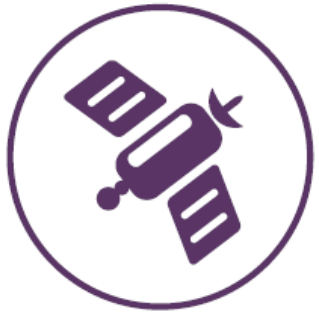
Specimen

Fossil

Material sample

Literature occurrence

Observation
Human observation
Living specimen



Machine observation

**sample-based data**

**Barcoding**
**Metagenomics**
**eDNA**

https://www.gbif.org/dataset-classes

# What is a data paper?

A data paper is a peer reviewed document describing a dataset, published in a peer reviewed journal. Data papers provide recognition for this effort by means of a scholarly article.

Searchable metadata document describing a dataset or groups of datasets

! no interpretation, analyses, or hypothesis testing

= data table + data story + journal link

# The data paper concept

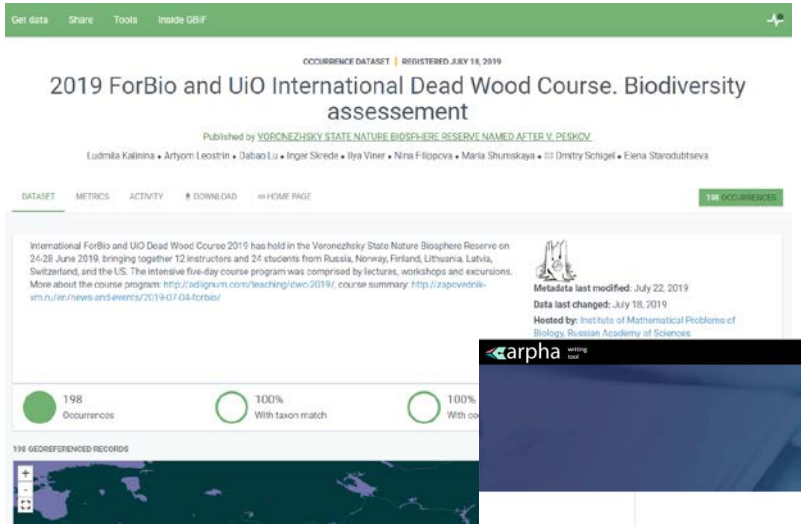
- Scholarly publication, established  scientific process
- Rigorous data stewardship to create and curate data as first class objects; peer-review.
- Recognize data publication as traditional scientific articles – impact factor, citation, etc.

# Three main actors in data paper process

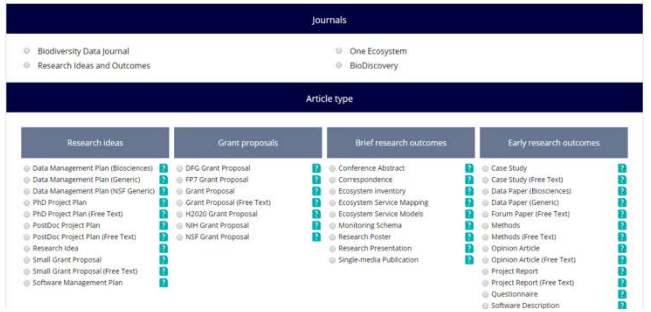
1. **Authors** – researchers, scholars, citizen scientists
2. **Publishers** – journals
3. **Data repositories** – e.g., GBIF, Figshare, Dryad, Zenodo, etc.

