

«Подготовка данных для публикации в Глобальной информационной системе о биоразнообразии GBIF»  
10 октября 2020 г., Екатеринбург

## Лекция 2

# Качество данных. Базовые инструменты для поиска ошибок в данных

Наталья Иванова

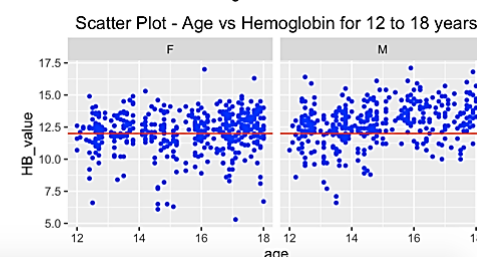
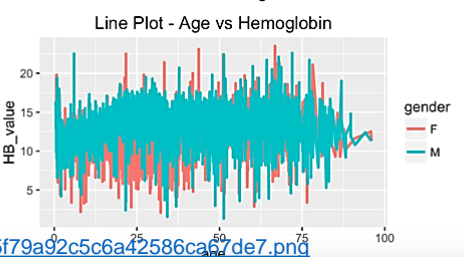
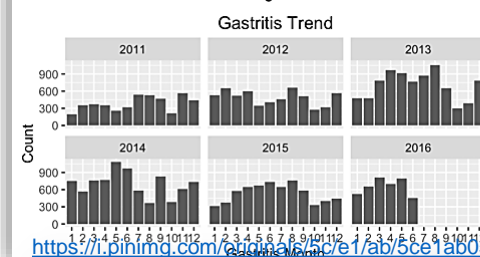
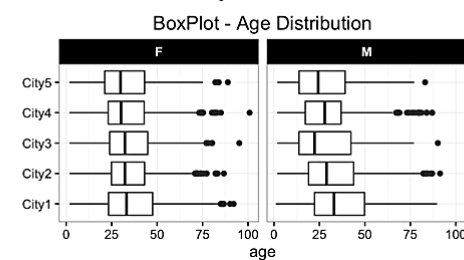
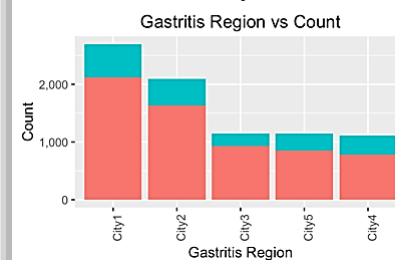
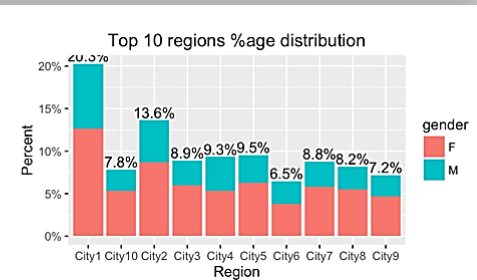
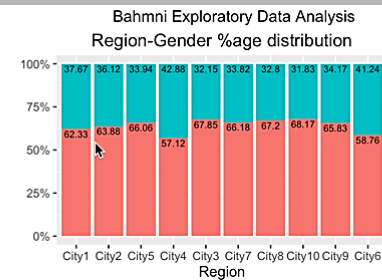
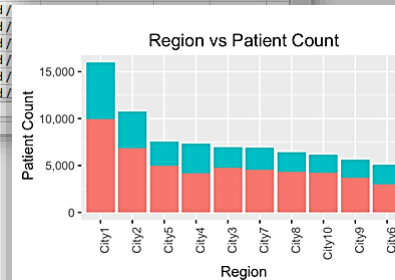
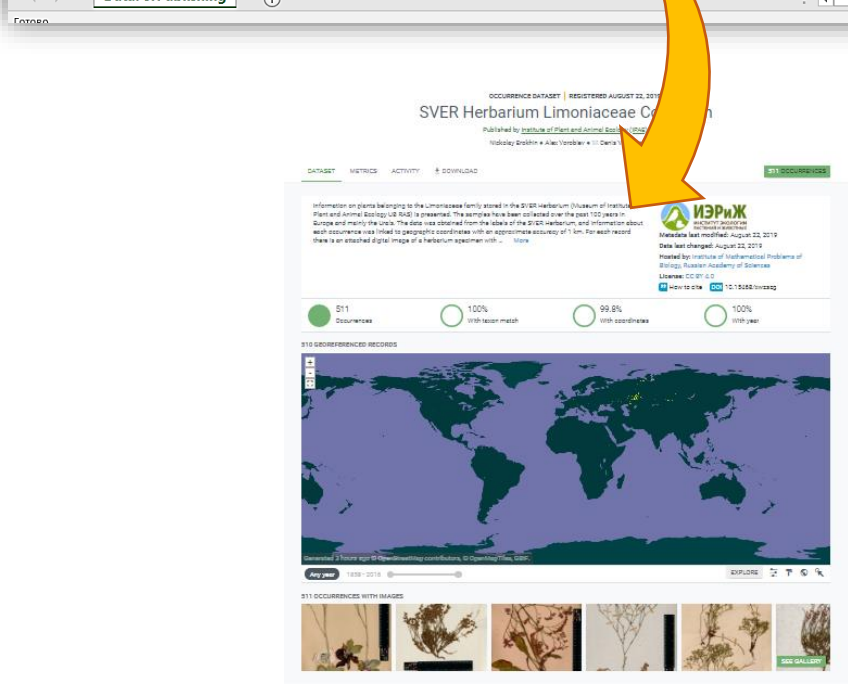
Институт математических проблем биологии РАН – филиал ИПМ им. М.В. Келдыша РАН



