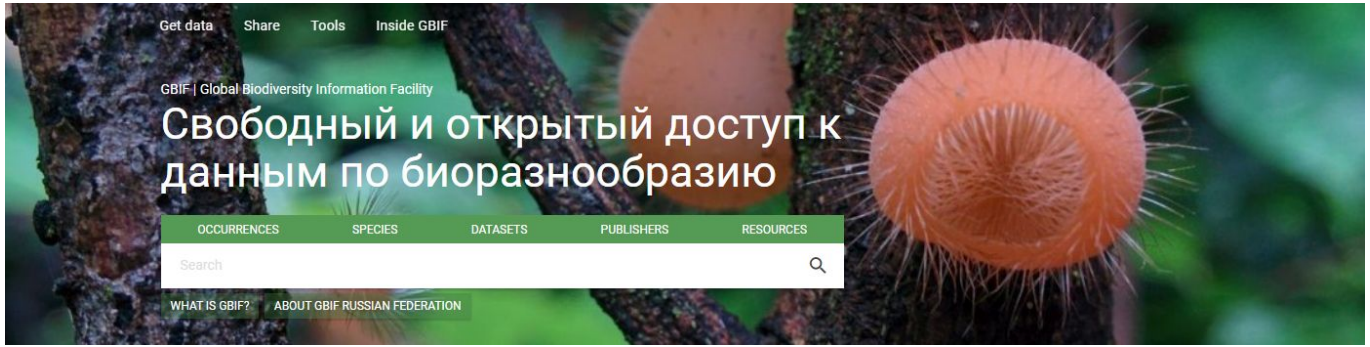
Разнообразие почвенных животных России: публикация и эффективное использование исходных данных
29–30 августа 2019 г., ИПЭЭ им. А.Н. Северцова РАН

Сессия 2.

Глобальная информационная система о биоразнообразии - GBIF

Максим Шашков

Институт математических проблем биологии РАН – филиал ИПМ им. М.В. Келдыша РАН
г. Пущино Московской области