# Automatically generate a data paper: example

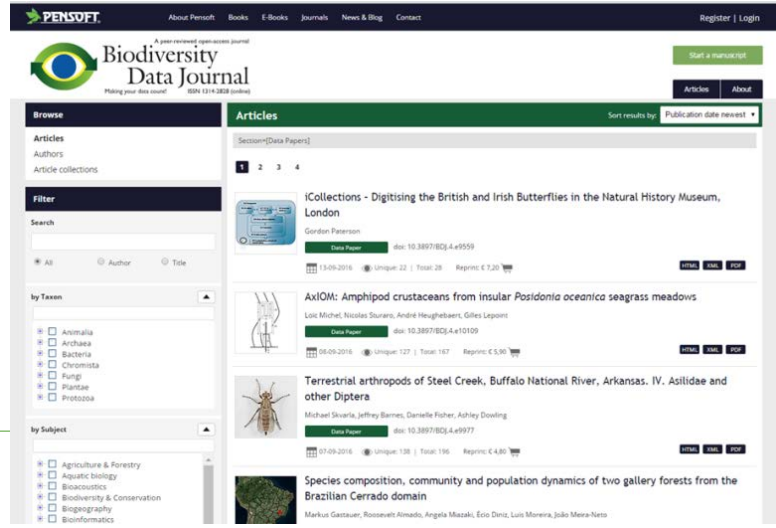
1. GBIF




2. Arpha


3. BDJ



GBIF | Global Biodiversity Information Facility

Intergovernmental open data infrastructure

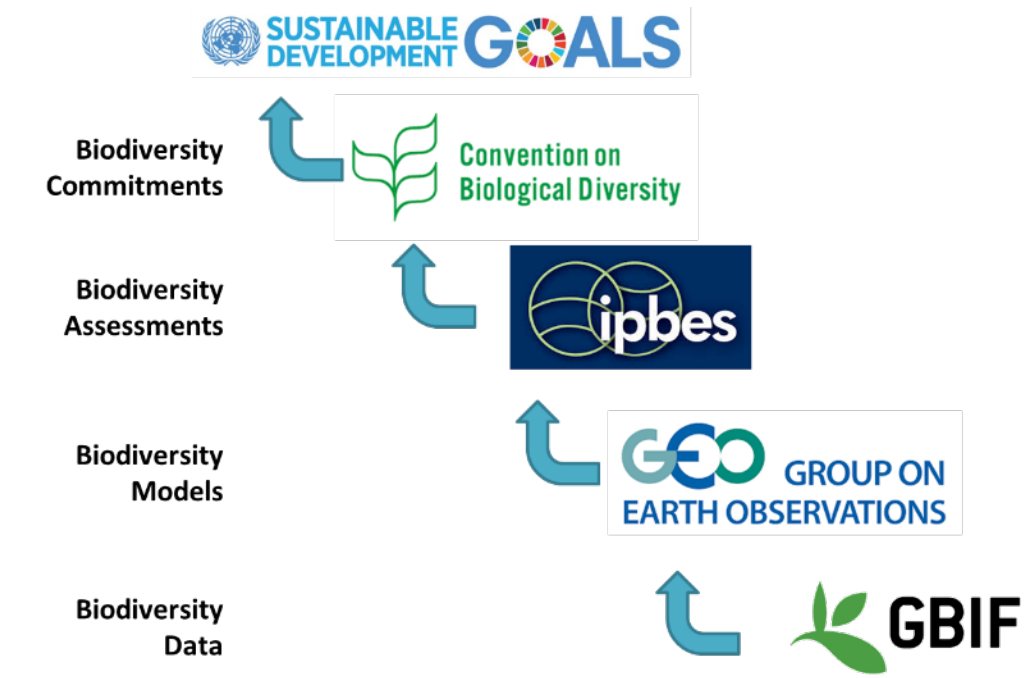
Established in 2001, OECD recommendation

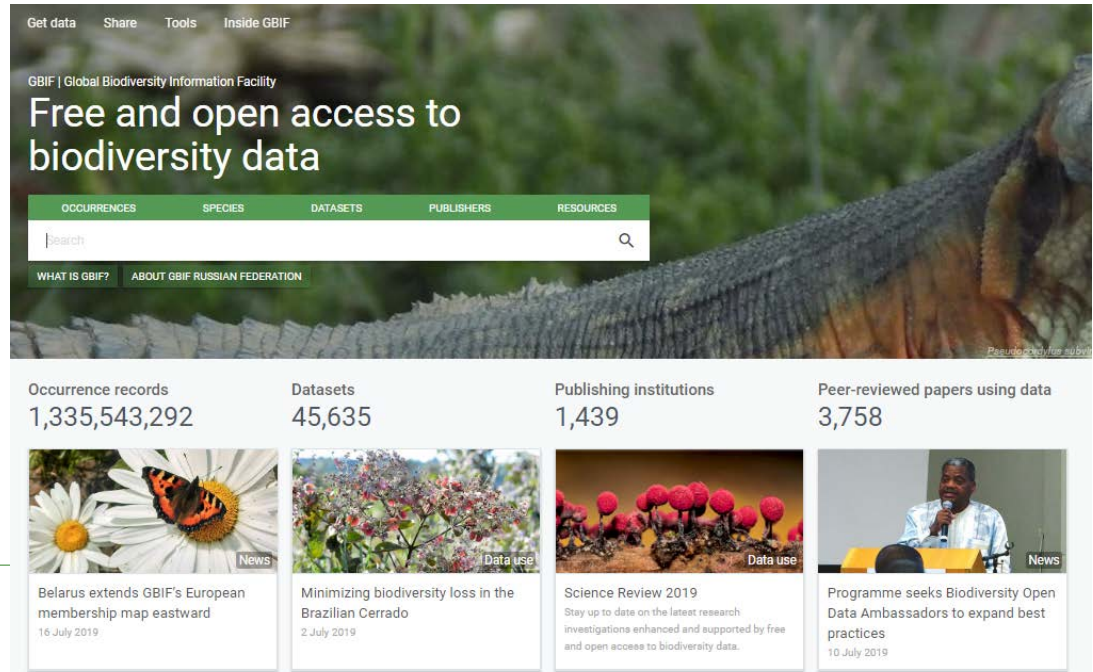
Network for **free and open access**
to biodiversity data from all countries

