

«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



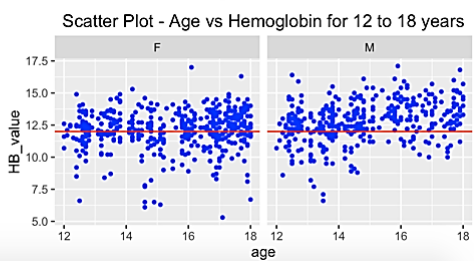
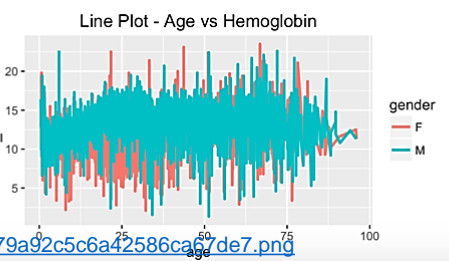
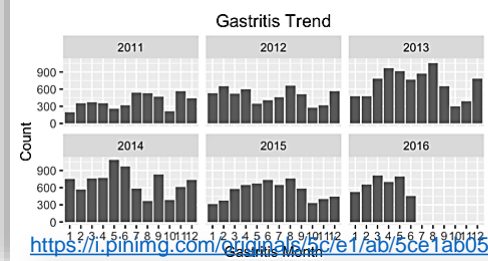
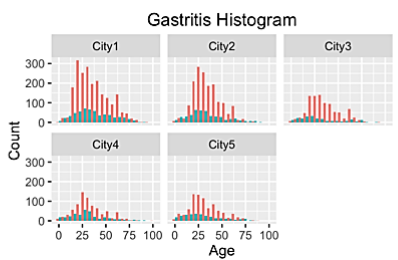
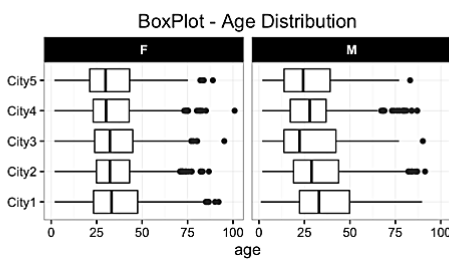
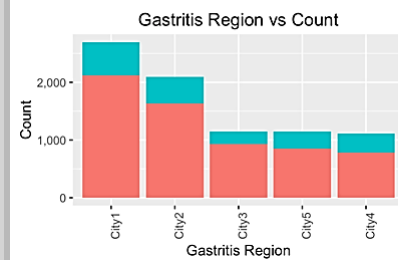
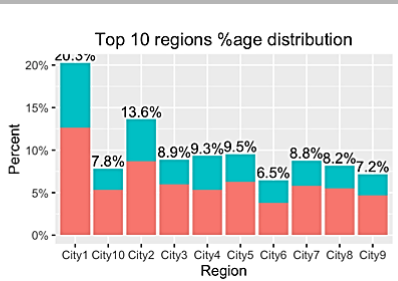
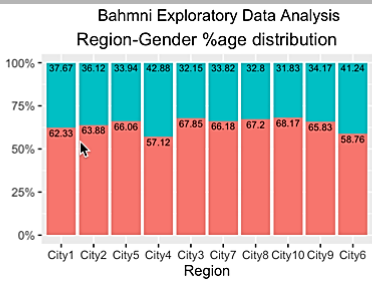
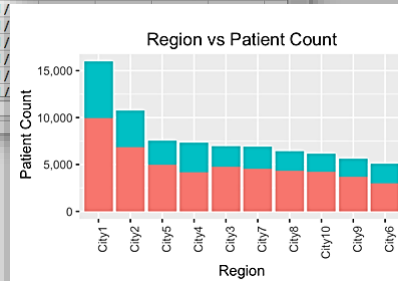
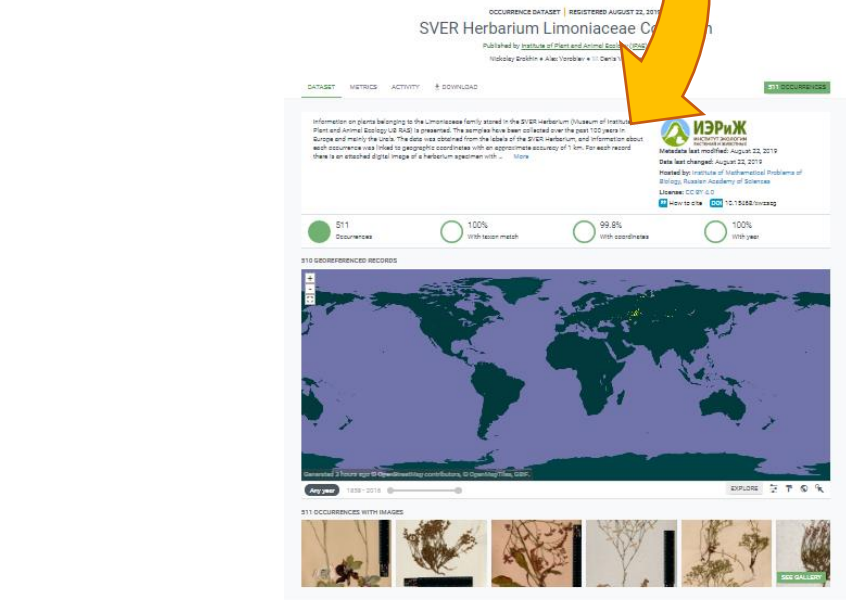
**Слайды CC BY:**