Слайды CC BY:
GBIF Secretariat
и Максим Шашков

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

Pseudocondylus subvil

Occurrence records
1,335,543,292

Datasets
45,635

Publishing institutions
1,439

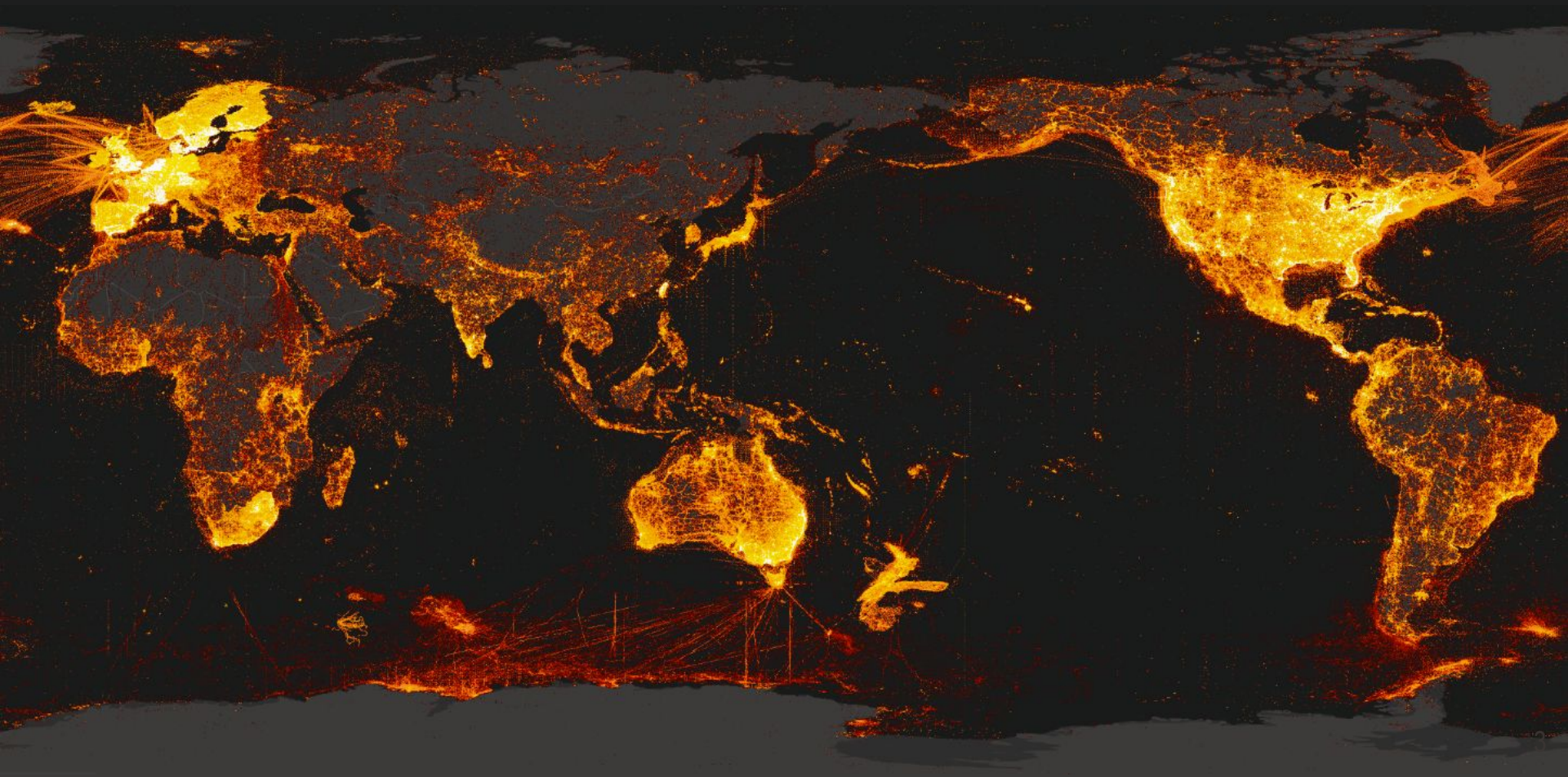
Peer-reviewed papers using data
3,758

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

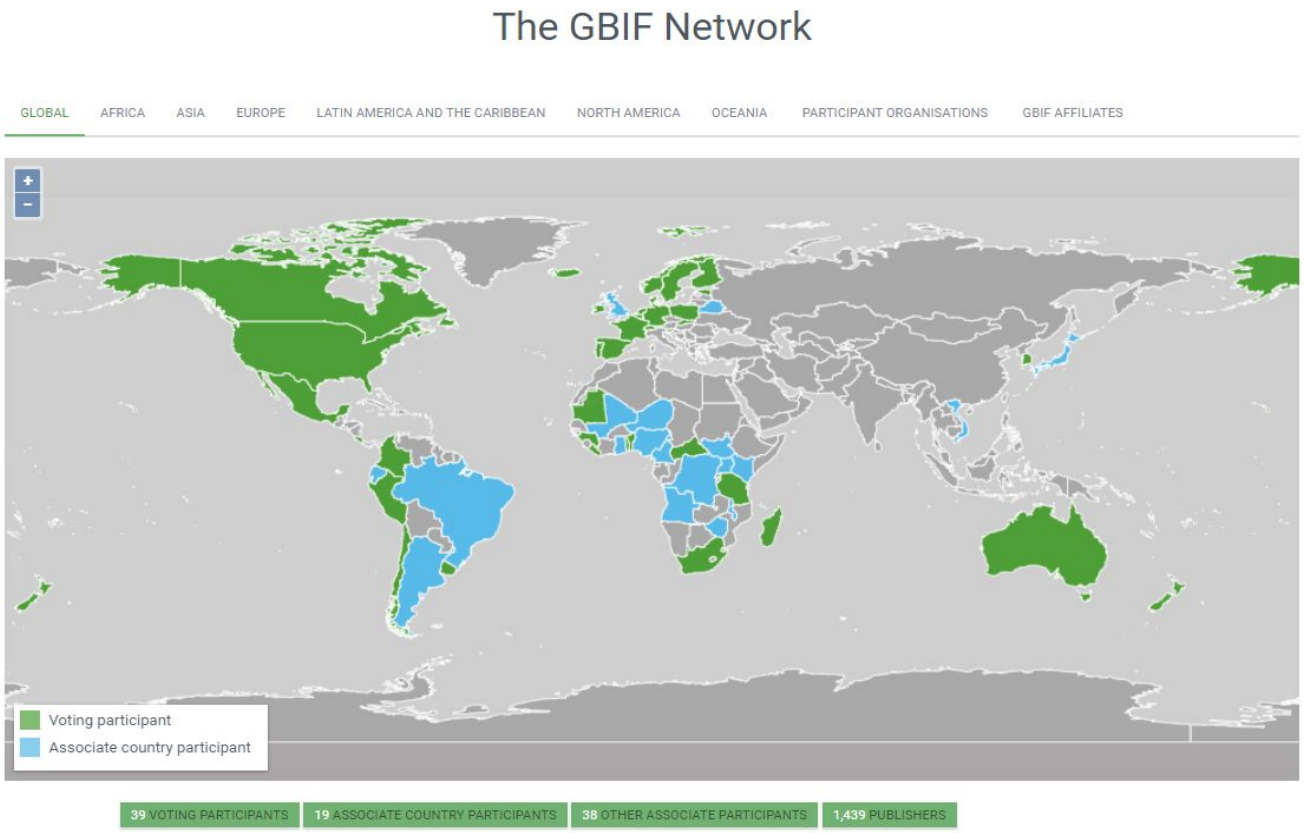


Страны-участники GBIF

Сеть GBIF

Недавно присоединилась **Беларусь**

Для того, чтобы публиковать данные, не обязательно работать в стране-участнике сети GBIF