Voluntary memorandum of understanding

Funds: governments of the participant countries

96 participants: 58 countries and 38 organizations




*https://www.gbif.org/the-gbif-network*

GBIF.org

Get data Share Tools Inside GBIF

GBIF | Global Biodiversity Information Facility

# Free and open access to biodiversity data

OCCURRENCES SPECIES DATASETS PUBLISHERS RESOURCES

Search

WHAT IS GBIF? ABOUT GBIF RUSSIAN FEDERATION

Occurrence records
1,335,543,292

Datasets
45,635

Publishing institutions
1,439

Peer-reviewed papers using data
3,758


Belarus extends GBIF's European membership map eastward
16 July 2019


Minimizing biodiversity loss in the Brazilian Cerrado
2 July 2019


Science Review 2019
Stay up to date on the latest research investigations enhanced and supported by free and open access to biodiversity data.


Programme seeks Biodiversity Open Data Ambassadors to expand best practices
10 July 2019


BIFA funds nine Asian data mobilization projects
26 June 2019

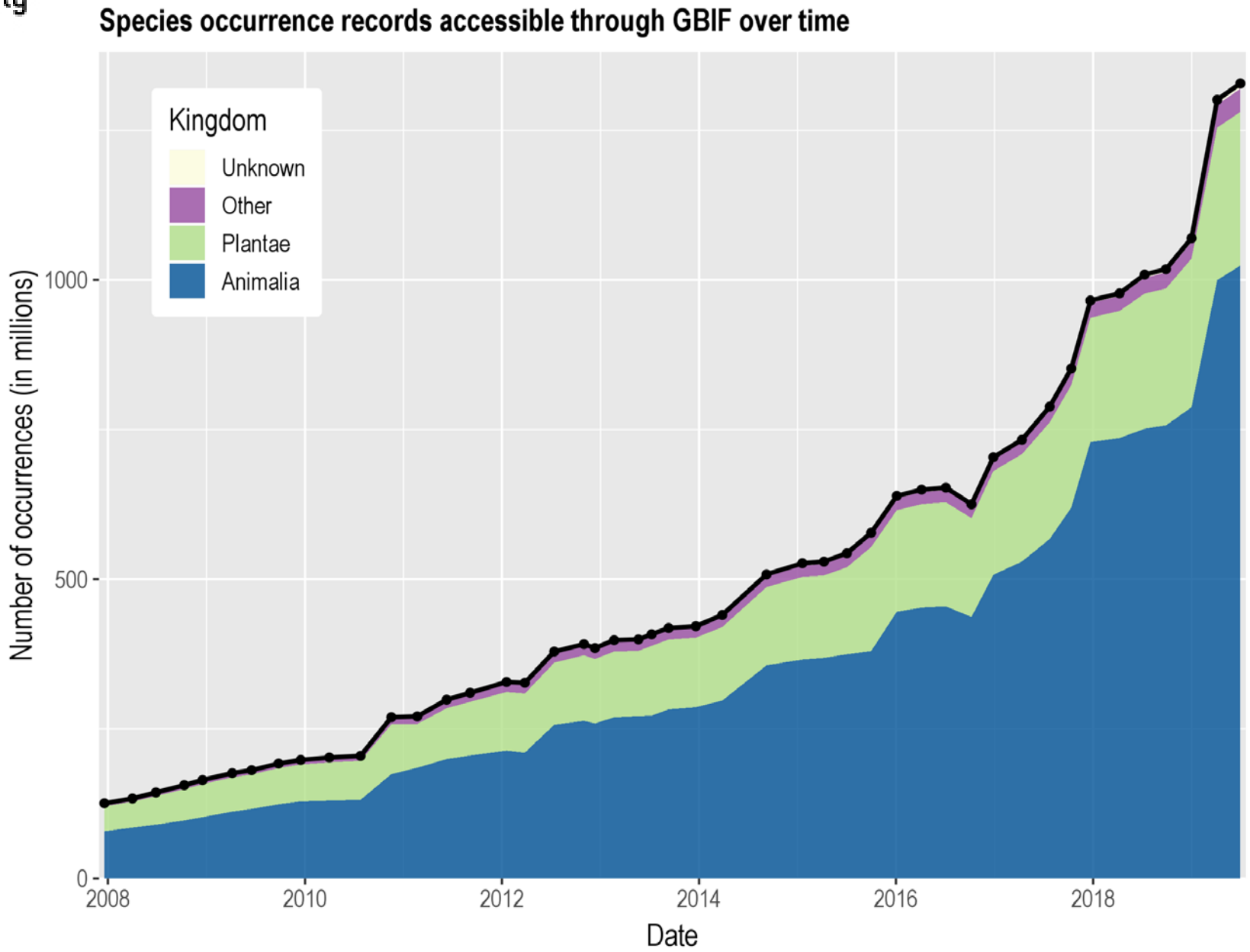

Atlantodesmus sierwaldae sp. nov.
Cladistic analysis and description of a new species of the Brazilian genus Atlantodesmus Hoffman, 2000 (Diplopoda: Polydesmida: Chelodesmidae)


The GBIF network
Dozens of countries and organizations working together to make species data findable, accessible, interoperable and reusable.


Establishing a national biodiversity information facility in Chile
Experiences of setting up and running a node in Latin America

GBIF Global Biodiversity Information Facility

# Data published through GBIF.org




data availability

www.gbif.org/analytics/global

# GBIF

## BY THE NUMBERS

*7 Aug 2019*

*Species occurrence records*

**1,335,646,137**

*Datasets*

