

3 апреля 2023

DATA FROM KAZAKHSTAN

94,781

Published occurrences

4

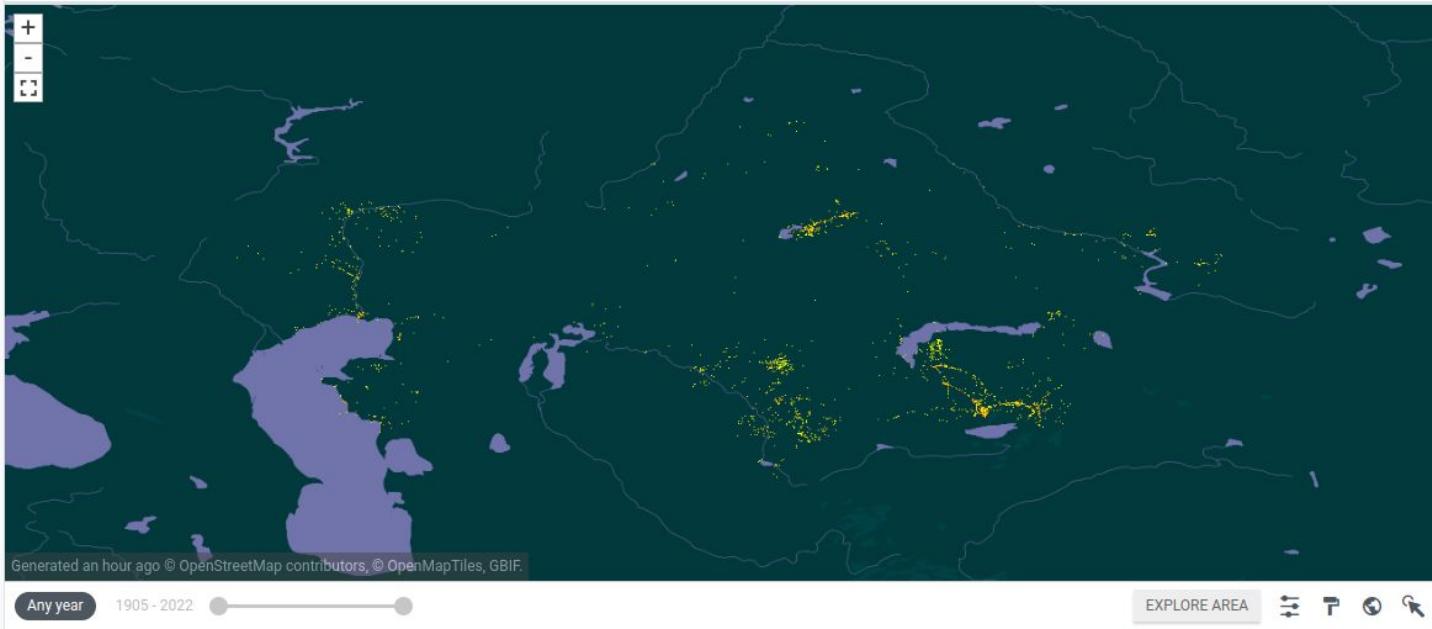
Published datasets

1

Countries and areas covered by  
data from Kazakhstan

7

Publishers from Kazakhstan



7 апреля 2023

DATA FROM KAZAKHSTAN

94,781

Published occurrences

12

Published datasets

1

Countries and areas covered by  
data from Kazakhstan

7

Publishers from Kazakhstan





Методы оцифровки данных по флоре и фауне  
и размещение на международной платформе биологического  
разнообразия Карагандинский университет  
имени академика Е.А. Букетова,  
3-15 апреля 2023 г.