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

DataCleaningExample - Excel (Сбой активации продукта)

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R
5	lee-2010-04	Achipteria coleoprata (Linnaeus, 1758)	SPECIES	2010	2010-06-0R	RU	Bezhanits	Polistovsk	57.0568	30.6463	WGS1984	10	5	ind / soil	c	aspen for	A.S. Zaitse	Zaitsev A. IEE RAS
6	lee-2010-05	Galluma obvia (Berlese, 1914)	SPECIES	2010	2010-06-0R	RU	Bezhanits	Polistovsk	57.0993	30.3815	WGS1984	10	3	ind / soil	c	og	A.S. Zaitse	Zaitsev A. IEE RAS
7	lee-2010-06	Tectocephus velatus (Michael 1880)	SPECIES	2010	2010-06-0R	RU	Bezhanits	Polistovsk	57.0993	30.3815	WGS1984	10	1	ind / soil	c	Bottomlar	A.S. Zaitse	Zaitsev A. IEE RAS
8	lee-2010-07	Achipteria coleoprata (Linnaeus, 1758)	SPECIES	2010	2010-06-0R	RU	Bezhanits	Polistovsk	57.1019	30.3891	WGS1984	10	25	ind / soil	c	mixed for	A.S. Zaitse	Zaitsev A. IEE RAS
9	lee-2010-08	Medioplia hygrophila (Mahunka 1987)	SPECIES	2010	2010-06-0R	RU	Bezhanits	Polistovsk	57.1032	30.3907	WGS1984	10	52	ind / soil	c	peatbog	A.S. Zaitse	Zaitsev A. IEE RAS
10	lee-2010-09	Scheloniates laevigatus (Koch, 1835)	SPECIES	2010	2010-06-0R	RU	Bezhanits	Polistovsk	57.4417	30.6801	WGS1984	10	36	ind / soil	c	meadow	A.S. Zaitse	Zaitsev A. IEE RAS
11	lee-2010-10	Microtritia minima (Berlese, 1904)	SPECIES	2010	2010-06-0R	RU	Bezhanits	Polistovsk	57.4417	30.6801	WGS1984	10	26	ind / soil	c	meadow	A.S. Zaitse	Zaitsev A. IEE RAS
12	lee-2010-11	Rhyotritia duplicata (Grandjean 1953)	SPECIES	2010	2010-06-0R	RU	Bezhanits	Polistovsk	57.4417	30.6801	WGS1984	10	4	ind / soil	c	meadow	A.S. Zaitse	Zaitsev A. IEE RAS
13	lee-2010-12	Scheloniates laevigatus (Koch, 1835)	SPECIES	2010	2010-06-0R	RU	Bezhanits	Polistovsk	57.102	30.4306	WGS1984	10	21	ind / soil	c	mixed for	A.S. Zaitse	Zaitsev A. IEE RAS
14	lee-2010-13	Parakalumnidae	SPECIES	2010	2010-06-0R	RU	Bezhanits	Polistovsk	57.057	30.6418	WGS1984	10	15	ind / soil	c	mixed for	A.S. Zaitse	Zaitsev A. IEE RAS
15	lee-2010-14	Platynothrus peltifer (Koch, 1840)	SPECIES	2010	2010-06-0R	RU	Bezhanits	Polistovsk	57.057	30.6418	WGS1984	10	20	ind / soil	c	mixed for	A.S. Zaitse	Zaitsev A. IEE RAS
16	lee-2010-15	Xenillus tegeocranus (Hermann 1804)	SPECIES	2010	2010-06-0R	RU	Bezhanits	Polistovsk	57.057	30.6418	WGS1984	10	2	ind / soil	c	mixed for	A.S. Zaitse	Zaitsev A. IEE RAS
17	lee-2010-16	Hoplophthiracarus illinoisensis (Ewing, 1909)	SPECIES	2010	2010-06-0R	RU	Bezhanits	Polistovsk	57.171	30.6404	WGS1984	10	24	ind / soil	c	Transition	A.S. Zaitse	Zaitsev A. IEE RAS
18	lee-2010-17	Tectocephus velatus (Michael 1880)	SPECIES	2010	2010-06-0R	RU	Bezhanits	Polistovsk	57.171	30.6404	WGS1984	10	3	ind / soil	c	Transition	A.S. Zaitse	Zaitsev A. IEE RAS
19	lee-2010-18	Trichoribates trimaculatus (C.L.Koch, 1836)	SPECIES	2010	2010-06-0R	RU	Bezhanits	Polistovsk	57.171	30.6404	WGS1984	10	1	ind / soil	c	Transition	A.S. Zaitse	Zaitsev A. IEE RAS
20	lee-2010-19	Oppliella nova (Oudemans, 1902)	SPECIES	2010	2010-06-0R	RU	Bezhanits	Polistovsk	57.1042	30.39	WGS1984	10	28	ind / soil	c	Raised pe	A.S. Zaitse	Zaitsev A. IEE RAS
21	lee-2010-20	Nanhermannia dorsalis (Banks, 1896)	SPECIES	2010	2010-06-0R	RU	Bezhanits	Polistovsk	57.1042	30.39	WGS1984	10	3	ind / soil	c	Raised pe	A.S. Zaitse	Zaitsev A. IEE RAS
22	lee-2010-21	Phthiracarus globosus (Koch, 1841)	SPECIES	2010	2010-06-0R	RU	Bezhanits	Polistovsk	57.1019	30.3891	WGS1984	10	1	ind / soil	c	Spruce for	A.S. Zaitse	Zaitsev A. IEE RAS
23	lee-2010-22	Oppliella nova (Oudemans, 1902)	SPECIES	2010	2010-06-0R	RU	Bezhanits	Polistovsk	57.0984	30.3809	WGS1984	10	2	ind / soil	c	Transition	A.S. Zaitse	Zaitsev A. IEE RAS
24	lee-2010-23	Tectocephus velatus (Michael 1880)	SPECIES	2010	2010-06-0R	RU	Bezhanits	Polistovsk	57.0984	30.3809	WGS1984	10	11	ind / soil	c	Transition	A.S. Zaitse	Zaitsev A. IEE RAS
25	lee-2010-24	Zetomimus furcatus (Warburton & Pearce 1902)	SPECIES	2010	2010-06-0R	RU	Bezhanits	Polistovsk	57.0984	30.3809	WGS1984	10	1	ind / soil	c	Transition	A.S. Zaitse	Zaitsev A. IEE RAS
26	lee-2010-25	Hoplophthiracarus illinoisensis (Ewing, 1909)	SPECIES	2010	2010-06-0R	RU	Bezhanits	Polistovsk	30.6452	57.1743	WGS1984	10	12	ind / soil	c	Transition	A.S. Zaitse	Zaitsev A. IEE RAS
27	lee-2010-26	Scheloniates laevigatus (Koch, 1835)	SPECIES	2010	2010-06-0R	RU	Bezhanits	Polistovsk	57.1743	30.6452	WGS1984	10	1	ind / soil	c	Transition	A.S. Zaitse	Zaitsev A. IEE RAS
28	lee-2010-27	Chamobates cuspidatus (Michael 1880)	SPECIES	2010	2010-06-0R	RU	Bezhanits	Polistovsk	57.3531	30.8125	WGS1984	10	1	ind / soil	c	Transition	A.S. Zaitse	Zaitsev A. IEE RAS
29	lee-2010-28	Scheloniates latipes (C.L.Koch, 1844)	SPECIES	2010	2010-06-0R	RU	Bezhanits	Polistovsk	57.3531	30.8125	WGS1984	10	6	ind / soil	c	Transition	A.S. Zaitse	Zaitsev A. IEE RAS
30	lee-2010-29	Minuthozetes seminufus (Koch, 1841)	SPECIES	2010	2010-06-0R	RU	Bezhanits	Polistovsk	57.3531	30.8125	WGS1984	10	4	ind / soil	c	Transition	A.S. Zaitse	Zaitsev A. IEE RAS