**45,632**

*Country Participants*

**58**

*Organizational Participants*

**38**

*Publishers*

**1,439**

# Data citation: tracking and display

GBIF Global Biodiversity Information Facility

data access and use

ure, how it's discovered and linked to GBIF-mediated data.

**Leptodactylus mystaceus (Spix, 1824): Distribution extension for the Brazilian Cerrado (Anura: Leptodactylidae)** Literature

Andreani, T. Borges, R. Santos, L. (2017) Herpetology Notes

Fifty-seven species of anurans belonging to the genus Leptodactylus are regi… and widely distributed throughout Brazil (Sá et al., 2014; SBH, 2016). Among these species, Leptodactylus mystace… …x, 1824) spreads throughout the whole Northern region and some areas of northeastern and central …

Journal Article

Data used in study DOI 10.15468/dl.s1xtf6

**DO ENVIRONMENTAL FACTORS AFFECT THE TAXONOMIC RELIABILITY OF LEAF CUTICULAR MICROMORPHOLOGICAL CHARACTERS? A CASE STUDY IN PODOCA… ACEAE** Literature

Clugston, J. Jeffree, C. Ahrends, A. Mill, R. (2017) Edinburgh Journal of Botany

Leaf cuticle micromorphology has been cited as an important set of taxonomic char… in gymnosperms, but previous studies have largely been based on small sample sizes. The premise of this study was to u… stand whether external factors affect cuticular micromorphology of Podocarpaceae. Two exam…

Journal Article

Data used in study DOI 10.15468/6e8nje DOI 10.15468/bkzv1l DOI 10.15468/dgbpla DOI 10.15468/dl… DOI 10.15468/hja69f DOI 10.15468/hnhrg3 DOI 10.15468/i9bj5r DOI 10.15468/ib5ypt DOI 10.15468/ly60bx DOI 10.15468/mug7kr DOI 10.15468/nc6rxy DOI 10.15468/ucmdjy DOI 10.15468/vtfbe3 DOI 10.15468/x5ucvh DOI 10.15468/xy0eoi DOI 10.15468/yo3mmu DOI 10.15468/ypoair DOI 10.15468/…