# Оцифровка научных биологических коллекций

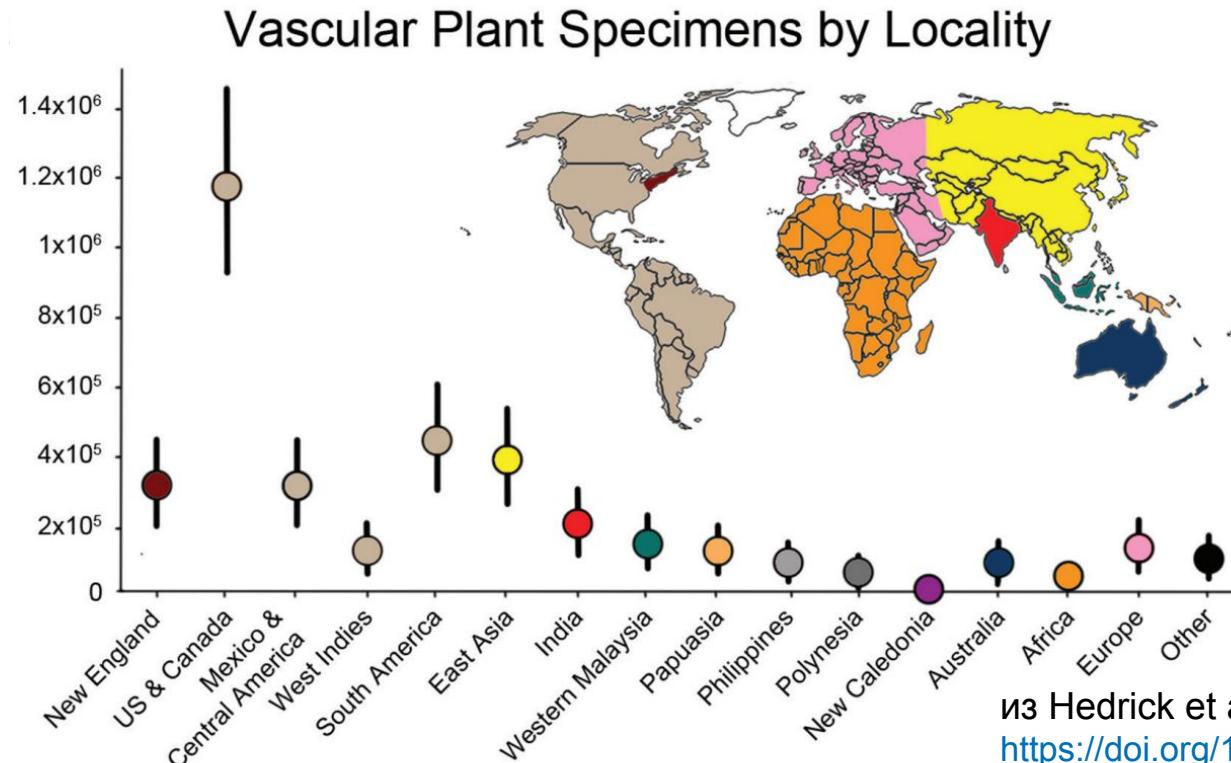
Наталья Иванова

# Сколько образцов хранится в научных биологических коллекциях мира?

В 2018 году в мире функционировало **3 095** гербариев, общее число образцов - **387 000 000** (Thiers, 2018). Эта оценка получена на основе [Index Herbariorum](#), т.е. включает только коллекции, у которых есть акроним.



# Сколько образцов хранится в научных биологических коллекциях мира?



# Сколько образцов хранится в научных биологических коллекциях мира?

По оценке Arturo H. Ariño (2010), в научных биологических коллекциях по всему миру хранится около **двух миллиардов экземпляров**  
(<https://journals.ku.edu/jbi/article/view/3991>)

По оценке Groom et al. (2019) общее число образцов составляет **1.2 миллиарда** экземпляров  
(<https://academic.oup.com/database/article/doi/10.1093/database/baz129/5670756>)

# Для чего оцифровывать коллекции



Национальный музей Бразилии, Рио-де-Жанейро

Учреждён королём Португалии Жуаном VI в 1818 году

20 миллионов образцов

# Для чего оцифровывать коллекции



Огнем уничтожено 92.5% образцов

[https://en.wikipedia.org/wiki/National\\_Museum\\_of\\_Brazil\\_fire](https://en.wikipedia.org/wiki/National_Museum_of_Brazil_fire)

Пожар в Национальном музее Бразилии  
2 сентября 2018



# Для чего оцифровывать коллекции

Небольшая часть утраченных коллекций была оцифрована и доступна на портале GBIF

**Coleção Entomológica do Museu Nacional / UFRJ**

Occurrence dataset

"Museu Nacional/UFRJ é vinculado ao Ministério da Educação. É a mais antiga instituição científica do Brasil e o maior museu de história natural e antropológica da América Latina. Criado por D...

Published by Museu Nacional / UFRJ

117,269 occurrences | 78 citations

**R - Herbario do Museu Nacional**

Occurrence dataset

O acervo está sendo organizado em armários compactados, em ordem alfabética de famílias Consta do Index Herbariorum com a sigla R, e estima-se que a coleção possua cerca de 550.000 exemplares, sendo 9...

Published by Museu Nacional / UFRJ

59,183 occurrences | 469 citations

**Mollusca Collection - Museu Nacional/UFRJ**

Occurrence dataset

The Mollusca Collection of the Museu Nacional (acronym MNRJ), one of the most important of its kind in South America, holds more than 40,000 registered lots and about 15,000 unregistered lots, which a...

Published by Museu Nacional / UFRJ

25,065 occurrences | 82 citations

**Coleção de Aves do Museu Nacional / UFRJ**

Occurrence dataset

Criado por D. João VI em 6 de junho de 1818, o Museu Nacional constitui um dos maiores e mais tradicionais centros de pesquisa da América Latina, sendo detentor de um dos mais vastos e representativos...

Published by Museu Nacional / UFRJ

32,060 occurrences | 145 citations

**Coleção Ictiológica (MNRJ), Museu Nacional (MN), Universidade Federal do Rio de Janeiro(UFRJ)**

Occurrence dataset

A Coleção Ictiológica (MNRJ) do Museu Nacional, Universidade Federal do Rio de Janeiro, é uma das maiores e mais antigas coleções científicas de peixes do Brasil, contando com um acervo iniciado no fi...