DataForPublishing



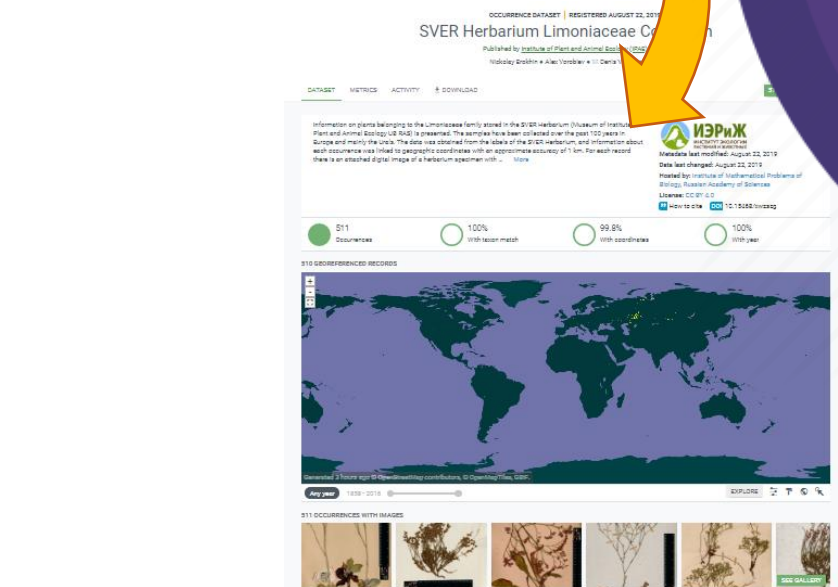
<https://i.pinimg.com/originals/e/1/a/b/5ce1ab055f79a92c5c6a42586ca7de7.png>