**Слайды CC BY:**

*Nicolas Noé, Sophie Pamerlon,  
Sharon Grant  
и Наталья Иванова*

| DataCleaningExample - Excel (Сбой активации продукта)   |             |   |         |      |             |           |            |         |         |         |    |    |            |              |
|---|-------------|---|---------|------|-------------|-----------|------------|---------|---------|---------|----|----|------------|--------------|
| Файл Главная Вставка Разметка страницы Формулы Данные Рецензирование Вид Настройки Что вы хотите сделать? Вход Общий доступ   |             |   |         |      |             |           |            |         |         |         |    |    |            |              |
| <div> <div>Вставить</div> <div> <div>Ж</div> <div>К</div> <div>Ч</div> </div> <div> <div>Ц</div> <div>А</div> <div>В</div> </div> <div> <div>Шрифт</div> <div>Выравнивание</div> <div>Число</div> </div> </div> <div> <div>Общий</div> <div>Условное форматирование</div> <div>Форматировать как таблицу</div> <div>Стили</div> </div> <div> <div>Вставить</div> <div>Удалить</div> <div>Сортировка</div> <div>Найти и выделить</div> </div> <div> <div>Формат</div> <div>Ячейки</div> <div>Редактирование</div> </div> |             |   |         |      |             |           |            |         |         |         |    |    |            |              |
| B2 fx Achipteria coleoptrata (Linnaeus, 1758)   |             |   |         |      |             |           |            |         |         |         |    |    |            |              |
| A   | B           | C   | D       | E    | F           | G         | H          | I       | J       | K       | L  | M  | N          | O            |
| 5   | lee-2010-04 | Achipteria coleoptrata (Linnaeus, 1758)       | SPECIES | 2010 | 2010-06-0RU | Bezhanits | Polistovsk | 57.0568 | 30.6463 | WGS1984 | 10 | 5  | ind / soil | caspian fore |
| 6   | lee-2010-05 | Galluma obvia (Berlese, 1914)                 | SPECIES | 2010 | 2010-06-0RU | Bezhanits | Polistovsk | 57.0993 | 30.3815 | WGS1984 | 10 | 3  | ind / soil | c bog        |
| 7   | lee-2010-06 | Tectocephus velatus (Michael 1880)            | SPECIES | 2010 | 2010-06-0RU | Bezhanits | Polistovsk | 57.0993 | 30.3815 | WGS1984 | 10 | 1  | ind / soil | Bottomlar    |
| 8   | lee-2010-07 | Achipteria coleoptrata (Linnaeus, 1758)       | SPECIES | 2010 | 2010-06-0RU | Bezhanits | Polistovsk | 57.1019 | 30.3891 | WGS1984 | 10 | 25 | ind / soil | mixed fore   |
| 9   | lee-2010-08 | Medioplia hygrophila (Mahunka 1987)           | SPECIES | 2010 | 2010-06-0RU | Bezhanits | Polistovsk | 57.1032 | 30.3907 | WGS1984 | 10 | 52 | ind / soil | peatbog      |
| 10  | lee-2010-09 | Scheloniates laevigatus (Koch, 1835)          | SPECIES | 2010 | 2010-06-0RU | Bezhanits | Polistovsk | 57.4417 | 30.6801 | WGS1984 | 10 | 36 | ind / soil | meadow       |
| 11  | lee-2010-10 | Microtritia minima (Berlese, 1904)            | SPECIES | 2010 | 2010-06-0RU | Bezhanits | Polistovsk | 57.4417 | 30.6801 | WGS1984 | 10 | 26 | ind / soil | meadow       |
| 12  | lee-2010-11 | Rhyotritia duplicata (Grandjean 1953)         | SPECIES | 2010 | 2010-06-0RU | Bezhanits | Polistovsk | 57.4417 | 30.6801 | WGS1984 | 10 | 4  | ind / soil | meadow       |
| 13  | lee-2010-12 | Scheloniates laevigatus (Koch, 1835)          | SPECIES | 2010 | 2010-06-0RU | Bezhanits | Polistovsk | 57.102  | 30.4306 | WGS1984 | 10 | 21 | ind / soil | mixed fore   |
| 14  | lee-2010-13 | Parakalumnidae                                | SPECIES | 2010 | 2010-06-0RU | Bezhanits | Polistovsk | 57.057  | 30.6418 | WGS1984 | 10 | 15 | ind / soil | mixed fore   |
| 15  | lee-2010-14 | Platynothrus peltifer (Koch, 1840)            | SPECIES | 2010 | 2010-06-0RU | Bezhanits | Polistovsk | 57.057  | 30.6418 | WGS1984 | 10 | 20 | ind / soil | mixed fore   |
| 16  | lee-2010-15 | Xenillus tegeocranus (Hermann 1804)           | SPECIES | 2010 | 2010-06-0RU | Bezhanits | Polistovsk | 57.057  | 30.6418 | WGS1984 | 10 | 2  | ind / soil | mixed fore   |
| 17  | lee-2010-16 | Hoplophthiracarus illinoisensis (Ewing, 1909) | SPECIES | 2010 | 2010-06-0RU | Bezhanits | Polistovsk | 57.171  | 30.6404 | WGS1984 | 10 | 24 | ind / soil | Transition   |
| 18  | lee-2010-17 | Tectocephus velatus (Michael 1880)            | SPECIES | 2010 | 2010-06-0RU | Bezhanits | Polistovsk | 57.171  | 30.6404 | WGS1984 | 10 | 3  | ind / soil | Transition   |
| 19  | lee-2010-18 | Trichoribates trimaculatus (C.L.Koch, 1836)   | SPECIES | 2010 | 2010-06-0RU | Bezhanits | Polistovsk | 57.171  | 30.6404 | WGS1984 | 10 | 1  | ind / soil | Transition   |
| 20  | lee-2010-19 | Oppliella nova (Oudemans, 1902)               | SPECIES | 2010 | 2010-06-0RU | Bezhanits | Polistovsk | 57.1042 | 30.39   | WGS1984 | 10 | 28 | ind / soil | Raised pe    |
| 21  | lee-2010-20 | Nanhermannia dorsalis (Banks, 1896)           | SPECIES | 2010 | 2010-06-0RU | Bezhanits | Polistovsk | 57.1042 | 30.39   | WGS1984 | 10 | 3  | ind / soil | Raised pe    |
| 22  | lee-2010-21 | Phthiracarus globosus (Koch, 1841)            | SPECIES | 2010 | 2010-06-0RU | Bezhanits | Polistovsk | 57.1019 | 30.3891 | WGS1984 | 10 | 1  | ind / soil | Spruce for   |
| 23  | lee-2010-22 | Oppliella nova (Oudemans, 1902)               | SPECIES | 2010 | 2010-06-0RU | Bezhanits | Polistovsk | 57.0984 | 30.3809 | WGS1984 | 10 | 2  | ind / soil | Transition   |
| 24  | lee-2010-23 | Tectocephus velatus (Michael 1880)            | SPECIES | 2010 | 2010-06-0RU | Bezhanits | Polistovsk | 57.0984 | 30.3809 | WGS1984 | 10 | 11 | ind / soil | Transition   |
| 25  | lee-2010-24 | Zetomimus furcatus (Warburton & Pearce 1902)  | SPECIES | 2010 | 2010-06-0RU | Bezhanits | Polistovsk | 57.0984 | 30.3809 | WGS1984 | 10 | 1  | ind / soil | Transition   |
| 26  | lee-2010-25 | Hoplophthiracarus illinoisensis (Ewing, 1909) | SPECIES | 2010 | 2010-06-0RU | Bezhanits | Polistovsk | 30.6452 | 57.1743 | WGS1984 | 10 | 12 | ind / soil | Transition   |
| 27  | lee-2010-26 | Scheloniates laevigatus (Koch, 1835)          | SPECIES | 2010 | 2010-06-0RU | Bezhanits | Polistovsk | 57.1743 | 30.6452 | WGS1984 | 10 | 1  | ind / soil | Transition   |
| 28  | lee-2010-27 | Chamobates cuspidatus (Michael 1880)          | SPECIES | 2010 | 2010-06-0RU | Bezhanits | Polistovsk | 57.3531 | 30.8125 | WGS1984 | 10 | 1  | ind / soil | Transition   |
| 29  | lee-2010-28 | Scheloniates latipes (C.L.Koch, 1844)         | SPECIES | 2010 | 2010-06-0RU | Bezhanits | Polistovsk | 57.3531 | 30.8125 | WGS1984 | 10 | 6  | ind / soil | Transition   |
| 30  | lee-2010-29 | Minuthozetes seminufus (Koch, 1841)           | SPECIES | 2010 | 2010-06-0RU | Bezhanits | Polistovsk | 57.3531 | 30.8125 | WGS1984 | 10 | 4  | ind / soil | Transition   |