**One for each ocean: revision of the Bursa granularis (Röding… 798) species complex (Gastropoda: Tonnoidea: Bursidae)** Literature

Sanders, M. Merle, D. Bouchet, P. Castelin, M. Beu, A. Samadi, … (2017) Journal of Molluscan Studies

Bursa granularis (Röding, 1798) is a tonnoidean gastrop… that is regarded as broadly distributed throughout the Indo-Pacific and tropical western Atlantic. Because of its variable sh… has received no less than thirteen names, now all synonymized under the name B. granularis. We sequenced a fra…

Journal Article

Data used in study DOI 10.15468/dl.ma8dle

OCCURRENCE DATASET | REGISTERED OCTOBER 13, 2016

## NMNH Extant Specimen Records

Published by National Museum of Natural History, Smithsonian Institution

Thomas Orrell

DATASET METRICS ACTIVITY DOWNLOAD HOME PAGE 7,571,188 OCCURRENCES 218 CITATIONS

Public records of accessioned specimens and observations curated by the National Museum of Natural History, Smithsonian Institution. These data are from the Departments of Botany, Entomology, Invertebrate Zoology, Vertebrate Zoology (Amphibians & Reptiles, Birds, Fishes, and Mammals) and include more than … primary type specimen records.

Metadata last modified … April 16, 2019

Data last changed … April 16, 2019

Hosted by … National Museum of Natural History, Smithsonian Institution

License: …

How to cite DOI 10.15468/hnhrg3

7,571,188 Occurrences

99.7% … exon match

…8% With coordinates

79% With year

…83 GEOREFERENCED RECORDS

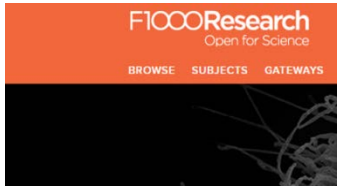
…ated 11 hours ago © OpenStreetMap contributors, © OpenMapTiles, GBIF

…year 1653 - 2018 EXPLORE

…03 OCCURRENCES WITH IMAGES

SEE GALLERY

# Diversity of biodiversity data journals




## Journals that publish biodiversity-related data papers

Please notify us at communication@gbif.org with corrections or updates.

| Journal | Publisher | Open Access (license) | APC estimate | Journal Impact Factor (2017) | Scopus CiteScore (2017) |
|---|---|---|---|---|---|
| Arxius de Miscel·lània Zoològica | Nat Hist Museum of Barcelona | CC BY | ? | - | 0.31 |
| Biodiversity Data Journal | Pensoft | CC BY | ? | - | 0.96 |
| Biogeographia | eScholarship | CC BY | ? | - | - |
| BioInvasions Records | REABIC Journals | CC BY | EUR 300 | 1.189 | 1.35 |
| BioRisk | Pensoft | Yes | EUR 300 | - | 6.00 |
| Biota Colombiana | Humboldt Institute, Colombia | CC BY-NC-ND | ? | - | - |
| BMC Ecology | Biomed Central | CC BY | GBP 1370 | 2.315 | 2.35 |

At least 30 journals

Impact Factor 0–8.7

Scopus CiteScore 0–8.1

APC 0–2014 USD

GBIF

# Benefits of openess

OPEN ACCESS Freely available online

PLoS one

## Sharing Detailed Research Data Is Associated with Increased Citation Rate

Heather A. Piwowar*, Roger S. Day, Douglas B. Fridsma

Department of Biomedical Informatics, University of Pittsburgh School of Medicine, Pittsburgh, Pennsylvania, United States of America

*Background.* Sharing research data provides benefit to the general scientific community, but the benefit is less obvious for the investigator who makes his or her data available. *Principal Findings.* We examined the citation history of 85 cancer microarray clinical trial publications with respect to the availability of their data. The 48% of trials with publicly available microarray data received 85% of the aggregate citations. Publicly available data was significantly ($p = 0.006$) associated with a 69% increase in citations, independently of journal impact factor, date of publication, and author country of origin using linear regression. *Significance.* This correlation between publicly available data and increased literature impact may further motivate investigators to share their detailed research data.