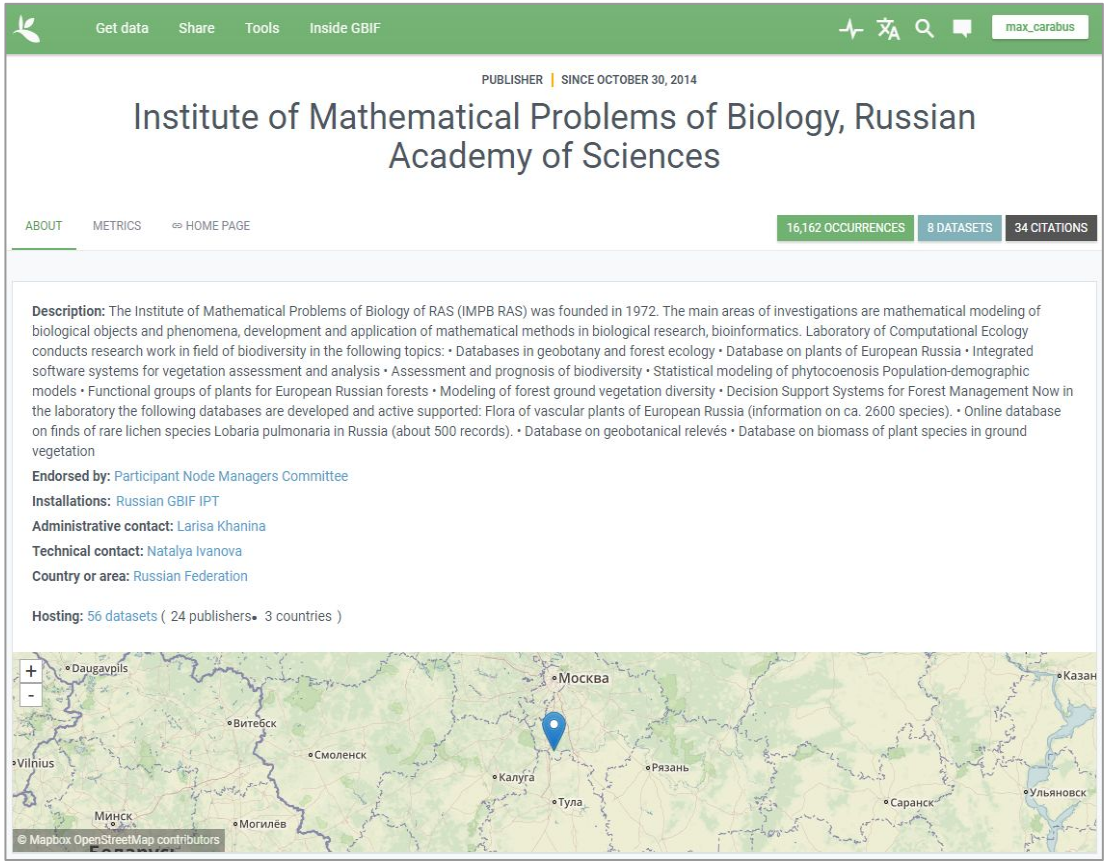
https://www.gbif.org/the-gbif-network

Организации, публикующий данные

Сеть GBIF

Данные в сети GBIF публикуются от имени организаций. Для этого организация должна быть зарегистрирована и одобрена какой-то уже зарегистрированной организацией.
https://www.gbif.org/become-a-publisher

Публикация данных происходит при помощи “Интегрированного инструмента для публикации данных (IPT)”. Свой IPT устанавливать не обязательно



Число организаций, публикующих данные

Сеть GBIF

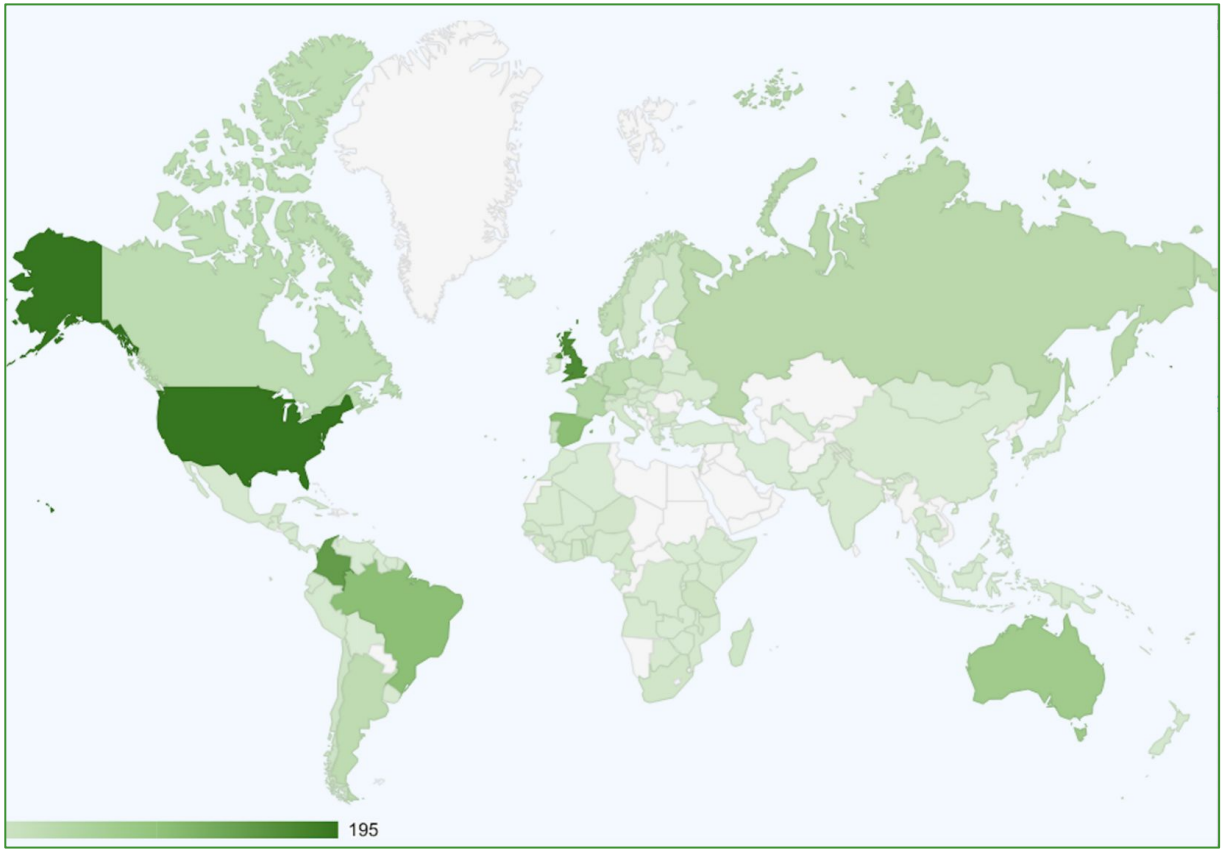
Всего 1444 организации

Россия - 88 организаций
26 августа 2019 года присоединился заповедник Брянский лес

