

## Проект мобилизации данных регионального гербария

Материалы для практических занятий на обучающем семинаре  
10 октября 2020 г., Уральский федеральный университет



Места сбора гербарных образцов на Территории Таджикистана.

*Материалы упражнений основаны на реальных данных, измененных в учебных целях. Концепция упражнений разработана для курсов повышения квалификации проекта BioDATA, справочная информация взята из [пособия](#) Буйволов и др., 2019.*

## Описание ситуации

Университет уездного города N является известным высшим учебным заведением России и национальным центром исследований биоразнообразия. Факультет Биологии растений имеет гербарий, содержащий около 1 000 000 образцов, включая сборы с территории Таджикистана (около 6000 листов).

Декан факультета Биологии растений недавно получил финансирование на два года для сканирования образцов ботанических коллекций Средней Азии и их публикации в сети Интернет. Участники проекта хотят использовать эту возможность для налаживания непрерывного процесса оцифровки и разработки методики публикации данных. Это повысит доступность коллекций Гербария и позволит привлечь дополнительное финансирование.

## Упражнение 1: Сбор данных о биоразнообразии

Представьте, что вам поручено перенести информацию с этикеток гербарных листов в таблицы.

1. Распакуйте файлы из архива 'USE CASE 1 TJ- Exercise 1 Base Material.zip'. Этот базовый материал содержит пять гербарных образцов, которые вы должны обработать. Всего в архиве 5 изображений, по 1 изображению на вид.
2. Используйте файл 'TemplateHerbarium.xlsx' для переноса информации, содержащейся на каждом изображении для каждого из пяти видов.

**Справочная информация.** Для представления данных об образцах биологических коллекций (Occurrence data) стандартом Darwin Core предусмотрено 4 обязательных для заполнения поля (см. таблицу ниже). Для каждого термина приводится русскоязычное описание и ссылка на оригинальную спецификацию на английском языке.

occurrenceID	Уникальный идентификатор находки. Вы можете составить его самостоятельно, или сгенерировать глобальный идентификатор GUID. Желательно при составлении ID не ограничиваться сплошной нумерацией записей (1, 2, 3, 4, 5 и т.д.), а включить в идентификатор акроним коллекции или аббревиатуру ООПТ, каталожный номер образца или другую подобную информацию. <a href="https://dwc.tdwg.org/terms/#dwc:occurrenceID">https://dwc.tdwg.org/terms/#dwc:occurrenceID</a>
basisOfRecord	В этом поле необходимо указать, что послужило основанием для появления этой записи. Для его заполнения предусмотрен фиксированный набор значений, необходимо использовать одно из них - PreservedSpecimen, FossilSpecimen, LivingSpecimen, MaterialSample, Event, HumanObservation, MachineObservation, Taxon, Occurrence. В случае коллекций необходимо выбирать значение <b>PreservedSpecimen</b> , при этом для всех терминов и значений важно сохранить оригинальное (слитное) написание. <a href="https://dwc.tdwg.org/terms/#dwc:basisOfRecord">https://dwc.tdwg.org/terms/#dwc:basisOfRecord</a>
scientificName	Научное название таксона. Заполняется по тем же правилам, что и для таксономических данных. Настоятельно рекомендуется проверить все названия с помощью инструмента Species matching. Для обучения работе с этим инструментом предусмотрено <u>отдельное упражнение во второй половине дня, выполняя это упражнение, заносите в поле scientificName названия, указанные на этикетках</u> . <a href="https://dwc.tdwg.org/terms/#dwc:scientificName">https://dwc.tdwg.org/terms/#dwc:scientificName</a>
eventDate	Дата сбора образца. Должна быть представлена в формате ГГГГ-ММ-ДД, т.е. дата 27 июня 2019 года в DwC будет иметь вид 2019-06-27. <a href="https://dwc.tdwg.org/terms/#dwc:eventDate">https://dwc.tdwg.org/terms/#dwc:eventDate</a>