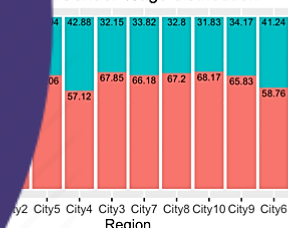
DataCleaningExample - Excel (Сбой активации продукта)

	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 1905)	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	

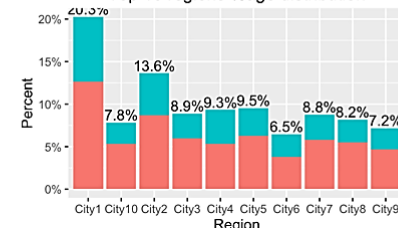
DataForPublishing



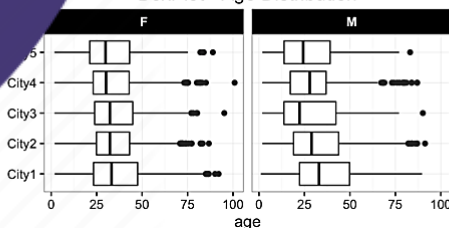
Omni Exploratory Data Analysis



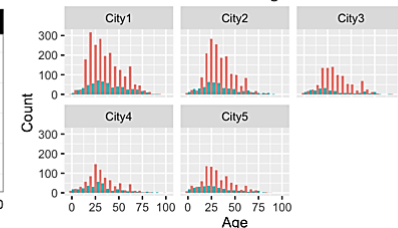
Top 10 regions %age distribution



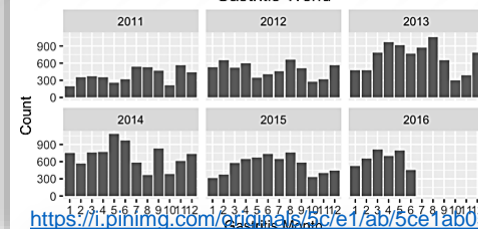
BoxPlot - Age Distribution



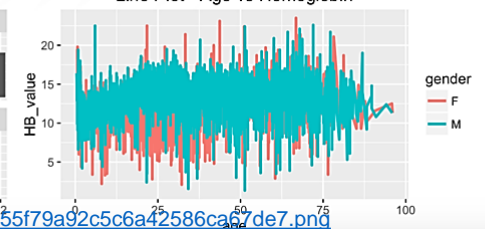
Gastritis Histogram



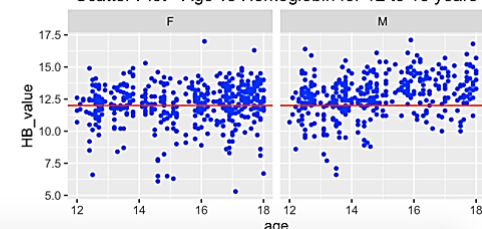
Gastritis Trend



Line Plot - Age vs Hemoglobin



Scatter Plot - Age vs Hemoglobin for 12 to 18 years



# 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)

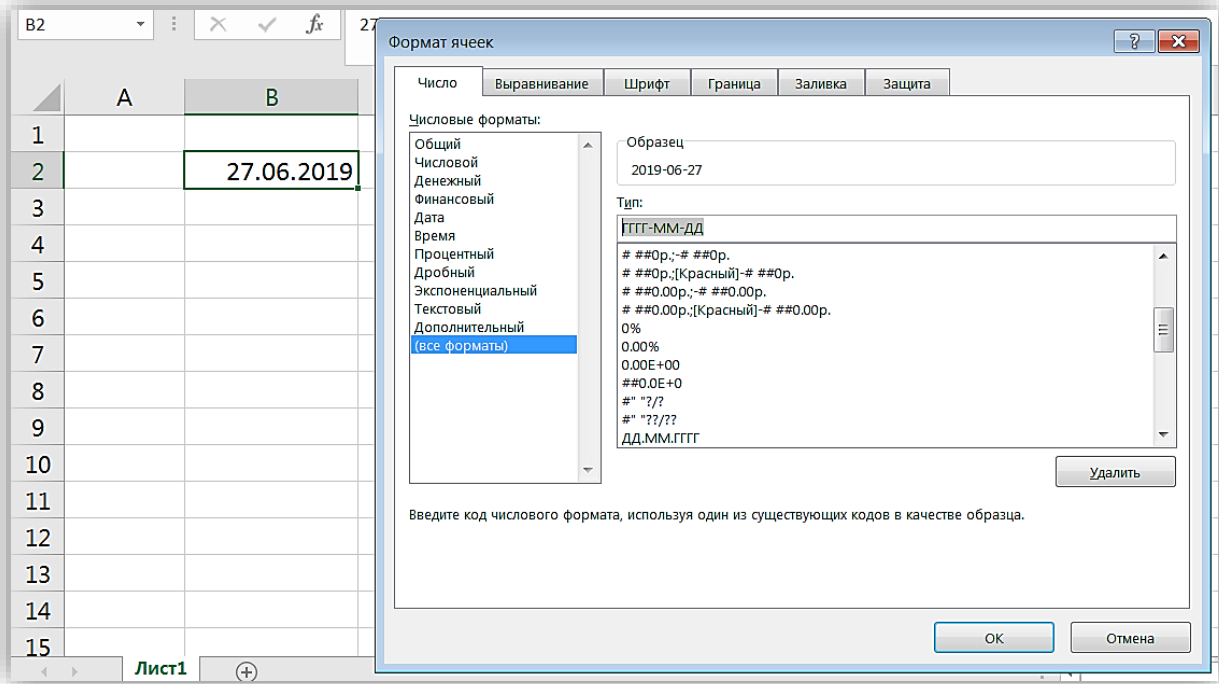
- R (командная строка)
- RStudio (графический пользовательский интерфейс)