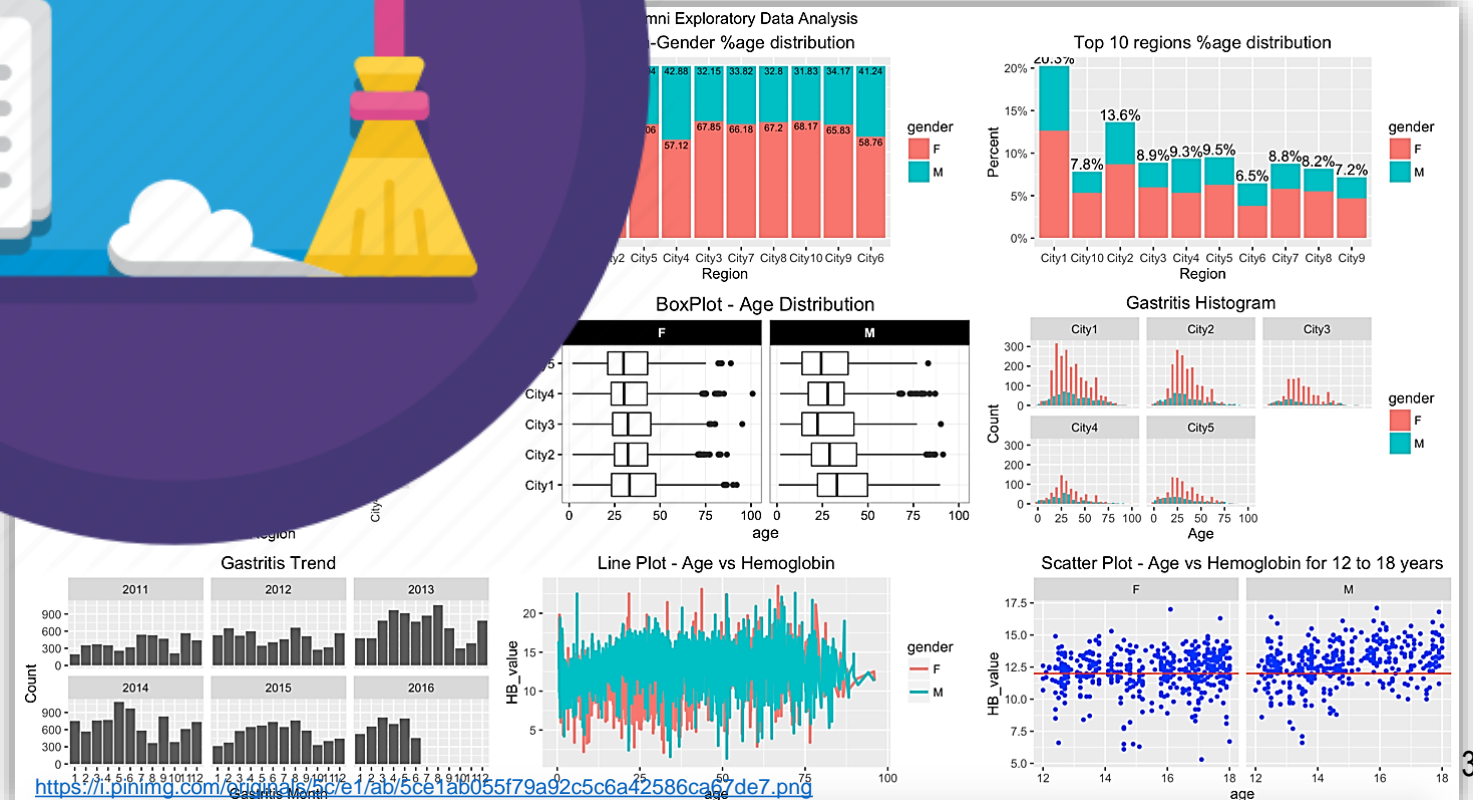
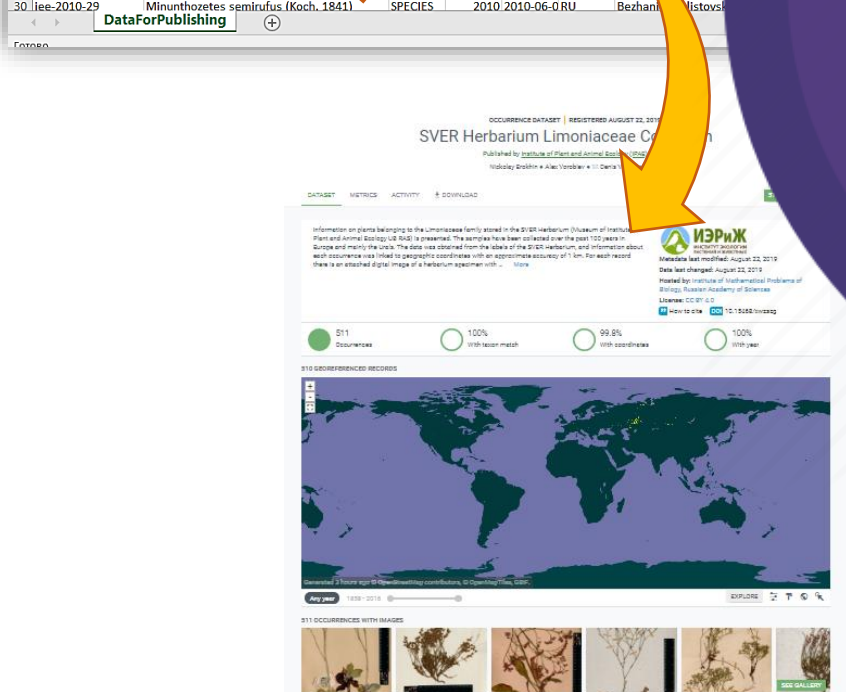