Украина - 6 организаций

Узбекистан - 4 организации

Беларусь - 2 организации



Данные, опубликованные через GBIF

Доступные данные

Данные, опубликованные через сеть GBIF может получить любой желающий

За предыдущие 10 лет объём опубликованных данных увеличился почти на порядок

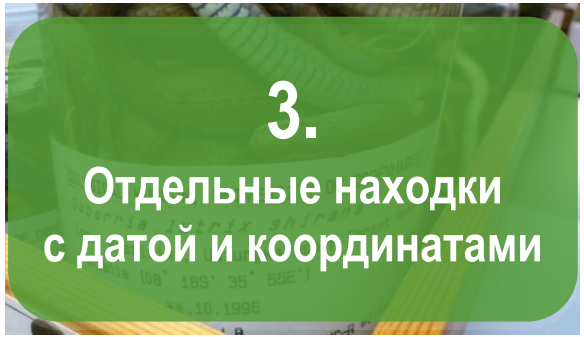
www.gbif.org/analytics/global

Доступные данные

Уровни подробности данных в GBIF

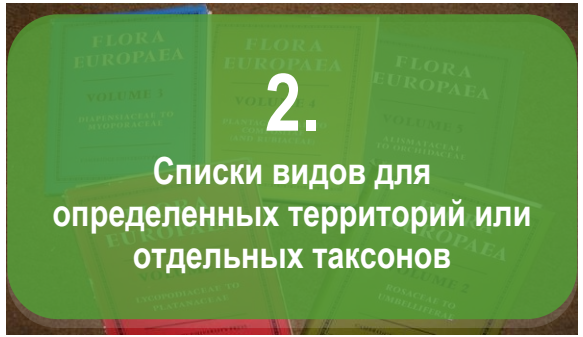
1.
Описательная информация

Dataset metadata



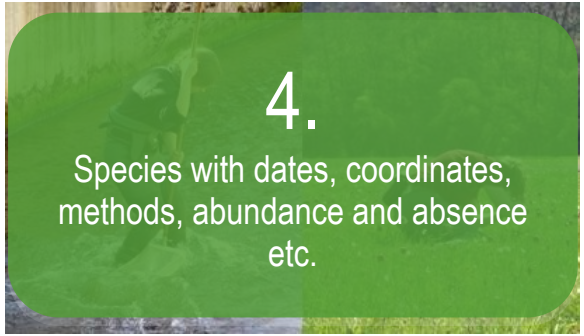
3.
Отдельные находки с датой и координатами