Для переноса данных с этикеток в таблицу, также могут потребоваться следующие поля:

institutionCode	Сокращенное название организации, в которой хранится коллекция <a href="http://rs.tdwg.org/dwc/terms/institutionCode">http://rs.tdwg.org/dwc/terms/institutionCode</a>
collectionCode	Акроним коллекции <a href="http://rs.tdwg.org/dwc/terms/collectionCode">http://rs.tdwg.org/dwc/terms/collectionCode</a>
recordedBy	В этом поле на первом месте указывается коллектор, а далее все лица, имеющие отношение к данной находке. Имена следует отделять вертикальной чертой, например Oleg Borodin   Alexey Petrovich Seregin <a href="http://rs.tdwg.org/dwc/terms/recordedBy">http://rs.tdwg.org/dwc/terms/recordedBy</a>

identifiedBy	В этом поле приводится список лиц, определивших образец. Правила заполнения аналогичны предыдущему полю. <a href="http://rs.tdwg.org/dwc/terms/identifiedBy">http://rs.tdwg.org/dwc/terms/identifiedBy</a>
taxonRank	Ранг таксона, до которого удалось определить образец (вид, род, семейство и др.). Это поле необходимо для корректной индексации набора данных на глобальном портале и сопоставления ваших названий с таксономией GBIF. Для заполнения этого поля необходимо использовать линневские ранги: царство, тип, класс, порядок, семейство, род, вид. Допускается написание рангов на латинском или английском языке. <a href="https://dwc.tdwg.org/terms/#dwc:taxonRank">https://dwc.tdwg.org/terms/#dwc:taxonRank</a>
kingdom	Царство. Его необходимо указывать для однозначного сопоставления ваших названий с таксономией GBIF. Как известно, существуют виды с одинаковыми названиями и относящиеся к разным таксономическим группам. Кроме того, в названиях ваших таксонов возможны опечатки, или другие неточности. Для заполнения этого поля необходимо указывать линневские ранги ("Animalia"), нельзя указывать таксономические группы ("Algae" или "Herbivora") <a href="https://dwc.tdwg.org/terms/#dwc:kingdom">https://dwc.tdwg.org/terms/#dwc:kingdom</a>
countryCode	Код страны согласно стандарту ISO 3166-1-alpha-2. Его необходимо приводить для однозначного указания страны. Код России RU. <a href="http://rs.tdwg.org/dwc/terms/countryCode">http://rs.tdwg.org/dwc/terms/countryCode</a>
stateProvince	Следующая после страны административно-территориальная единица. Как правило, в этом поле указывают название области, края или республики. <a href="http://rs.tdwg.org/dwc/terms/stateProvince">http://rs.tdwg.org/dwc/terms/stateProvince</a>
county	Следующая после указанной в предыдущем поле административно-территориальная единица. Как правило, в этом поле указывают название района. <a href="http://rs.tdwg.org/dwc/terms/county">http://rs.tdwg.org/dwc/terms/county</a>
locality	Место сбора. Обычно указывается расстояние и направление до населенного пункта. Информацию в этом поле лучше приводить на английском языке. Важно сохранить и русскоязычное описание места сбора, эту информацию следует поместить в поле verbatimLocality. <a href="http://rs.tdwg.org/dwc/terms/locality">http://rs.tdwg.org/dwc/terms/locality</a>
verbatimLocality	Описание места сбора в соответствии с исходными данными. Приводится на языке оригинала в точном соответствии этикетке. <a href="http://rs.tdwg.org/dwc/terms/verbatimLocality">http://rs.tdwg.org/dwc/terms/verbatimLocality</a>
decimalLatitude	Географическая широта в десятичных градусах. Для северного полушария значения широты положительные, для южного - отрицательные. <a href="http://rs.tdwg.org/dwc/terms/decimalLatitude">http://rs.tdwg.org/dwc/terms/decimalLatitude</a>
decimalLongitude	Географическая долгота в десятичных градусах. Для восточного полушария значения долготы положительные, для западного - отрицательные. <a href="http://rs.tdwg.org/dwc/terms/decimalLongitude">http://rs.tdwg.org/dwc/terms/decimalLongitude</a>
geodeticDatum	Эллипсоид, геодезическая система координат или пространственная система отсчета (SRS), на которой основаны географические координаты. <a href="http://rs.tdwg.org/dwc/terms/geodeticDatum">http://rs.tdwg.org/dwc/terms/geodeticDatum</a>
coordinatePrecision	Точность определения координат, выраженная в долях градуса. Например, показания GPS навигатора 54.75375, 35.36574, их точность 0.00001; точность координат 54, 35 составляет 1.0 <a href="http://rs.tdwg.org/dwc/terms/coordinatePrecision">http://rs.tdwg.org/dwc/terms/coordinatePrecision</a>