DataCleaningExample - Excel (Сбой активации продукта)

|    | A           | B   | C       | D    | E         | F  | G         | H          | I       | J       | K       | L |
|----|-------------|---|---------|------|-----------|----|-----------|------------|---------|---------|---------|---|
| 5  | lee-2010-04 | Achipteria coleoprata (Linnaeus, 1758)        | SPECIES | 2010 | 2010-06-0 | RU | Bezhanits | Polistovsk | 57.0568 | 30.6463 | WGS1984 |   |
| 6  | lee-2010-05 | Galluma obvia (Berlese, 1914)                 | SPECIES | 2010 | 2010-06-0 | RU | Bezhanits | Polistovsk | 57.0993 | 30.3815 | WGS1984 |   |
| 7  | lee-2010-06 | Tectocephus velatus (Michael 1880)            | SPECIES | 2010 | 2010-06-0 | RU | Bezhanits | Polistovsk | 57.0993 | 30.3815 | WGS1984 |   |
| 8  | lee-2010-07 | Achipteria coleoprata (Linnaeus, 1758)        | SPECIES | 2010 | 2010-06-0 | RU | Bezhanits | Polistovsk | 57.1019 | 30.3891 | WGS1984 |   |
| 9  | lee-2010-08 | Medioplia hygrophila (Mahunka 1987)           | SPECIES | 2010 | 2010-06-0 | RU | Bezhanits | Polistovsk | 57.1032 | 30.3907 | WGS1984 |   |
| 10 | lee-2010-09 | Scheloribates laevigatus (Koch, 1835)         | SPECIES | 2010 | 2010-06-0 | RU | Bezhanits | Polistovsk | 57.4417 | 30.6801 | WGS1984 |   |
| 11 | lee-2010-10 | Microtritia minima (Berlese, 1904)            | SPECIES | 2010 | 2010-06-0 | RU | Bezhanits | Polistovsk | 57.4417 | 30.6801 | WGS1984 |   |
| 12 | lee-2010-11 | Rhysotritia duplicata (Grandjean 1953)        | SPECIES | 2010 | 2010-06-0 | RU | Bezhanits | Polistovsk | 57.4417 | 30.6801 | WGS1984 |   |
| 13 | lee-2010-12 | Scheloribates laevigatus (Koch, 1835)         | SPECIES | 2010 | 2010-06-0 | RU | Bezhanits | Polistovsk | 57.102  | 30.6801 | WGS1984 |   |
| 14 | lee-2010-13 | Parakalumnidae                                | SPECIES | 2010 | 2010-06-0 | RU | Bezhanits | Polistovsk | 57.057  | 30.6801 | WGS1984 |   |
| 15 | lee-2010-14 | Platynothus peltifer (Koch, 1840)             | SPECIES | 2010 | 2010-06-0 | RU | Bezhanits | Polistovsk | 57.057  | 30.6801 | WGS1984 |   |
| 16 | lee-2010-15 | Xenillus tegeocranus (Hermann 1804)           | SPECIES | 2010 | 2010-06-0 | RU | Bezhanits | Polistovsk | 57.057  | 30.6801 | WGS1984 |   |
| 17 | lee-2010-16 | Hoplophthiracarus illinoisensis (Ewing, 1909) | SPECIES | 2010 | 2010-06-0 | RU | Bezhanits | Polistovsk | 57.057  | 30.6801 | WGS1984 |   |
| 18 | lee-2010-17 | Tectocephus velatus (Michael 1880)            | SPECIES | 2010 | 2010-06-0 | RU | Bezhanits | Polistovsk | 57.057  | 30.6801 | WGS1984 |   |
| 19 | lee-2010-18 | Trichoribates trimaculatus (C.L.Koch, 1836)   | SPECIES | 2010 | 2010-06-0 | RU | Bezhanits | Polistovsk | 57.057  | 30.6801 | WGS1984 |   |
| 20 | lee-2010-19 | Oppliella nova (Oudemans, 1902)               | SPECIES | 2010 | 2010-06-0 | RU | Bezhanits | Polistovsk | 57.057  | 30.6801 | WGS1984 |   |
| 21 | lee-2010-20 | Nanhermannia dorsalis (Banks, 1896)           | SPECIES | 2010 | 2010-06-0 | RU | Bezhanits | Polistovsk | 57.057  | 30.6801 | WGS1984 |   |
| 22 | lee-2010-21 | Phthiracarus globosus (Koch, 1841)            | SPECIES | 2010 | 2010-06-0 | RU | Bezhanits | Polistovsk | 57.057  | 30.6801 | WGS1984 |   |
| 23 | lee-2010-22 | Opiella nova (Oudemans, 1902)                 | SPECIES | 2010 | 2010-06-0 | RU | Bezhanits | Polistovsk | 57.057  | 30.6801 | WGS1984 |   |
| 24 | lee-2010-23 | Tectocephus velatus (Michael 1880)            | SPECIES | 2010 | 2010-06-0 | RU | Bezhanits | Polistovsk | 57.057  | 30.6801 | WGS1984 |   |
| 25 | lee-2010-24 | Zetomimus furcatus (Warburton & Pearce 1902)  | SPECIES | 2010 | 2010-06-0 | RU | Bezhanits | Polistovsk | 57.057  | 30.6801 | WGS1984 |   |
| 26 | lee-2010-25 | Hoplophthiracarus illinoisensis (Ewing, 1909) | SPECIES | 2010 | 2010-06-0 | RU | Bezhanits | Polistovsk | 57.057  | 30.6801 | WGS1984 |   |
| 27 | lee-2010-26 | Scheloribates laevigatus (Koch, 1835)         | SPECIES | 2010 | 2010-06-0 | RU | Bezhanits | Polistovsk | 57.057  | 30.6801 | WGS1984 |   |
| 28 | lee-2010-27 | Chamobates cuspidatus (Michael, 1880)         | SPECIES | 2010 | 2010-06-0 | RU | Bezhanits | Polistovsk | 57.057  | 30.6801 | WGS1984 |   |
| 29 | lee-2010-28 | Scheloribates latipes (C.L.Koch, 1844)        | SPECIES | 2010 | 2010-06-0 | RU | Bezhanits | Polistovsk | 57.057  | 30.6801 | WGS1984 |   |
| 30 | lee-2010-29 | Minunthozetes seminifus (Koch, 1841)          | SPECIES | 2010 | 2010-06-0 | RU | Bezhanits | Polistovsk | 57.057  | 30.6801 | WGS1984 |   |

Formula bar: B2 = Achipteria coleoprata (Linnaeus, 1758)



<https://i.pinimg.com/originals/5f/79/a9/5f79a92c5c6a42586ca9de7.png>

# Данные всегда содержат ошибки

Качество данных – это относительная концепция, которая зависит от способа использования этих данных



# Наиболее распространенные ошибки в данных

**Технические ошибки:** опечатки, пропущенные значения, лишние пробелы, корректность диапазонов для дат, соответствие типа данных полю, в котором они содержатся

Ошибки формата данных

**Согласованность данных:** соответствие даты сбора, идентификации, обновления и оцифровки, координаты всех точек находятся в указанном регионе, точки находок сухопутных видов находятся на суше и т.д.