Published by Museu Nacional / UFRJ

Occurrence dataset

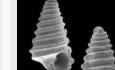
Tools | Community | About

Search | Login

View | XA |  |  |  | 



## Для чего оцифровывать коллекции

Оцифровка повышает доступность информации

*Lepidium karataviense* Regel & Schmalh. | Клоповник каратавский

В музее Естественной истории (Лондон) хранится 1 образец, собранный в Казахстане

Перелет Астана - Лондон ~ 490 000 ₸  
Расстояние 5 900 км

Отсканированный образец и информация этикетки бесплатно доступны [на сайте музея](#) и на портале [GBIF](#)



<https://www.gbif.org/occurrence/1799005094>

## Оцифровка - перенос данных с бумажного носителя на электронный



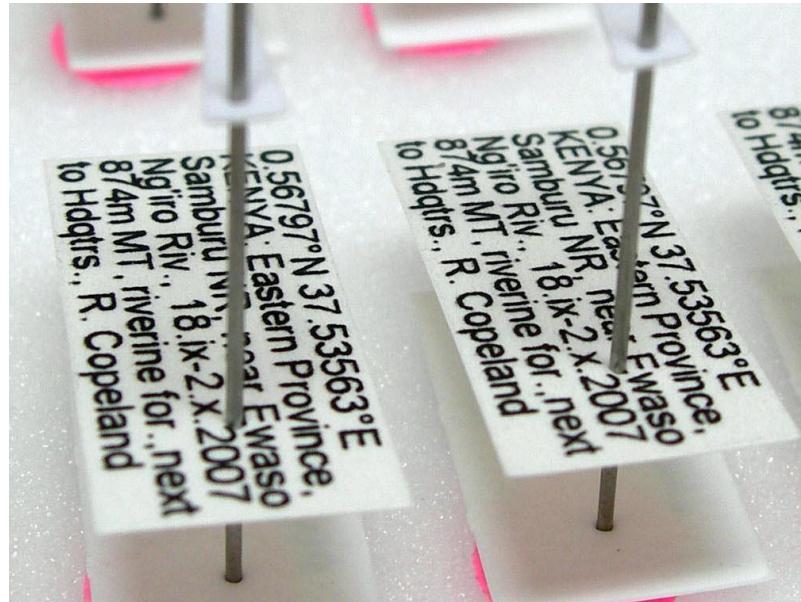
# Оцифровка научных биологических коллекций

Физический образец  
(само насекомое)



Этикетка

# Что важнее - физический образец или этикетка?



ГЕРБАРИЙ КОМИ НАУЧНОГО ЦЕНТРА  
УРАЛЬСКОГО ОТДЕЛЕНИЯ РАН (SYKO)

*Dieramum drumondii* Müll. Hal.

Республика Коми: Княжпогостский р-н,  
р. Бычава, приокр р. Киселовка,  
6 км до п. Киселово. У водок.

(62°20' - 50°50')

16.39

21.07.2010

Leg.: Тегерюк Б.Ю.

№ 12

№ общий 52388

Det.: Жемэлова Г.В.

Минимум данных, желательный для  
регистрации находки в GBIF

Оцифрованные коллекции не  
всегда предоставляют доступ к  
изображениям образцов, но  
обязательно - к этикеточным  
данным

# Какие задачи необходимо решить в процессе оцифровки коллекций



Образец из коллекции  
Ohio State University

- Подготовка образцов к оцифровке (проверка, баркоды)
- Получение изображений
- Обработка изображений
- Сбор электронных данных (информация этикеток)
- Географическая привязка
- Связь изображений, баз данных коллекций и научных публикаций с физическими образцами и коллекторами

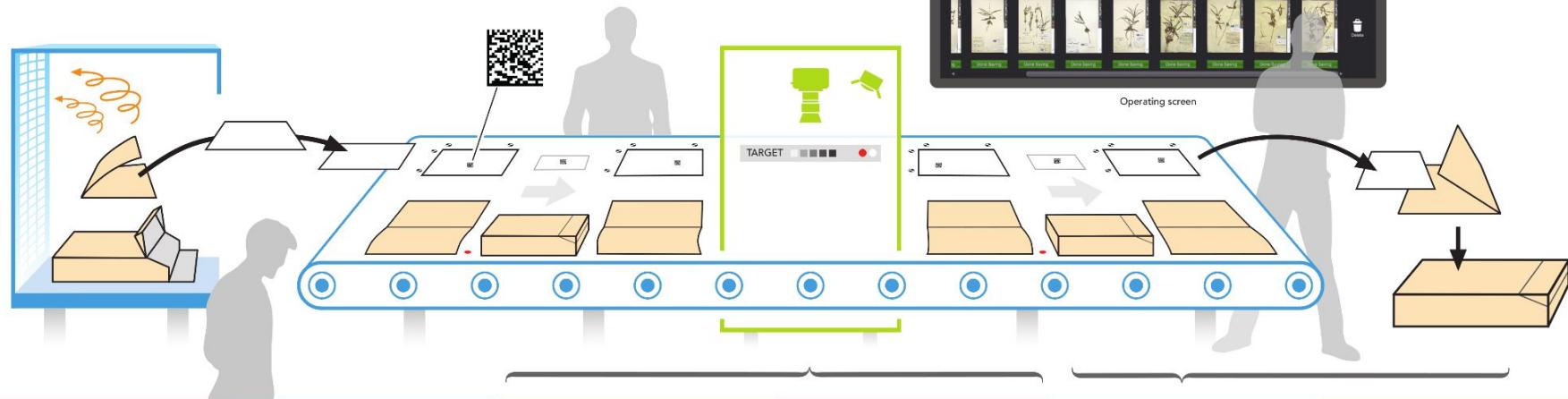
Nelson et al., 2012  
Hudson et al., 2015

<https://doi.org/10.1371/journal.pone.0143402>  
<https://zookeys.pensoft.net/article/2926/list/9/>