**Выберите то, что удобно вам!**

Как задать необходимый формат даты в MS Excel

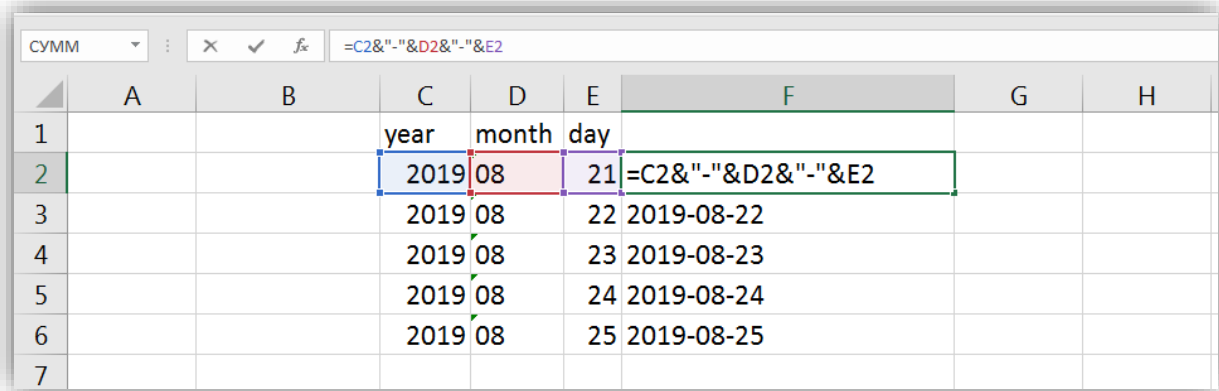
Способ 1



Ошибки формата: даты

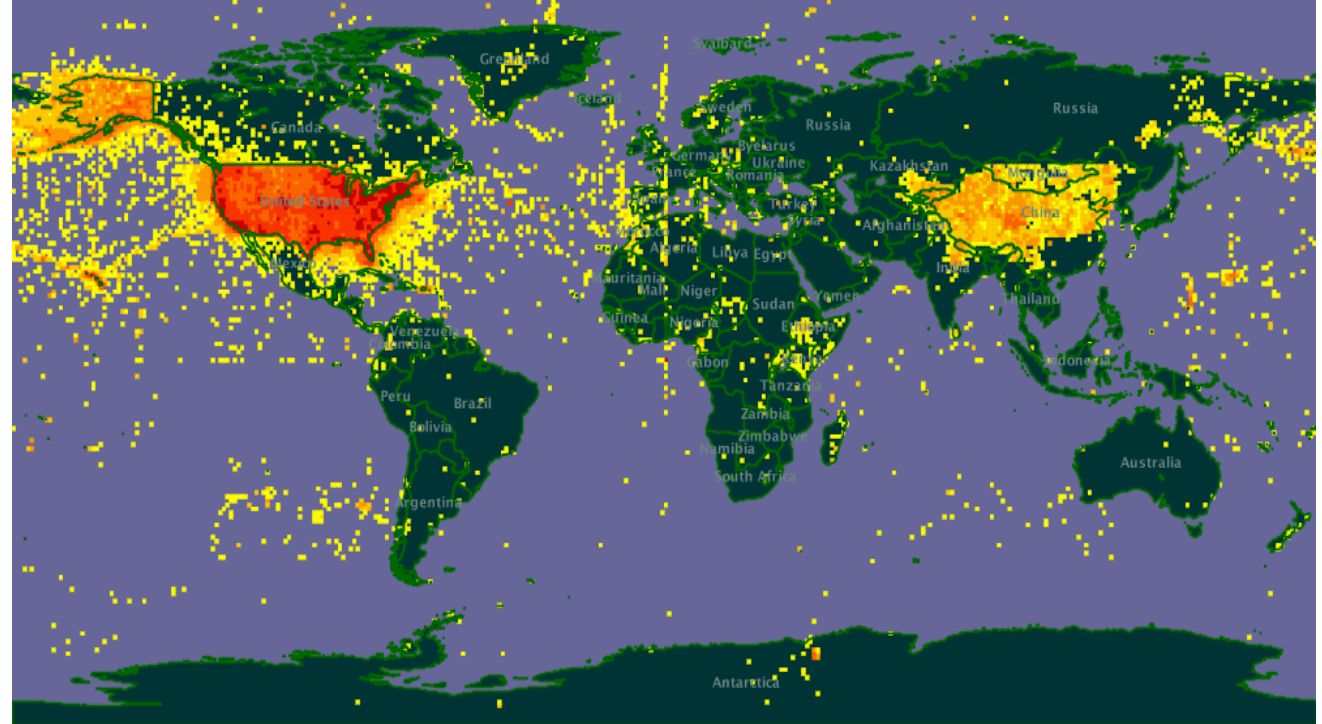
eventDate	verbatimEventDate
2019-08-27	27 авг 2019
	27 VIII 2019
	27.08.2019
2019-08-29/30	29-30 августа 2019