Номенклатурные ошибки:  
соответствие названия таксонов выбранному справочнику  
Соответствие других значений справочным



# Инструменты для поиска и исправления технических ошибок и ошибок в данных

Текстовые редакторы

- BBEdit (Mac)
- Notepad++ (Windows)
- Emacs, vi (Unix, Linux)

- R (командная строка)
- RStudio (графический пользовательский интерфейс)



**Выберите то, что удобно вам!**

# Кодировка

Каким способом набор кодов символов переводится в байты для обработки компьютером.  
Каждому символу компьютер присваивает уникальный номер.

**A = 065 = 41 = 01000001**

| Glyph | ASCII | ISO-8859 Western Europe | UTF-8 | UTF-16 | UTF-32      | Binary   |
|-------|-------|-------------------------|-------|--------|-------------|----------|
| A     | 065   | 041                     | 41    | 00 41  | 00 00 00 41 | 01000001 |

# Кодировка

Почему это важно?

To be, or not to be, that is the question.  
(Быть или не быть, вот в чем вопрос)

Être, ou ne pas être, c'est là la question.



# Кодировка

Почему это важно?

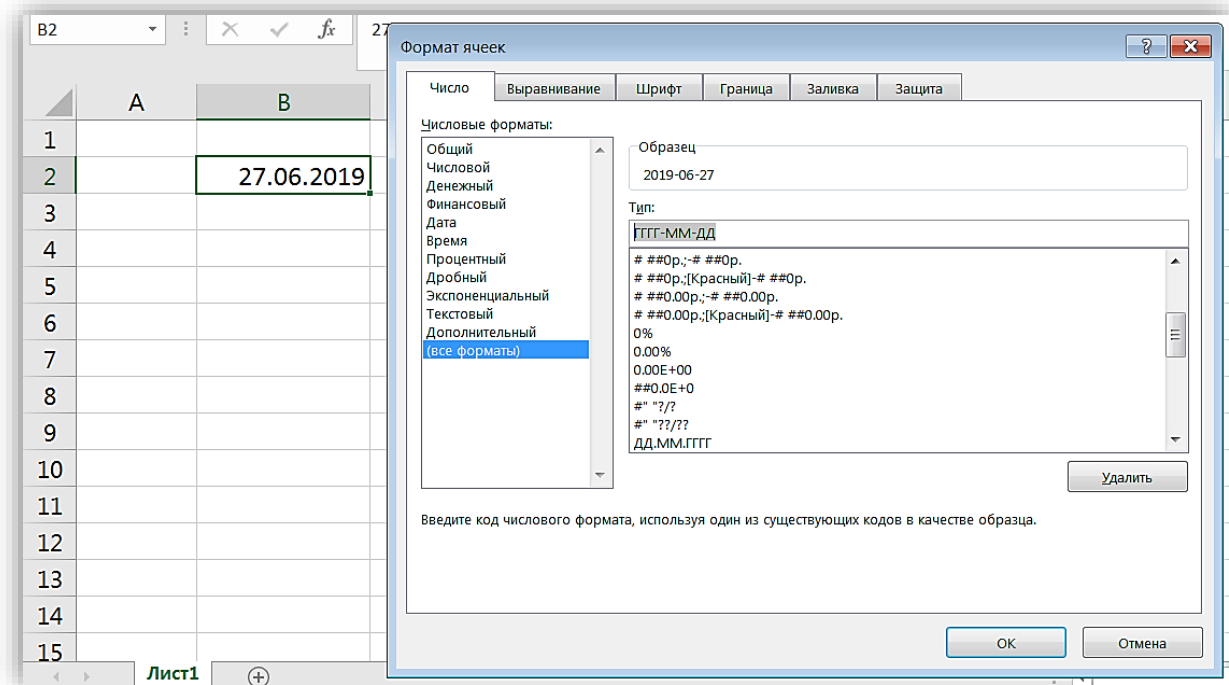
To be, or not to be, that is the question.

Être, ou ne pas être, c'est là la question.

❖tre, ou ne pas ❖tre, c❖est l❖ la question.