Occurrence data



2.
Списки видов для определенных территорий или отдельных таксонов

Species checklists



4.
Species with dates, coordinates, methods, abundance and absence etc.

Sampling-event data

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

КОЛИЧЕСТВЕННЫЕ ПОКАЗАТЕЛИ

на 27 августа 2019 года

Доступные данные

Всего находок биологических видов

1 351 861 301

наборов данных

45 855

официальных участников

58 *стран*

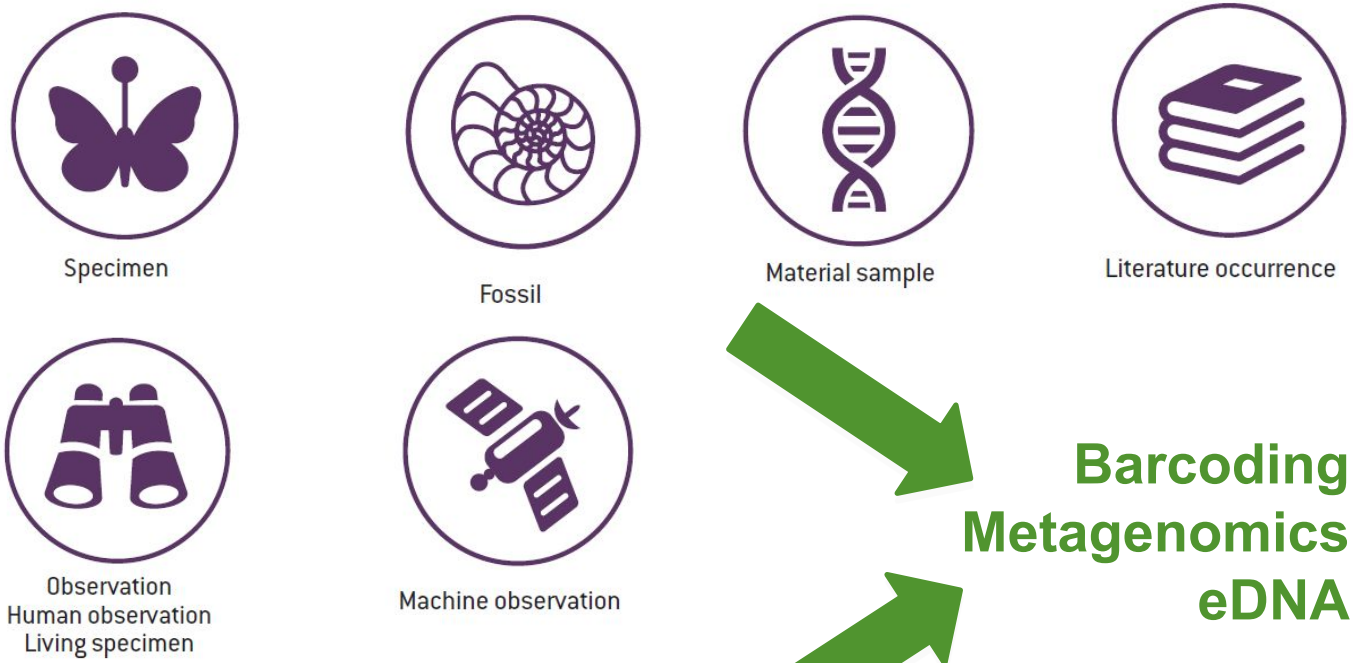
38 *организаций*

организаций

1 444

Типы данных, распространяемых через GBIF

Доступные данные



Barcoding
Metagenomics
eDNA

sample-based data

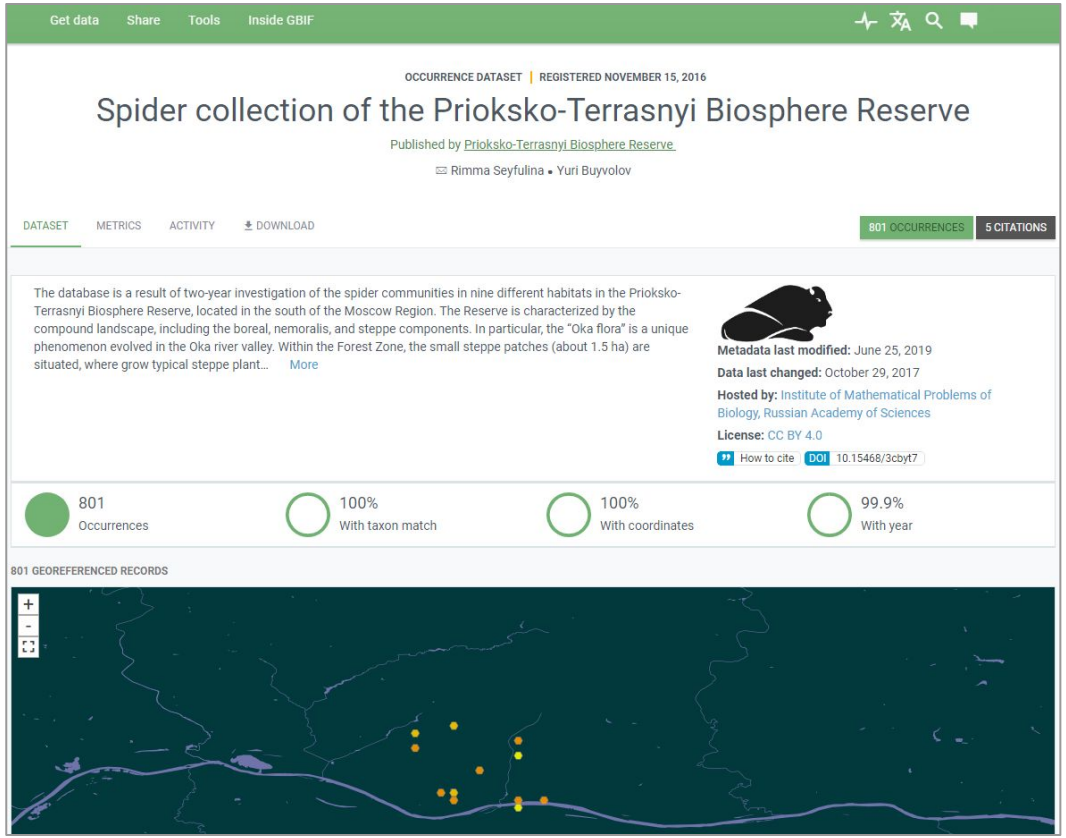
Опубликованный набор данных

Доступные данные

На сегодняшний день (2019-08-28) 88 российскими организациями опубликовано 91 набор данных