PICTURAE

# Этапы оцифровки



## 1 Preparation

- Selection of boxes
- Extract vapor
- Apply box barcode
- Apply cover barcode

## 2 Place material on conveyor belt

- Spread on the conveyor belt
- Multisheet token
- Straighten
- Apply sheet barcode

## 3 Digitizing

- Read barcode
- Multisheet yes/no
- Apply ICC profile
- Readout color
- Readout sharpness
- Feedback → retake

## 4 Processing

- Rotating
- Cropping
- Readout target
- Multisheet color code
- Merge metadata to CSV file format
- Save deliverables

## 5 Packing

- Order remains intact
- Logistic management
- Return material

## 6 Metadata entry

- Cover & label description
- Look-up lists
- Linking to databases
- Multisheet processing

# Обязательные элементы на оцифрованном гербарном листе



- (1) Цветовая палитра | Colour Chart
- (2) Масштабная линейка | Scale Bar
- (3) Баркод (штрихкод) | Barcode
- (4) Этикетка | Labels
- (5) Название института | Institution Name

Nieva de la Hidalga et al., 2020

<https://bdj.pensoft.net/article/47051/list/9/>

# Баркод (штрихкод) | Barcode



Идентификатор  
(уникальный номер, ID)  
образца в коллекции

# Оцифровка гербария



Установка для фотографирования  
образцов в гербарии Arkansas  
State University (STAR)  
Harris & Marsico, 2017  
[doi: 10.3732/apps.1600125](https://doi.org/10.3732/apps.1600125)

# Оцифровка гербария



Установка для конвейерной  
оцифровки образцов

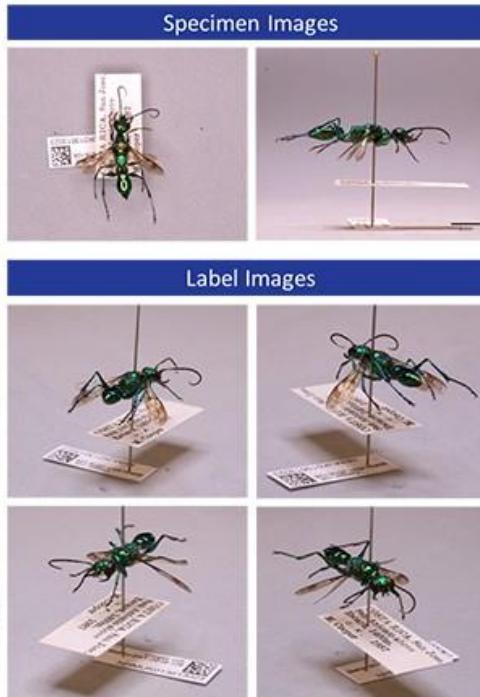
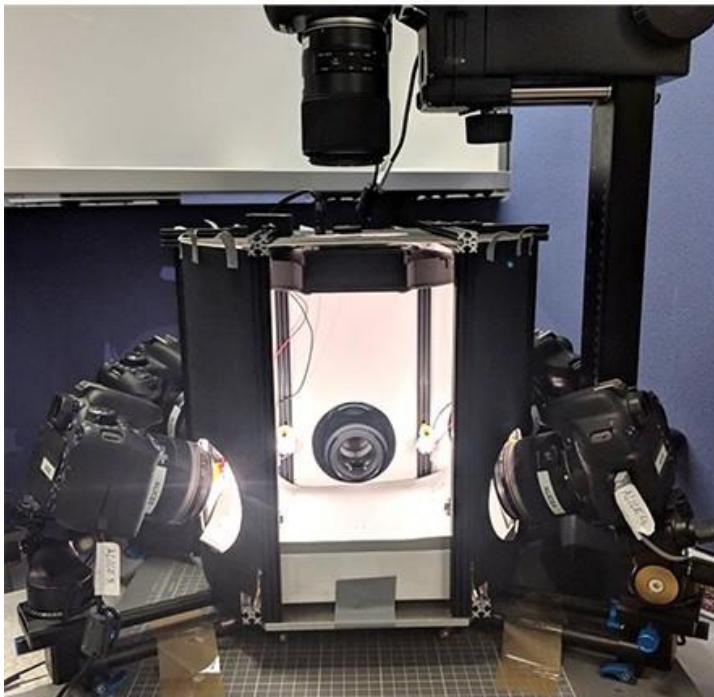
# Оцифровка энтомологических коллекций: объемные объекты