# Как задать необходимый формат даты в MS Excel

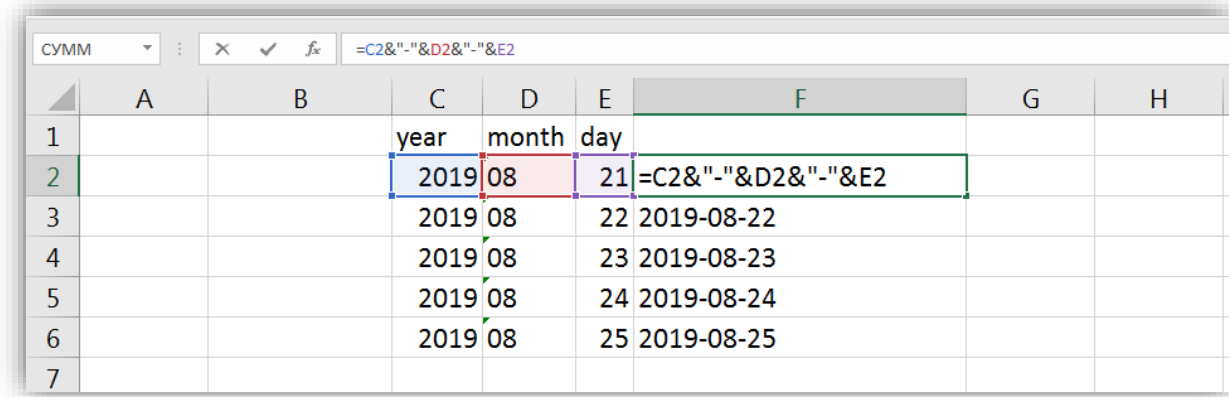
Способ 1



## Ошибки формата: даты

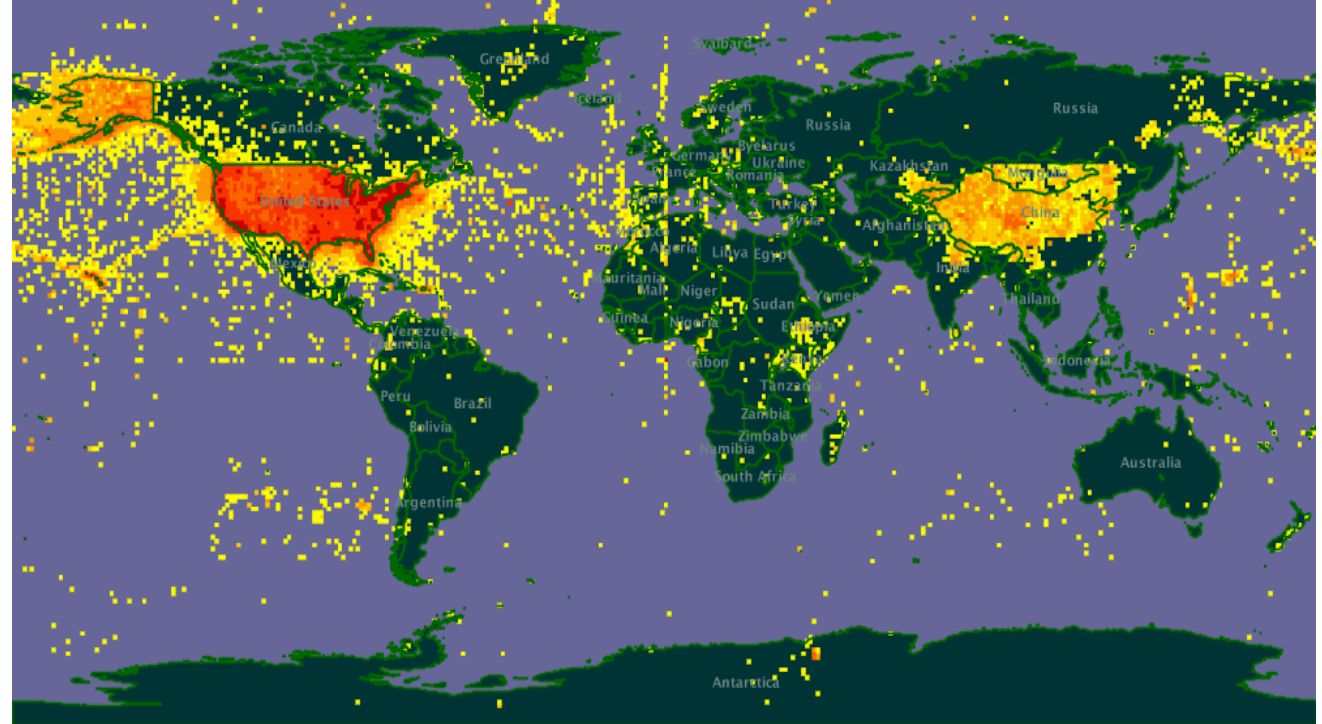
| eventDate     | verbatimEventDate  |
|---------------|--------------------|
| 2019-08-27    | 27 авг 2019        |
|               | 27 VIII 2019       |
|               | 27.08.2019         |
| 2019-08-29/30 | 29-30 августа 2019 |

Способ 2



# Пространственные данные: наиболее распространенные технические ошибки

- **Широта и долгота перепутаны местами**
- Неправильно указано полушарие
- Нулевые значения
- Неизвестная система координат
- Ошибки преобразования координат из одной системы в другую или из одной формы представления в другую



Ранняя GBIF карта, иллюстрирующая данные из США, с широко распространенными ошибками:

- Координаты 0,0 (Гринвичский меридиан и Экватор)
- Неправильно указано полушарие (точки с неверной (восточной) долготой попадают в Китай, с неверной (южной) широтой - в Чили).

# QGIS: открытая ГИС



- Настольная (локальная) геоинформационная система (ГИС)
- Для трансформации, анализа, визуализации, проверки и верификации и т.д.
- <http://www.qgis.org>





# Пространственные данные: ошибки формата

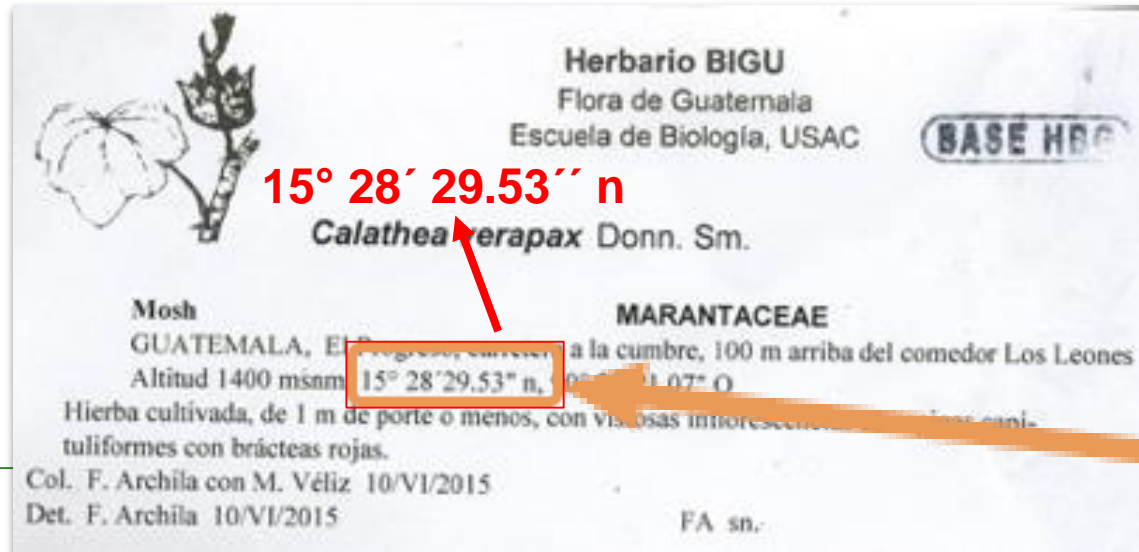
Градусы Минуты Секунды Полушарие →  
десятичные градусы