Citation: Piwowar HA, Day RS, Fridsma DB (2007) Sharing Detailed Research Data Is Associated with Increased Citation Rate. PLoS ONE 2(3): e308. doi:10.1371/journal.pone.0000308

INTRODUCTION

Sharing information facilitates science. Publicly sharing detailed research data–sample attributes, clinical factors, patient outcomes, DNA sequences, raw mRNA microarray measurements–with other researchers allows these valuable resources to contribute far beyond their original analysis[1]. In addition to being used to confirm original results, raw data can be used to explore related or new hypotheses, particularly when combined with other publicly available data sets. Real data is indispensable when investigating and developing study methods, analysis techniques, and software implementations. The larger scientific community also benefits: sharing data encourages multiple perspectives, helps to identify errors, discourages fraud, is useful for training new researchers, and increases efficient use of funding and patient population resources by avoiding duplicate data collection.

Believing that that these benefits outweigh the costs of sharing research data, many initiatives actively encourage investigators to make their data available. Some journals, including the *PLoS* family, require the submission of detailed biomedical data to publicly available databases as a condition of publication[2–4]. Since 2003, the NIH has required a data sharing plan for all large funding grants. The growing open-access publishing movement will perhaps increase peer pressure to share data.

However, while the general research community benefits from shared data, much of the burden for sharing the data falls to the study investigator. Are there benefits for the investigators themselves?

A currency of value to many investigators is the number of times their publications are cited. Although limited as a proxy for the scientific contribution of a paper[5], citation counts are often used in research funding and promotion decisions and have even been assigned a salary-increase dollar value[6]. Boosting citation rate is thus is a potentially important motivator for publication authors.

In this study, we explored the relationship between the citation rate of a publication and whether its data was made publicly available. Using cancer microarray clinical trials, we addressed the following questions: Do trials which share their microarray data receive more citations? Is this true even within lower profile trials? What other data-sharing variables are associated with an increased citation rate? While this study is not able to investigate causation, quantifying associations is a valuable first step in understanding these relationships. Clinical microarray data provides a useful environment for the investigation: despite being valuable for reuse and extremely costly to collect, is not yet universally shared.

RESULTS

We studied the citations of 85 cancer microarray clinical trials published between January 1999 and April 2003, as identified in a systematic review by Ntzani and Ioannidis[7] and listed in Supplementary Text S1. We found 41 of the 85 clinical trials (48%) made their microarray data publicly available on the internet. Most data sets were located on lab websites (28), with a few found on publisher websites (4), or within public databases (6 in the Stanford Microarray Database (SMD)[8], 6 in Gene Expression Omnibus (GEO)[9], 2 in ArrayExpress[10], 2 in the NCI GeneExpression Data Portal (GEDP)(gedp.nci.nih.gov); some datasets in more than one location). The internet locations of the datasets are listed in Supplementary Text S2. The majority of datasets were made available concurrently with the trial publication, as illustrated within the WayBackMachine internet archives (www.archive.org/web/web.php) for 25 of the datasets and mention of supplementary data within the trial publication itself for 10 of the remaining 16 datasets. As seen in Table 1, trials published in high impact journals, prior to 2001, or with US authors were more likely to share their data.

The cohort of 85 trials was cited an aggregate of 6239 times in 2004–2005 by 3153 distinct articles (median of 1.0 cohort citation per article, range 1–23). The 48% of trials which shared their data received a total of 5334 citations (85% of aggregate), distributed as shown in Figure 1.

Academic Editor: John Ioannidis, University of Ioannina School of Medicine, Greece

Received December 13, 2006; Accepted February 26, 2007; Published March 21, 2007



Funding: HAP was supported by NLM Training Grant Number 5T15-LM007059-19. The NIH had no role in study design, data collection or analysis, writing the paper, or the decision to submit it for publication. The publication contents are solely the responsibility of the authors and do not necessarily represent the official views of the NIH.

Competing Interests: The authors have declared that no competing interests exist.