# Оцифровка энтомологических коллекций: объемные объекты



# Оцифровка энтомологических коллекций: объемные объекты



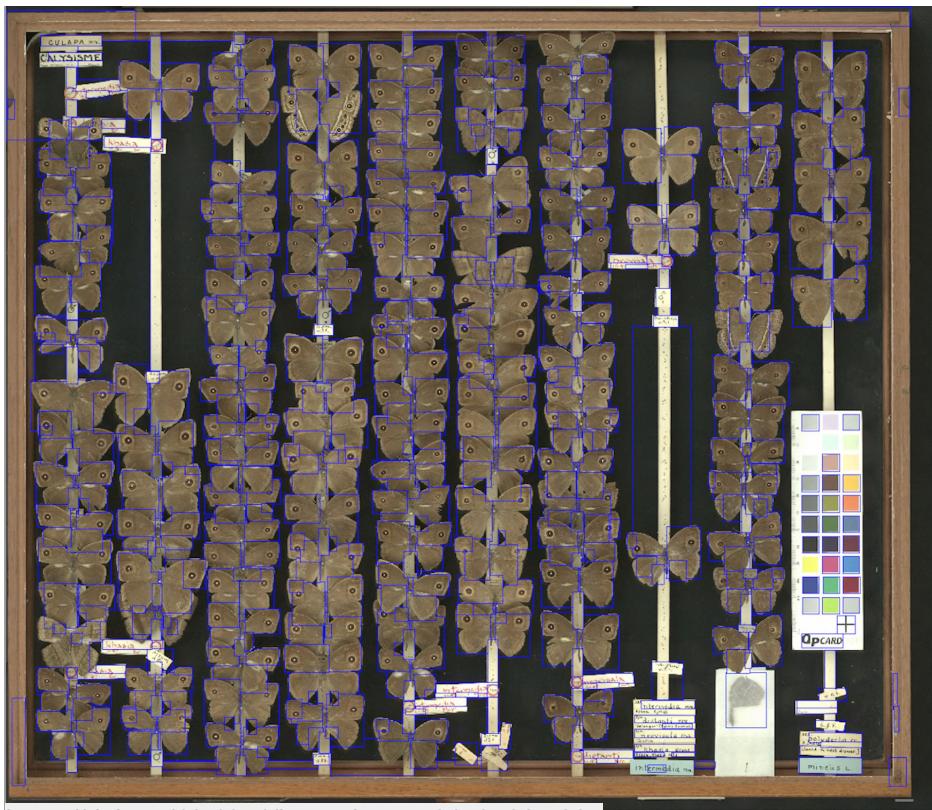
Музей естественной истории,  
Лондон

Метод оцифровки без снятия  
этикеток 'Angled Label Image  
Capture and Extraction' (ALICE)

6 фотографий с разных ракурсов  
позволяют получить изображение  
образца и реконструировать  
этикетку

Price et al., 2018  
<https://osf.io/s2p73/>

## Оцифровка энтомологических коллекций: объемные объекты



## Переход от энтомологической коробки к отдельным образцам

Inselect - открытое программное обеспечение для сегментации отдельных образцов на материалах массовой оцифровки

Hudson et al., 2015

<https://doi.org/10.1371/journal.pone.0143402>

<https://doi.org/10.1371/journal.pone.0143402.g007>

# Оцифровка орнитологических коллекций





## Где хранить изображения?

Файл с изображением этого образца занимает 521.5 kB

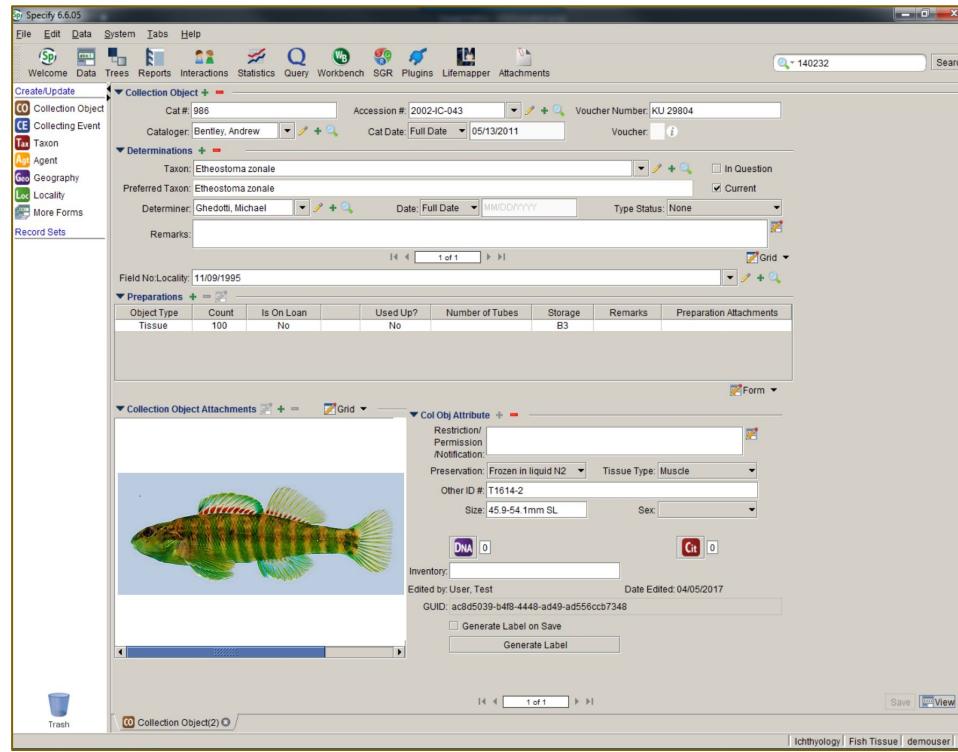
В настоящий момент в Цифровом гербарии ЦСБС СО РАН\* 21 655 оцифрованных гербарных листов

В целом в гербарии высших растений, лишайников и грибов (NS, NSK)" хранится 808 550 листов и пакетов

\*Центральный сибирский ботанический сад, Новосибирск, Россия

<http://www.csbg.nsc.ru/gerbarij/tematika-14.html>

# Готовые программные продукты для внесения данных и ведения цифровых коллекций





## Коллекции, доступные на портале “Ноев ковчег”

Раздел	Число коллекций	Число образцов
Животные	13	191 704
Растения	14	1 215 788
Микроорганизмы и грибы	24	25 629

А также Биоматериалы человека, Культурные растения, Дрозофилы, Химическая энзимология



ДЕПОЗИТАРИЙ  
живых систем  
«НОЕВ КОВЧЕГ»

Микроорганизмы и грибы

Растения

Животные

Биоматериалы человека

Био. информация

...