$$\text{ГГ} = (\text{Г} + \text{М}/60 + \text{С}/3600) * [\text{Полушарие}]$$

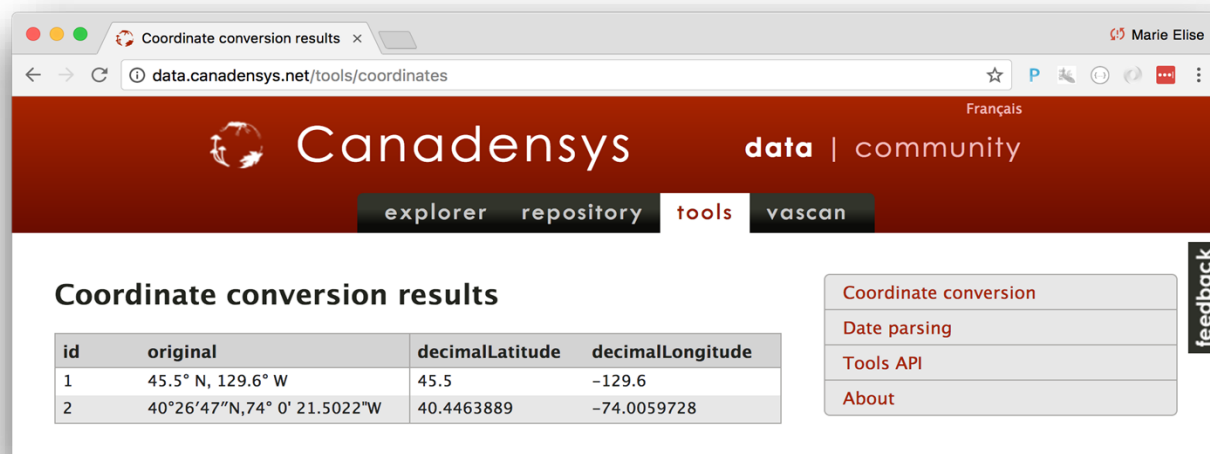
Полушарие: западное = -1; восточное = 1

$$\text{ГГ} = (15 + 28/60 + 29.53/3600) * 1$$

$$\text{ГГ} = 15.47487$$



# Автоматический пересчет координат из ГГ ММ СС в ГГ.ГГГГГ



The screenshot shows a web browser window with the URL [data.canadensys.net/tools/coordinates](http://data.canadensys.net/tools/coordinates). The page has a red header with the Canadensys logo and navigation links: explorer, repository, tools (selected), and vascan. Below the header, the title "Coordinate conversion results" is displayed. A table shows the conversion results for two entries:

| id | original                     | decimalLatitude | decimalLongitude |
|----|------------------------------|-----------------|------------------|
| 1  | 45.5° N, 129.6° W            | 45.5            | -129.6           |
| 2  | 40°26'47"N, 74° 0' 21.5022"W | 40.4463889      | -74.0059728      |

To the right of the table is a sidebar with links: Coordinate conversion, Date parsing, Tools API, and About. A vertical "feedback" button is on the far right.

<http://data.canadensys.net/tools/coordinates?lang=en>



The screenshot shows the GIS-LAB website. The header includes the GIS-LAB logo, the text "Географические информационные системы и дистанционное зондирование", and social media links for Twitter, Facebook, and Google+. A search bar is on the right. Below the header, a navigation menu lists: Статьи, Документация, Геоданные, О GIS-Lab, С чего начать?, Форум, Блог, and Реклама. The main content area features a title "Конвертация значений координат в формате DDMMSS в формат DD.DDDD" and a description "Как переводить координаты из одного числового формата в другой". There is a "Select Language" dropdown and a note "Powered by Google Translate". A blue button labeled "Обсудить в форуме" (5 comments) and a link to "Редктировать в вики" are also visible. The GIS-LAB logo is in the bottom right corner.

<http://gis-lab.info/qa/dms2dd.html>

# Проверка корректности данных

Массив данных содержит образцы окаменелостей Триасового периода.

Представлены записи для образца рода *Thismia*.