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

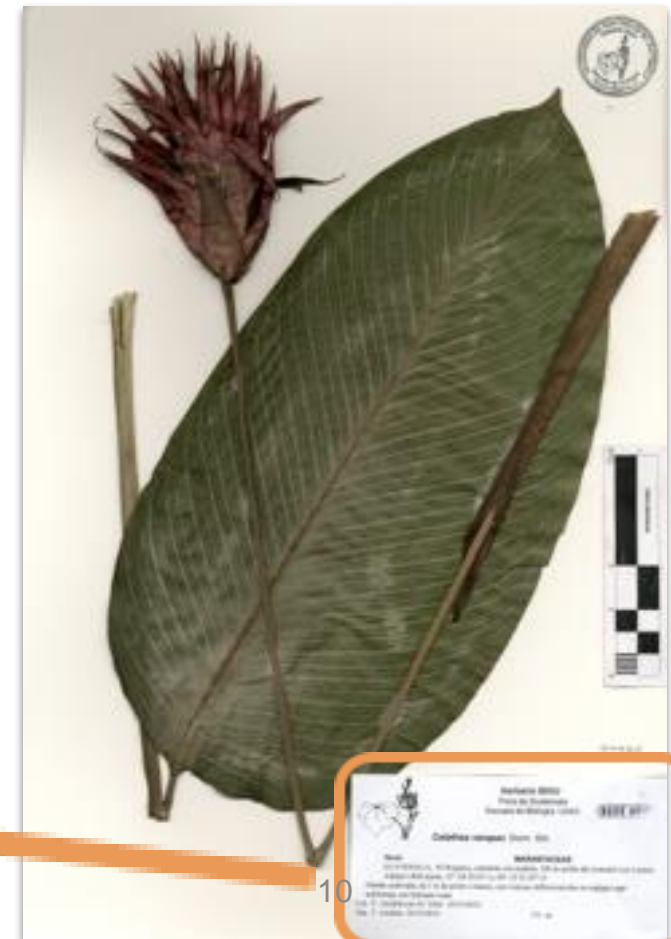
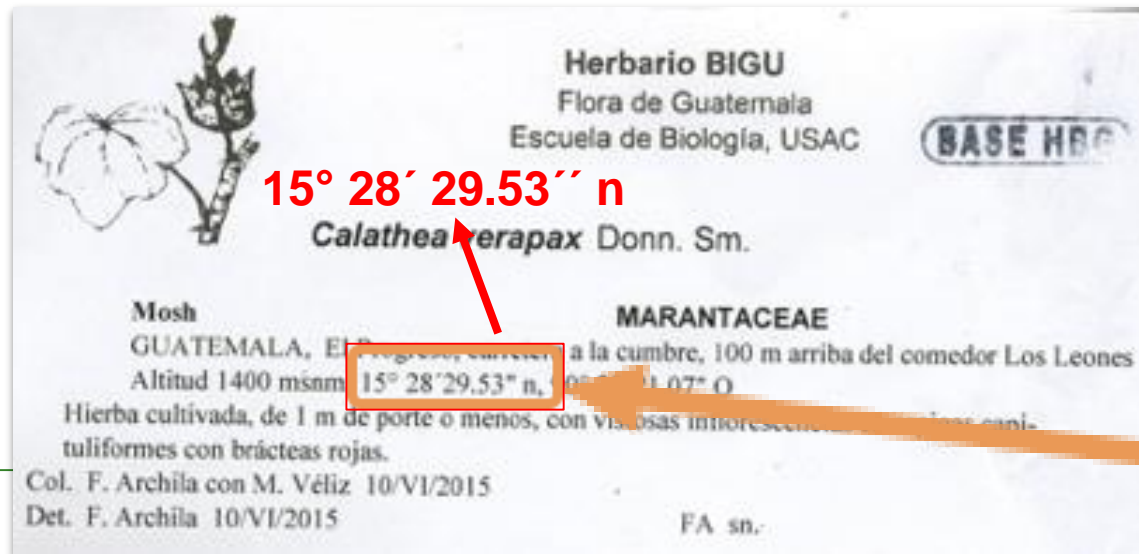
Градусы Минуты Секунды Полушарие →  
десятичные градусы