1 836 687 находок, охватывающих 181 страну и территорию

для территории России опубликовано 2 904 659 находок



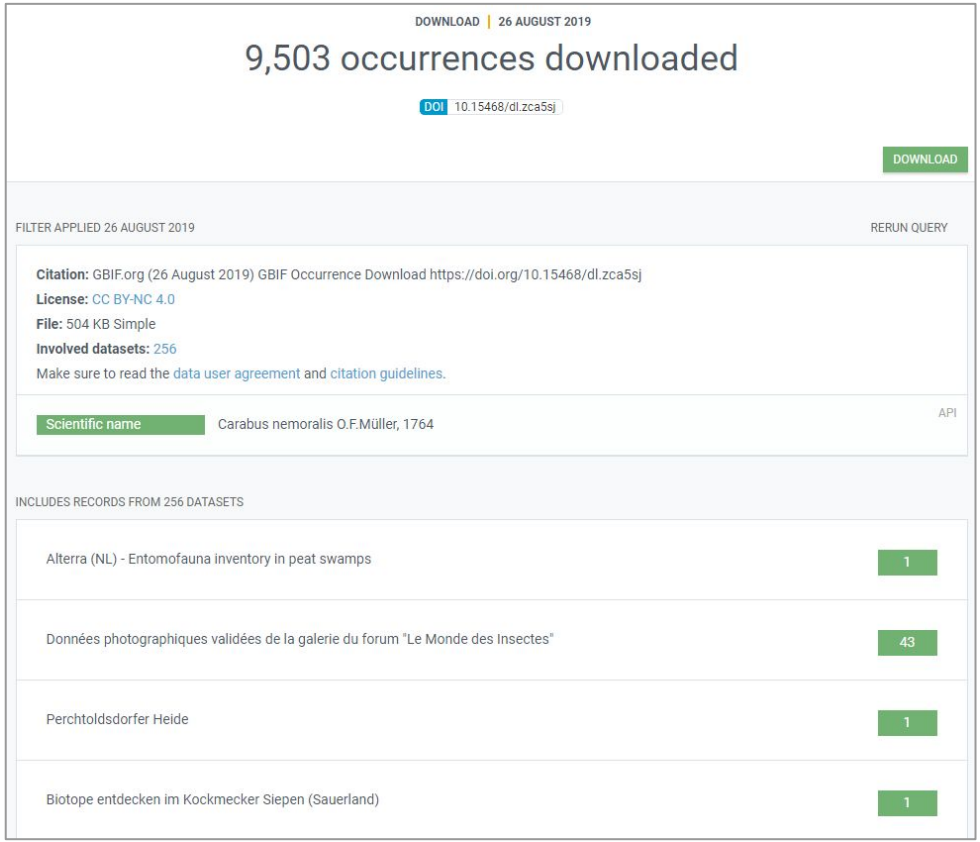
DOI для загружаемых данных

Использование данных

Каждая загрузка данных по пользовательскому запросу получает идентификатор цифрового объекта - DOI

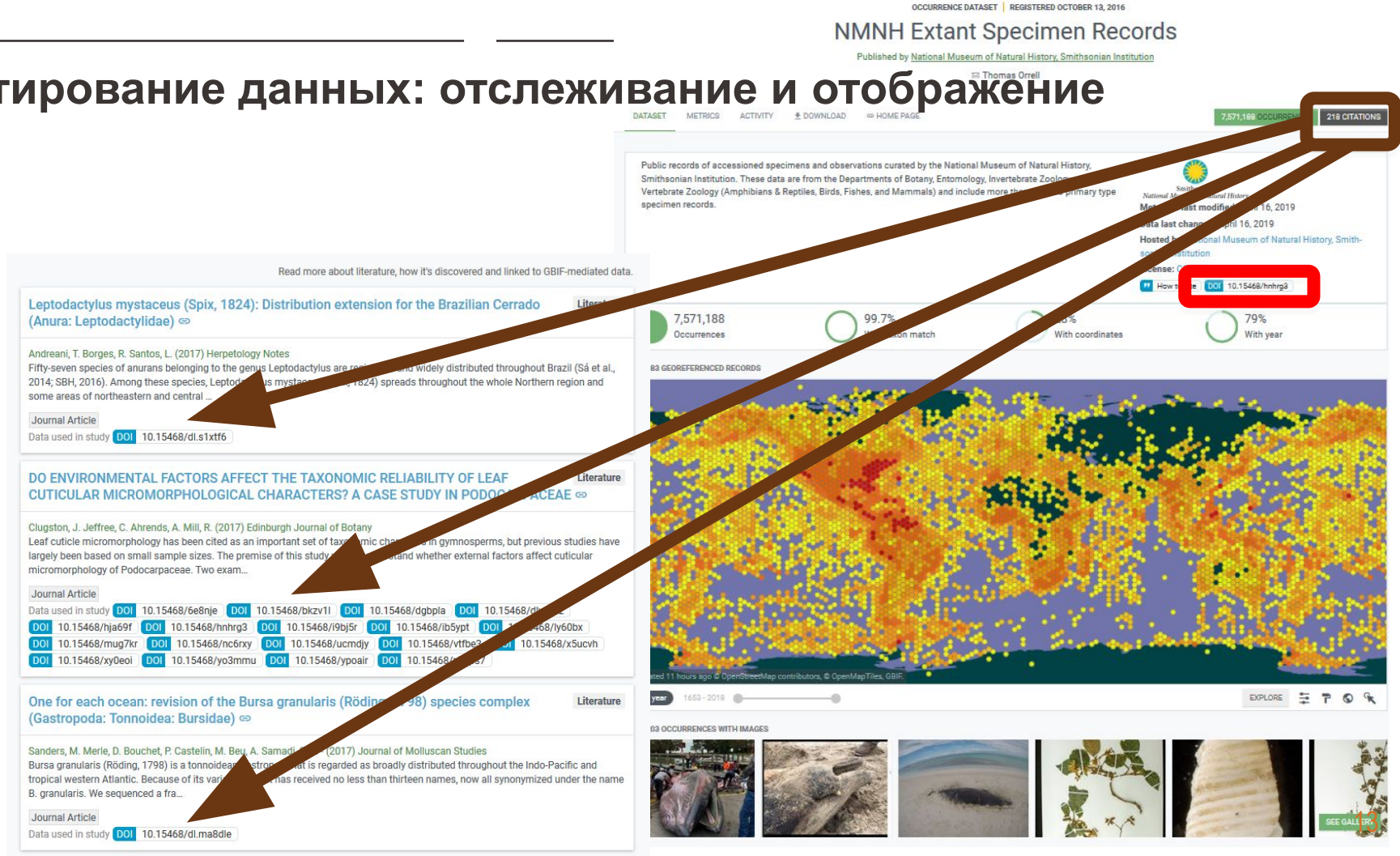
На странице ресурса доступны

- Дата и размер загрузки
- Критерии поиска
- Ссылка для повторной загрузки
- Сведения об исходных наборах данных