* To whom correspondence should be addressed. E-mail: hpiwowar@cbmi.pitt.edu



- Increases the efficiency of research
- Promotes scholarly rigor and quality of research
- Enables tracking of data use and data citation through DOIs
- Expands the spectrum of academic products through data papers
- Enhances visibility and scope for engagement
- Enables researchers to ask new research questions
- Enhances collaboration and community-building
- Increases the economic and social impact of research
- International conventions and requirements from funding agencies

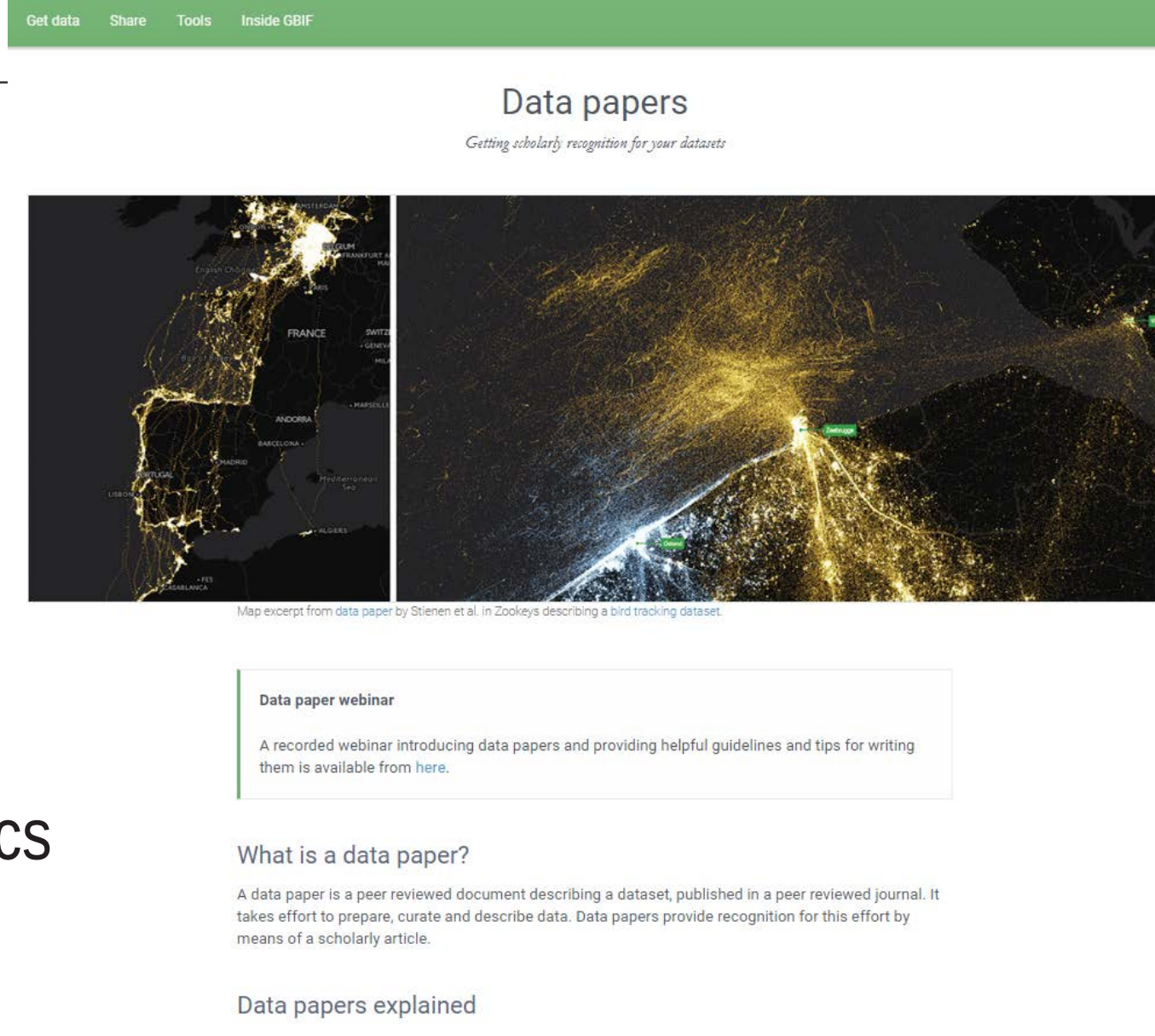
*Piwowar et al. 2007*

# Potential benefits of publishing data

- Enable validity and reproducibility of research
- Increase the visibility and use of research
- Research outputs to answer new questions
- Increase cost-efficiency and data reuse
- Access to new communities - public, policymakers, industry professionals, etc.

# GBIF.org: Data papers

- What is it?
- Webinar video
- Why publish?
- Data paper tools
- Journals with links, costs & metrics
- helpdesk@gbif.org




https://www.gbif.org/data-papers


# THANK YOU

dschigel@gbif.org

www.gbif.org