RU

EN



Вход в систему

[О системе](#) [Коллекции](#) [Контакты](#) [Ссылки](#) [Инфраструктура](#) [Цитировать](#)

Сейчас в базе данных (гербарий, образцы ДНК, фотографии растений в природе):



Образцов: [1215788](#)



Изображений: [1194601](#)



Видов: [39460](#)



Геопривязок: [804582](#)



Этикеток + OCR: [556798](#) + [656143](#)

## Национальный банк-депозитарий живых систем

### Цифровой гербарий МГУ

Проект Московского университета "Ноев ковчег" посвящен созданию многофункционального сетевого хранилища биологического материала.

Планируется работа с материалом всех возможных типов - от отдельных биологических молекул до целых живых организмов.

Создание депозитария позволит сохранить биоразнообразие нашей планеты и создать новые способы полезного использования биологического материала.

Сегодня Цифровой гербарий МГУ - это консорциум нескольких российских университетских и академических гербариев, которые вносят [свой вклад](#) в документацию флоры России. Эти данные доступны также в GBIF.

### Атлас флоры России

В разделе «Атлас флоры России» в паспорте каждого образца дана предварительная сеточная карта по квадратам 100×100 км на основе датасета FLORUS. Этот массив данных включает в себя предварительно очищенные данные GBIF (в том числе сведения из Цифрового гербария МГУ) и ряда других источников. Общий объем исходных данных по флоре России – около 6,5 млн точек (700 тыс. указаний для отдельных квадратов). Мы занимаемся активной чисткой карт и проверкой исходных данных.

Хочешь принять участие в создании "Атласа флоры России"? Загружай свои фотографии растений в природе и точку съемки на [iNaturalist](#), где они станут частью нашего нового проекта "Флора России | Flora of Russia".



Алексей Серегин, д.б.н.,  
Куратор гербария MW  
(МГУ, Москва)

<https://plant.depo.msu.ru/>

# Вклад участников консорциума Цифрового гербариев МГУ

## MW (Гербарий Московского университета)

 Образцов: 1036251  Изображений: 1015242  Видов: 39083  Геопривязок: 667843  Этикеток + OCR: 436278 + 597127

## MHA (Гербарий Главного ботанического сада РАН, г. Москва), с 1.04.2019

 Образцов: 97702  Изображений: 97517  Видов: 3364  Геопривязок: 75864  Этикеток + OCR: 50053 + 47648

## IRKU (Гербарий Иркутского государственного университета), с 11.09.2020

 Образцов: 42254  Изображений: 42154  Видов: 1112  Геопривязок: 27889  Этикеток + OCR: 30901 + 11353

## KUZ (Гербарий Кузбасского ботанического сада СО РАН, г. Кемерово), с 14.05.2020

 Образцов: 19014  Изображений: 19104  Видов: 1445  Геопривязок: 19004  Этикеток + OCR: 19009 + 5

## TUL (Гербарий Тульского государственного педагогического университета имени Л.Н. Толстого), с 25.12.2019

 Образцов: 9000  Изображений: 9024  Видов: 1164  Геопривязок: 8956  Этикеток + OCR: 9000 + 0

## TULGU (Гербарий Тульского государственного университета), с 07.2021

 Образцов: 3921  Изображений: 3921  Видов: 678  Геопривязок: 794  Этикеток + OCR: 3921 + 0

## KULPOL (Гербарий Музея-заповедника «Куликово поле», г. Тула), с 07.2021

 Образцов: 3330  Изображений: 3329  Видов: 576  Геопривязок: 850  Этикеток + OCR: 3330 + 0

## MAG (Гербарий Института биологических проблем Севера ДВО РАН, г. Магадан), с 22.10.2020

 Образцов: 2604  Изображений: 2604  Видов: 106  Геопривязок: 2537  Этикеток + OCR: 2598 + 6

## TKM (Гербарий Тульского областного краеведческого музея), с 07.2021

 Образцов: 1712  Изображений: 1706  Видов: 750  Геопривязок: 845  Этикеток + OCR: 1708 + 4

how many digitized specimens are there

[All](#) [Images](#) [News](#) [Books](#) [More](#)

Tools

About 389,000,000 results (0.49 seconds)

<https://phys.org> › Biology › Plants & Animals

### Southeastern US herbaria digitize three million specimens ...

Jul 23, 2021 — Many collections include **specimens** that are now extinct in the wild, while others have yielded the discovery of entirely new **species**. And as ...

<https://beta.nsf.gov> › news › southeastern-us-herbaria-d...

### Southeastern U.S. herbaria digitize 3 million specimens, now ...

Southeastern U.S. herbaria digitize 3 million **specimens**, now freely available online. Effort involved more than 100 herbaria spread out across the ...