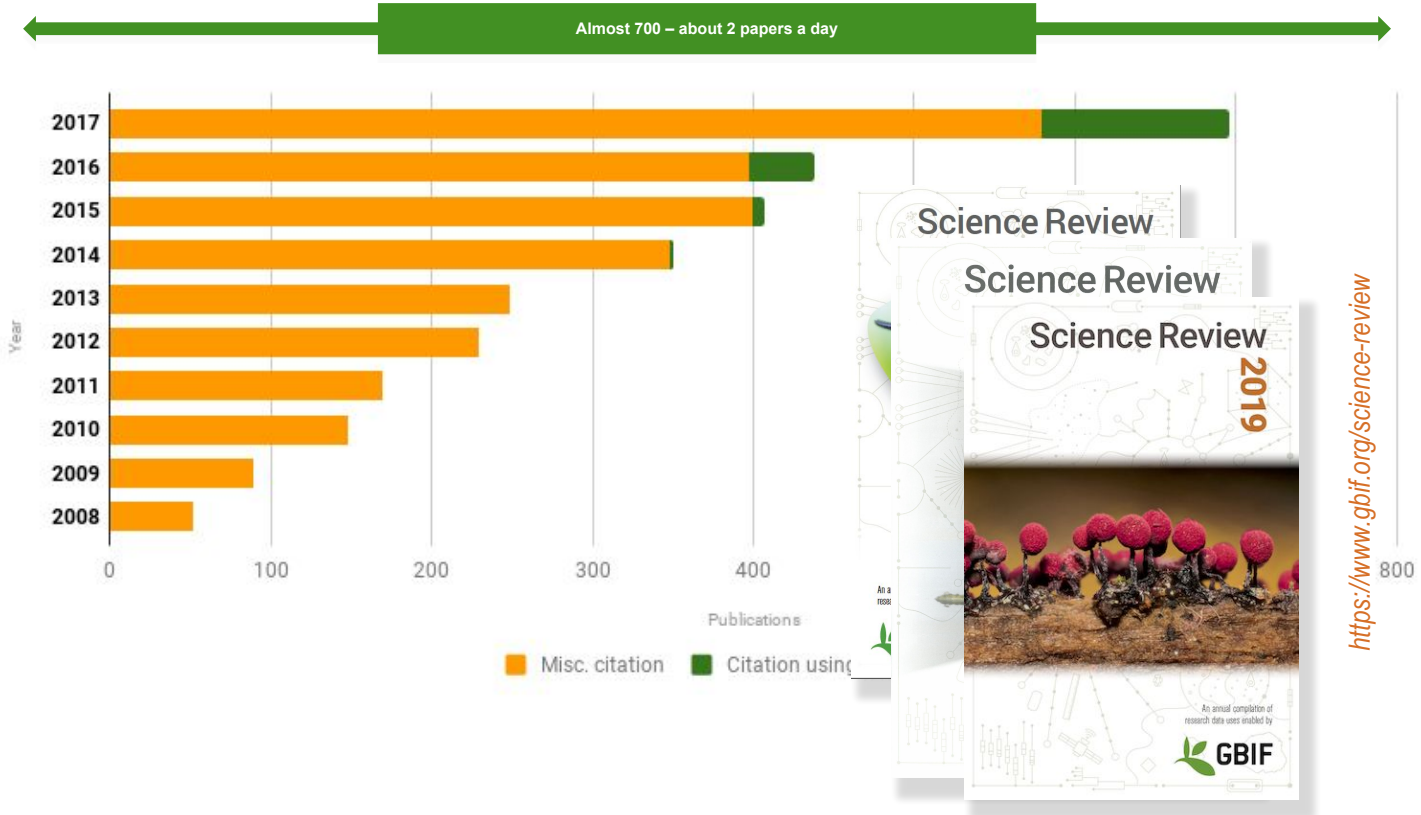
Использование данных

Цитирование данных: отслеживание и отображение



Использование данных

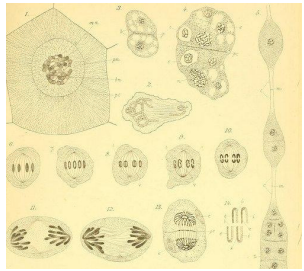
Рецензируемые статьи с использование данных GBIF



https://www.gbif.org/science-review

Разные способы документировать биоразнообразие

Данные о биоразнообразии



Латинские названия: Линнеевская номенклатура

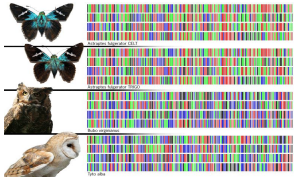
OTU коды

GBIF Global Biodiversity Information Facility

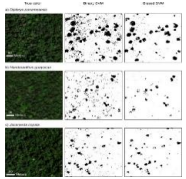
Images: Noah Strycker, Biodiversity Heritage Library, Dmitry Schigel

Пять главных вопросов для данных о биоразнообразии

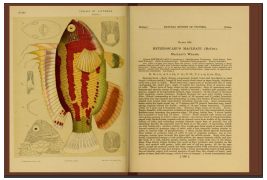
Данные о биоразнообразии



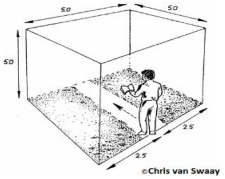
Genomics



Remote-sensing

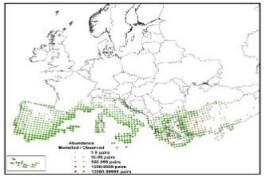


Literature

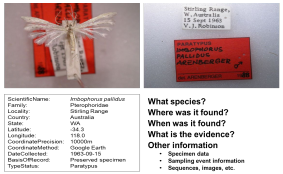


Survey data

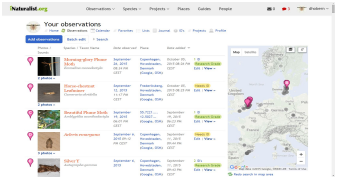
Что это (вид)?
Где?
Дата и время?
Доказательство?
Обилие?