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

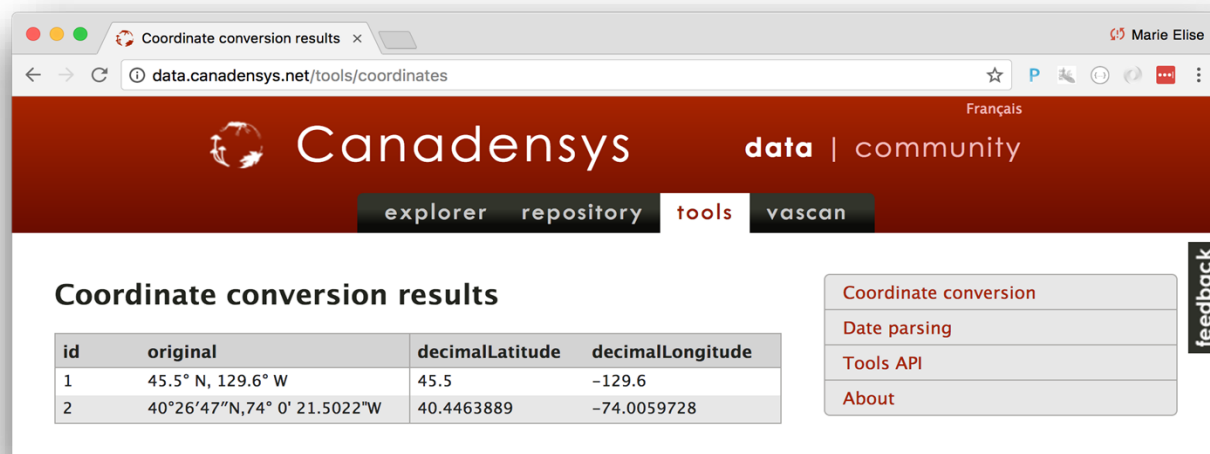
Полушарие: западное = -1; восточное = 1

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

$$\text{ГГ} = 15.47487$$



# 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 is a navigation menu with links: Статьи, Документация, Геоданные, О GIS-Lab, С чего начать?, Форум, Блог, and Реклама. The main content area displays a FAQ entry titled "Конвертация значений координат в формате DDMMSS в формат DD.DDDD". Below the title is a brief description: "Как переводить координаты из одного числового формата в другой". On the right, there is a "Select Language" dropdown and a note "Powered by Google Translate". At the bottom right, there are buttons for "Обсудить в форуме" (5 comments) and "Редактировать в вики". The GIS-LAB logo is also present in the bottom right corner.

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



## Проверка корректности данных

Массив данных содержит образцы окаменелостей Триасового периода.

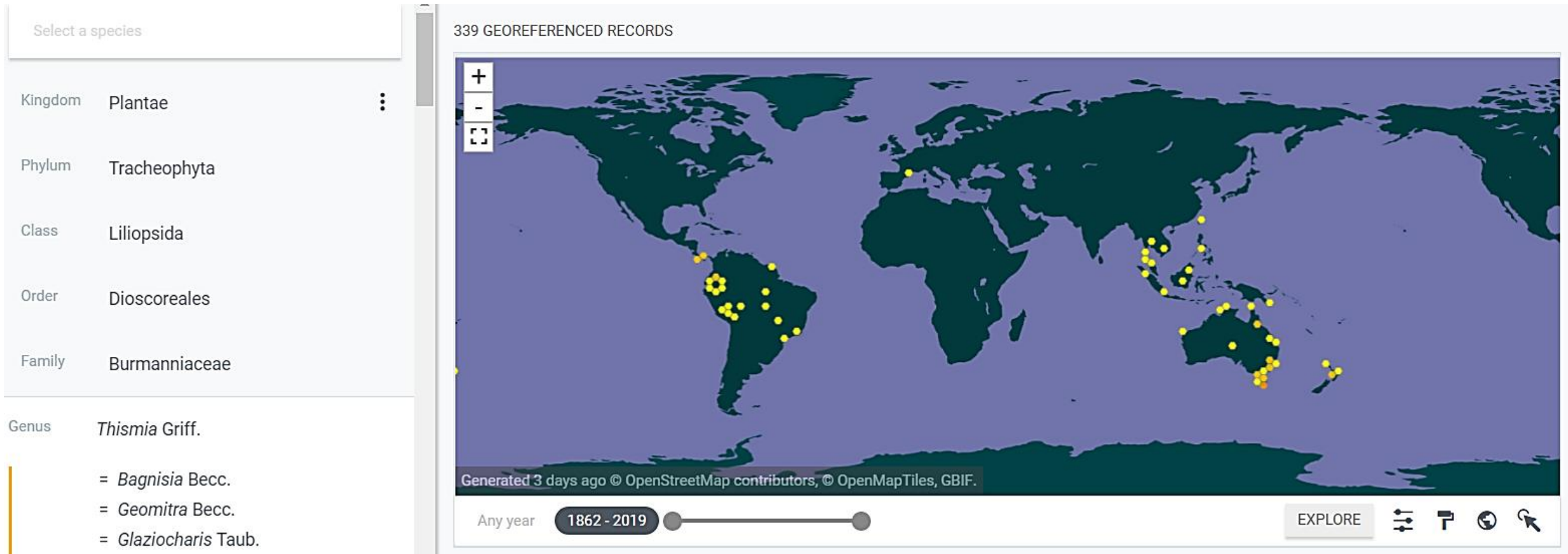
Представлены записи для образца рода *Thismia*.