<https://herbarium.natsci.msu.edu> › Collections

### Digitization - MSU Herbarium

The MSU Herbarium is in the process **digitizing specimens**; that is imaging them, transcribing **their labels**, and then making these data **available** online.

### People also ask



What is herbarium digitization?



How do you make a digital herbarium?



Who initiated the art of herbarium?

[Feedback](#)<https://www.researchgate.net> › figure › Number-of-scanne...

### Number of scanned herbarium specimens in the world's ...

... at least 61 herbaria have over 1M physical specimens. World leaders in herbarium digitisation are P, L, NY, PE and US (Table 1). This list is a compilation ...

9 из 10 первых результатов - источник по оцифровке гербарных коллекций

# Сколько образцов из научных биологических коллекций оцифровано?

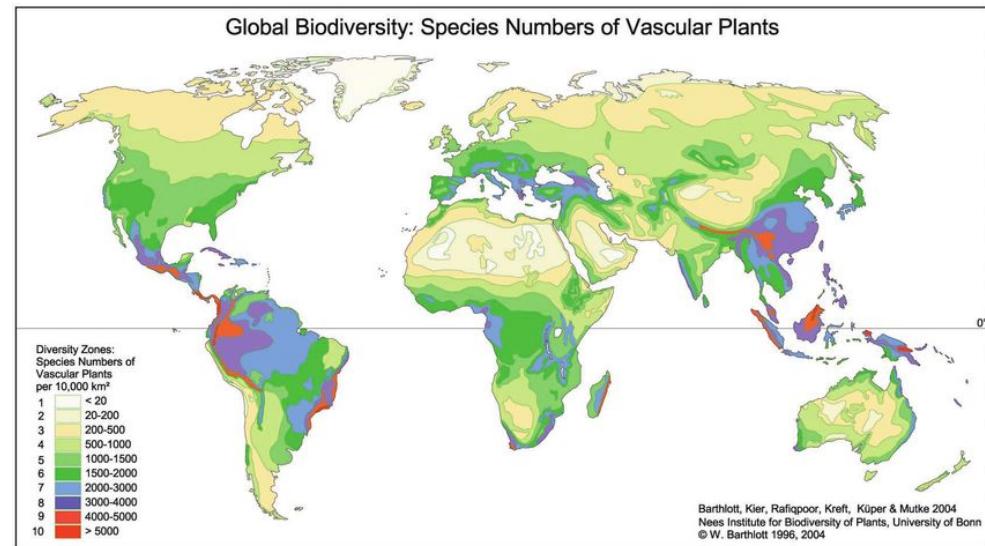
Коллекция	Всего образцов	Оцифровано
Музей естественной истории, Лондон	80 000 000	13 000 000
Зоологический институт РАН, Санкт-Петербург	60 000 000	~28 000
Naturalis Biodiversity Center, Лейден, Нидерланды	37 000 000	7 000 000
Музей природы Канады, Оттава	14 600 000	3 000 000

Borsch T et al. (2020) **A complete digitization of German herbaria is possible, sensible and should be started now.** Research Ideas and Outcomes 6: e50675. <https://doi.org/10.3897/rio.6.e50675>

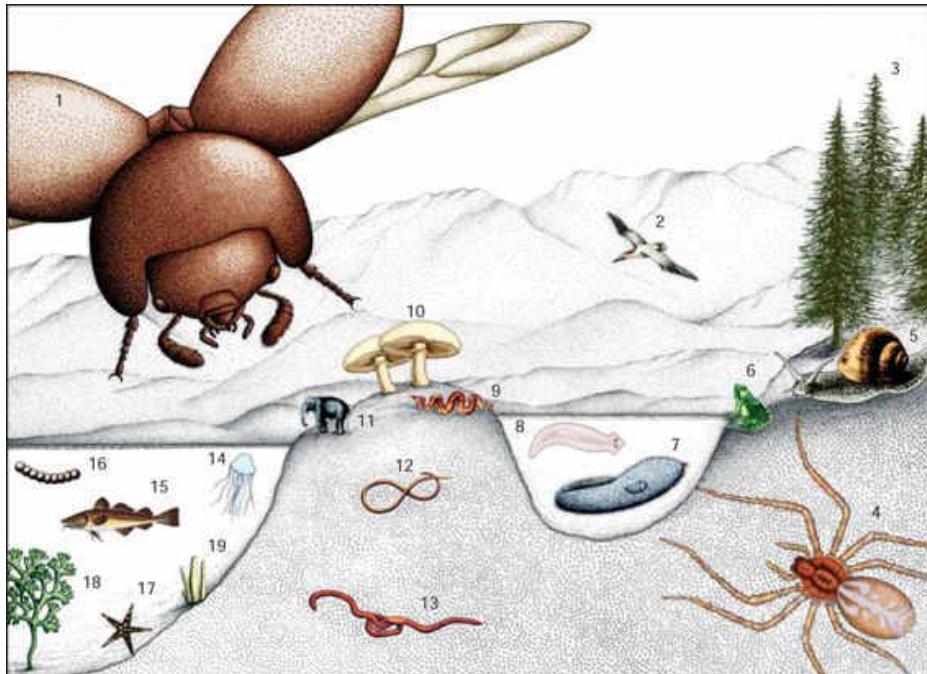
# Оцифрованные данные и разнообразие видов в природе