Atlas data



Specimens



Citizen science

Animal tracking

http://e360.yale.edu/features/dna_technology_discovering_new_species - Suz Bateson, University of Guelph

doi: 10.1371/journal.pone.0118403 / http://www.biodiversitylibrary.org/bibliography/4821 / http://www.ebba2.info/ Roche

Зачем открывать данные

Данные о биоразнообразии

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 3133 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 20, 2007; **Published** March 21, 2007

Copyright: © 2007 Piwowar et al. This is an open-access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

Funding: HAP was supported by NLM Training Grant Number ST15-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

Piwowar et al. 2007

ОТКРЫТЫЕ ДАННЫЕ ...

Повышают эффективность исследований

Усиливают “научную строгость” и качество исследований

Позволяют отслеживать использование данных (DOI)

Формируют новый вид научного результата - Data paper

Повышают заметность вашего исследования

Способствуют возникновению новых вопросов

Способствуют сотрудничеству и развитию сообщества

Усиливают экономическое и социальное значение исследований

Оказывают влияние на международные конвенции

Совместное использование данных

Данные о биоразнообразии

Краткосрочное — Долгосрочное

перспектива

ФАЗА ИССЛЕДОВАНИЯ

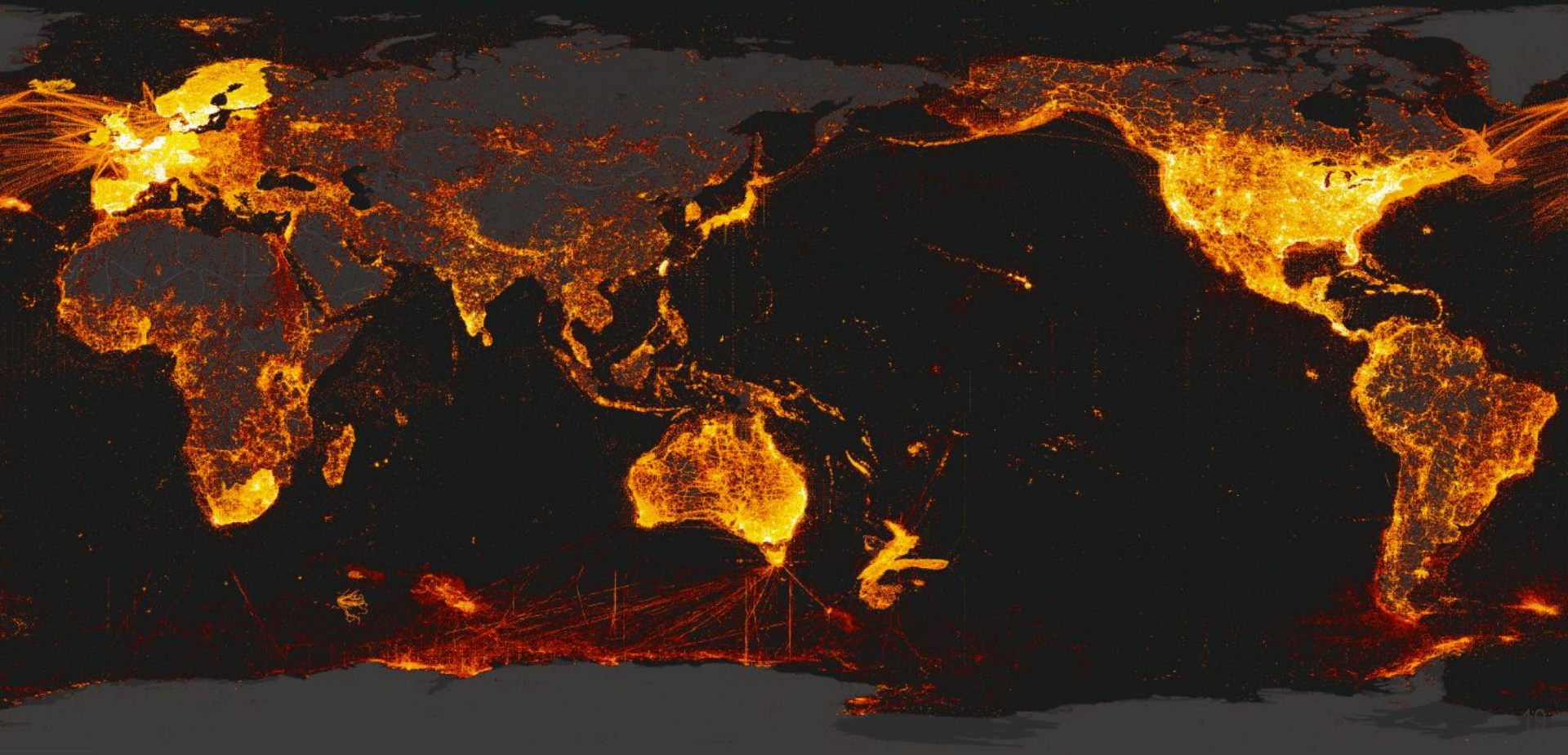
- формат файла
- права на данные
- метаданные
- где хранить
- резервные копии

ФАЗА РАСПРОСТРАНЕНИЯ

- делиться ли?
- с кем делиться?
- лицензия
- метаданные

ФАЗА ХРАНЕНИЯ

- хранилище?
- долгосрочное управление ?

Sophie Kay 2013 (Open Science Training Initiative) CC-BY 3.0.

Спасибо за внимание !!!