***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

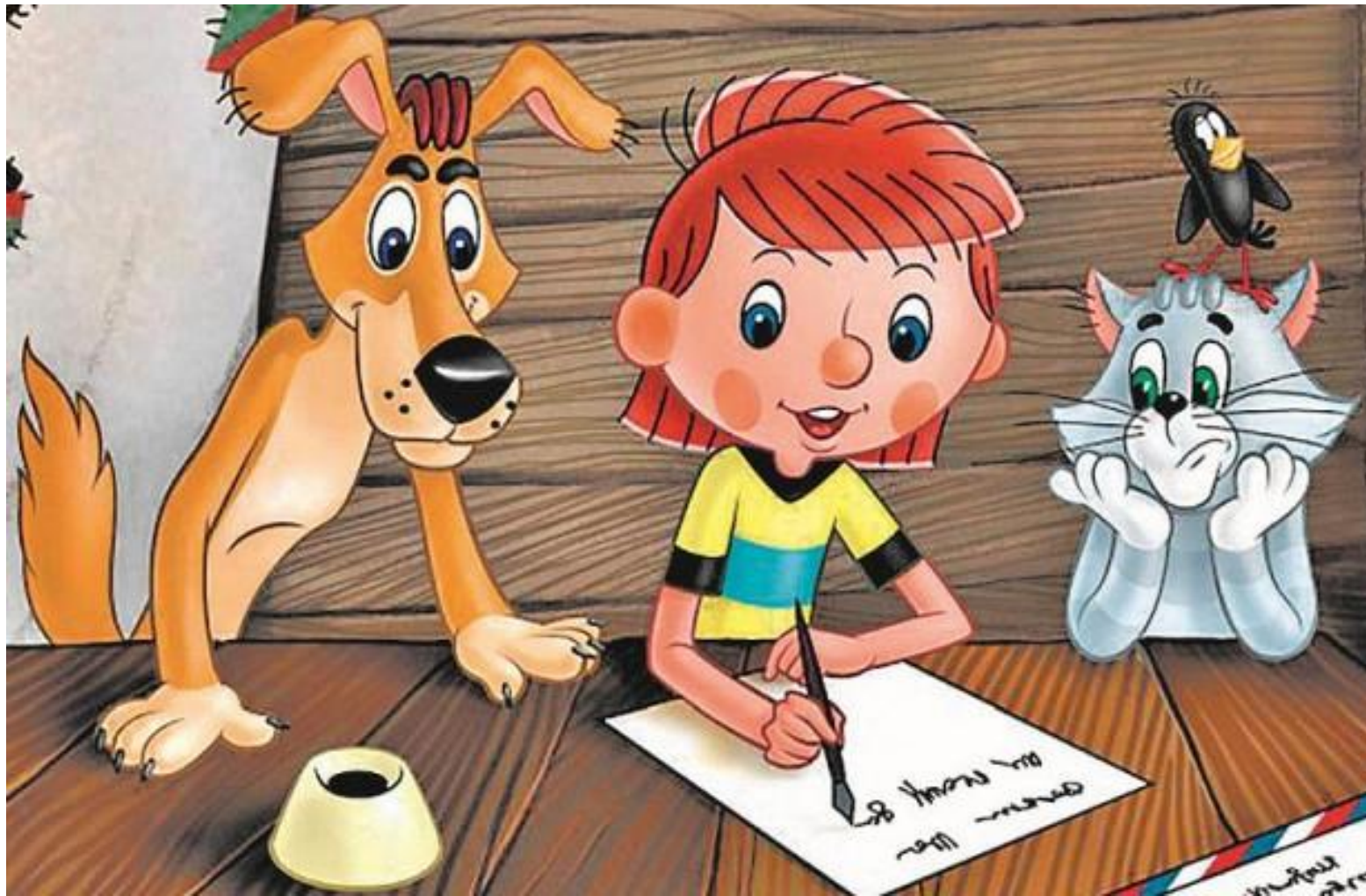


## Как вносить исправления?

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