<https://insider.si.edu/wp-content/uploads/2017/12/herbarium.jpg>



# Оцифрованные данные и разнообразие видов в природе

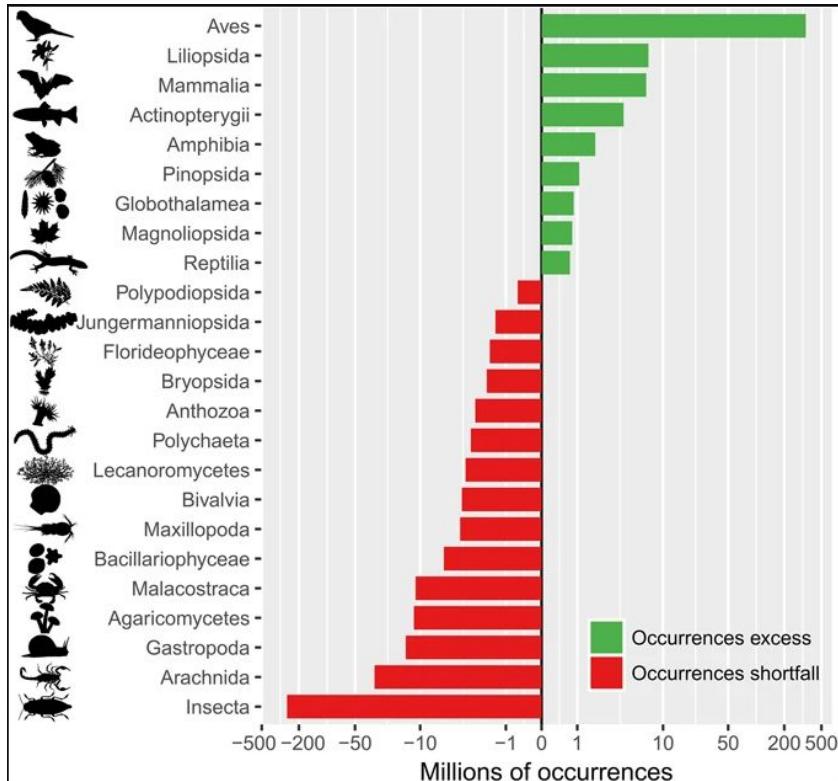


Troudet et al. (2017) проанализировали данные крупнейшего мирового репозитория о биоразнообразии GBIF и выяснили, что разнообразие видов внутри классов не соответствует существующему в природе.

Troudet et al., 2017

<https://doi.org/10.1038/s41598-017-09084-6>

# Оцифрованные данные и разнообразие видов в природе



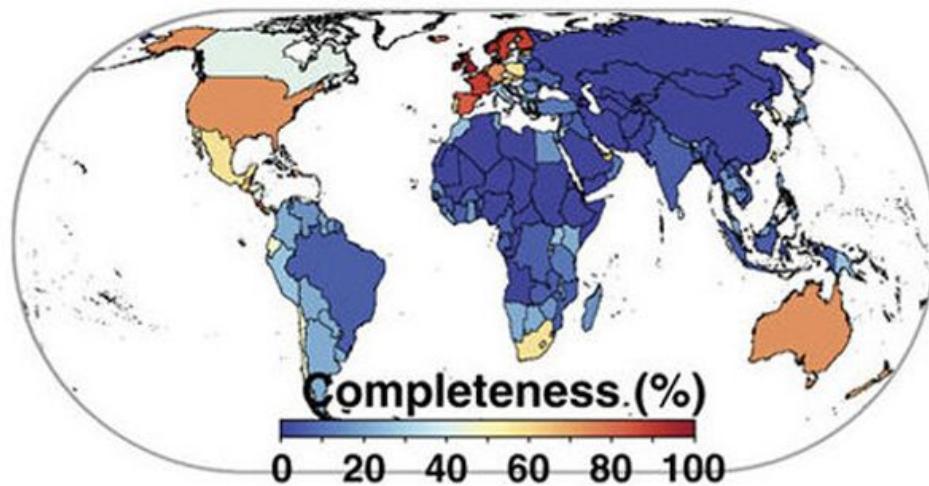
- Наиболее представленный класс - птицы, наименее - насекомые.
- Обнаружено увеличение таксономической ошибки с течением времени, в основном за счет быстрого накопления данных о встречах птиц.
- В последнее время данные по большинству классов накапливались быстрее, чем когда-либо прежде.

Troudet et al., 2017

<https://doi.org/10.1038/s41598-017-09084-6>

## Не только сдвиги (bias), но и пробелы (gaps) в данных

Полнота цифровых данных о биоразнообразии в разных странах согласно порталу GBIF  
(Meyer et al., 2015)



# В больших странах данные распределены неравномерно

Цифровые данные о биоразнообразии, доступные через портал GBIF

Плотность находок сосудистых растений



Плотность находок птиц



## Заключение

- Оцифровка повышает сохранность и доступность данных о биоразнообразии. Многие мировые коллекции сегодня доступны онлайн.
- Оцифровка гербариев идет быстрее, чем оцифровка зоологических коллекций.
- Разработаны методы и технологии, позволяющие автоматизировать многие этапы оцифровки, но в целом процесс идет медленно и большинство образцов в мире остаются неоцифрованными
- В цифровых данных о биоразнообразии есть таксономические сдвиги, соотношение видов не соответствует существующему в природе.
- Также существуют пробелы в данных как для территорий, так и для разных таксонов.