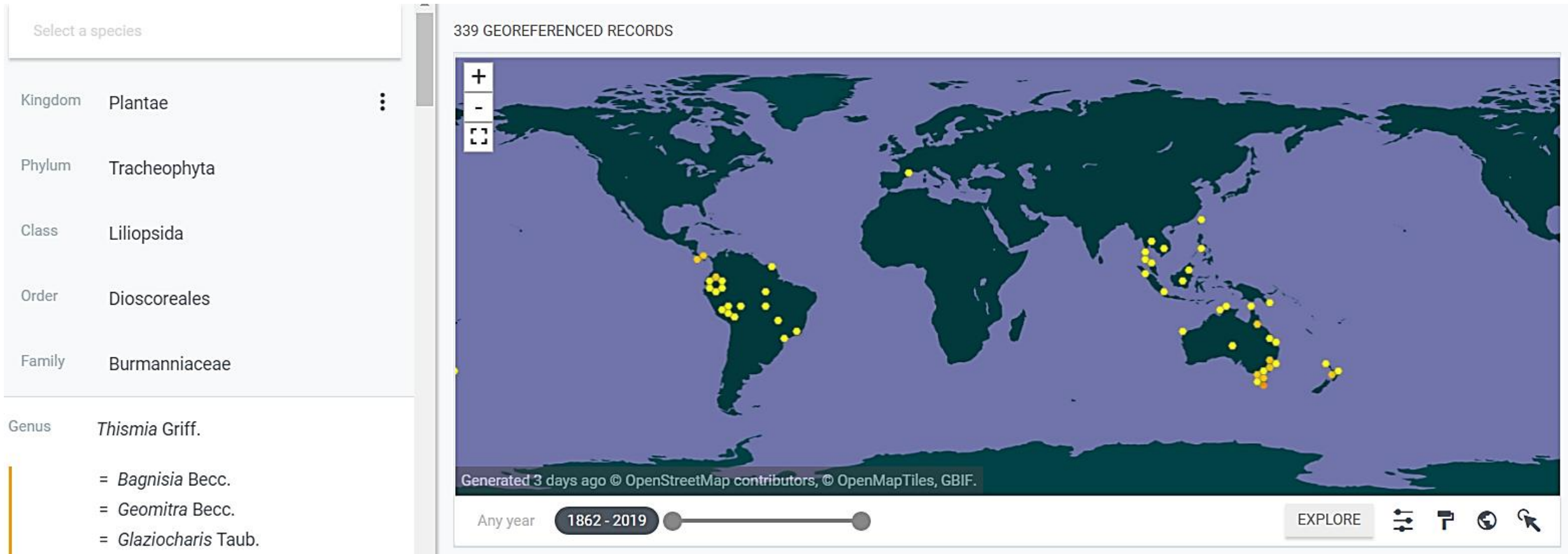
***Thismia* – это ископаемый вид?**



















# Проверка корректности данных

## *Thismia* – род современных цветковых растений





# Проверка данных на соответствие базовой таксономии GBIF: поиск номенклатурных ошибок с помощью Species matching

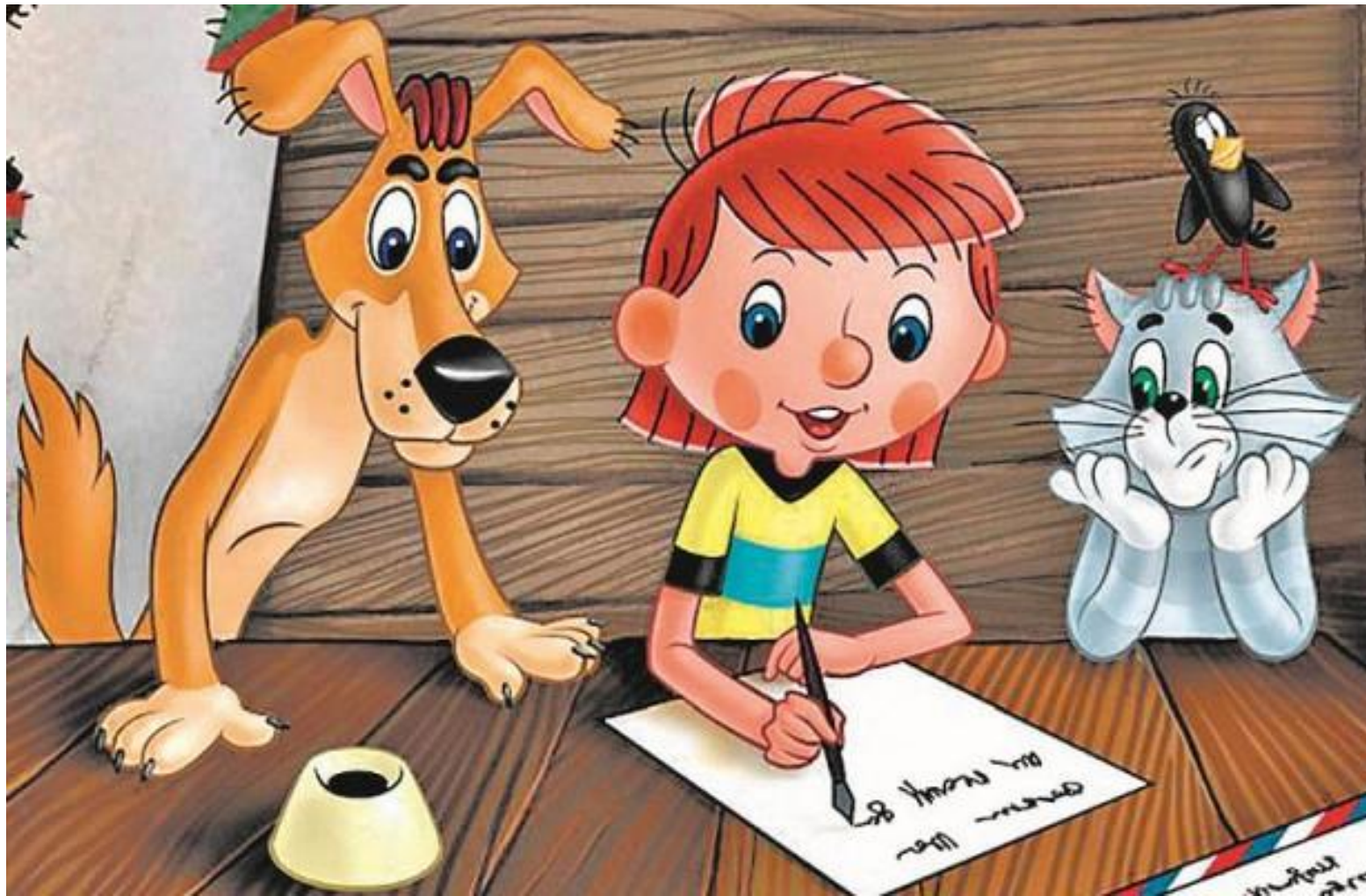
| <div>Get dataShareToolsInside GBIF<div></div>Login</div> |                 |            |            |   |          |         |
|---|-----------------|------------|------------|---|----------|---------|
| TOOLS   LOOK UP   |                 |            |            |   |          |         |
| verbatimScientificName  | preferedKingdom | matchType  | confidence | scientificName (editable)   | status   | rank    |
| Achipteria coleoptrata (Linnaeus, 1758)   | animalia        | EXACT      | 100        |  Achipteria coleoptrata (Linnaeus, 1758)         | ACCEPTED | Species |
| Belba corynopus (Hermann, 1804)   | animalia        | EXACT      | 100        |  Belba corynopus (Hermann, 1804)                 | ACCEPTED | Species |
| Cepheus cepheiformis (Nicolet, 1855)  | animalia        | EXACT      | 100        |  Cepheus cepheiformis (Nicolet, 1855)            | ACCEPTED | Species |
| Chamobates cuspidatus (Michael, 1884)   | animalia        | EXACT      | 100        |  Chamobates cuspidatus (Michael, 1884)           | ACCEPTED | Species |
| Conchogneta willmanni (Dyrdowska, 1929)   | animalia        | EXACT      | 100        |  Conchogneta willmanni (Dyrdowska, 1929)         | ACCEPTED | Species |
| Eupelops acromios (Hermann, 1804)   | animalia        | EXACT      | 100        |  Eupelops acromios (Hermann, 1804)             | ACCEPTED | Species |
| Galluma obvia (Berlese, 1914)   | animalia        | HIGHERRANK | 99         |  Animalia                                      | ACCEPTED | Kingdom |
| Galumna obvia (Berlese, 1914)   | animalia        | EXACT      | 100        |  Galumna obvia (Berlese, 1914)                 | ACCEPTED | Species |
| Hoplophthiracarus illinoisensis (Ewing, 1909)   | animalia        | EXACT      | 100        |  Hoplophthiracarus illinoisensis (Ewing, 1909) | ACCEPTED | Species |

## Как вносить исправления?

| ID записи | Ошибка                                    | Какое исправление<br>сделано           | Кто внес исправление | Дата       |
|-----------|---|--|----------------------|------------|
| ISEE-1245 | Неправильно указана<br>широта 45° 71.345' | Исправлено на 45.71345                 | Сидоров И.И.         | 2010-08-05 |
| ISEE-8354 | Дата сбора<br>30 февраля                  | Удалено                                | Пахомов А.Е.         | 2013-12-25 |
| ISEE-0507 | Дубль записи ISEE-<br>05077               | Запись 05077 удалена                   | Боровиков Н.Н.       | 2015-05-10 |
| ISEE-8932 | Фамилия коллектора<br>указана неверно     | Исправлено с Пономарев на<br>Понамарев | Волков А.А.          | 2017-03-18 |

**Тщательное документирование**  
**Сохранение исходных данных (с ошибками)**

# Зачем документировать исправления?



«Подготовка данных для публикации в Глобальной информационной системе о биоразнообразии GBIF»  
10 октября 2020 г., Екатеринбург

## Лекция 2

# Качество данных. Базовые инструменты для поиска ошибок в данных

Наталья Иванова

Институт математических проблем биологии РАН – филиал ИПМ им. М.В. Келдыша РАН



**Слайды CC BY:**

*Nicolas Noé, Sophie Pamerlon,  
Sharon Grant  
и Наталья Иванова*