coordinateUncertaintyInMeters	Оценка погрешности определения координат, выраженная в метрах. Этот термин очень важен, т.к. координаты могут быть определены с высокой точностью (5 знаков после запятой), но реальное место сбора не всегда будет соответствовать этим координатам. Например, при использовании обычного туристического GPS приемника погрешность определения координат, как правило, составляет 3-5 метров. При определении координат по топографическим картам общедоступных масштабов 1:200 000 или 1:500 000, погрешность привязки, в лучшем случае, составляет сотни метров. Если значение погрешности неизвестно, или его невозможно оценить (например, когда координаты не указаны), поле следует оставить пустым. “0” не является допустимым значением. <a href="http://rs.tdwg.org/dwc/terms/coordinateUncertaintyInMeters">http://rs.tdwg.org/dwc/terms/coordinateUncertaintyInMeters</a>
individualCount	Число обнаруженных особей, выраженное в штуках. Например, <b>37</b> . <a href="http://rs.tdwg.org/dwc/terms/individualCount">http://rs.tdwg.org/dwc/terms/individualCount</a>
organismQuantity	Этот термин используется, если для оценки обилия выбрано не число особей, а другая величина. В этом случае значение обилия помещают в этом поле. <a href="http://rs.tdwg.org/dwc/terms/organismQuantity">http://rs.tdwg.org/dwc/terms/organismQuantity</a>
organismQuantityType	Размерность величины, указанной в предыдущем поле. Например, <i>r</i> в organismQuantity и <b>Braun Blanquet Scale</b> в organismQuantityType. <a href="http://rs.tdwg.org/dwc/terms/organismQuantityType">http://rs.tdwg.org/dwc/terms/organismQuantityType</a>
year	Год сбора образца <a href="http://rs.tdwg.org/dwc/terms/year">http://rs.tdwg.org/dwc/terms/year</a>
month	Месяц сбора образца <a href="http://rs.tdwg.org/dwc/terms/month">http://rs.tdwg.org/dwc/terms/month</a>
day	День сбора образца <a href="http://rs.tdwg.org/dwc/terms/day">http://rs.tdwg.org/dwc/terms/day</a>

## Упражнение 2. Качество данных о биоразнообразии. Поиск и исправление ошибок в данных.

Ваш институт участвует в проекте “*Цифровая флора Таджикистана*”. Руководитель проекта получил финансирование на публикацию в открытом доступе обновленной информации о флоре и предложил вашему Гербарию участвовать в этом проекте. Вам требуется предоставить высококачественные данные об имеющихся у вас образцах из Таджикистана. В вашей гербарии хранится около 6000 листов из этой страны, поэтому вы думаете, что могли бы внести существенный вклад в проект.

В этом небольшом упражнении мы сфокусируемся на поиске и исправлении ошибок средствами EXCEL и выполним базовую проверку данных на наличие в них технических ошибок и ошибок согласованности.

1. Используйте файл ‘**Data Cleaning\_DATA EXAMPLE TJ.csv**’.
2. Откройте CSV-файл в EXCEL.
3. С помощью инструмента *фильтр* найдите и исправьте ошибки (см. таблицу ниже) вручную.
4. Используйте таблицу для документирования ваших действий.

Обратите внимание, что названия столбцов в таблице не всегда соответствуют терминам DwC.



### Предлагаемые действия по поиску и исправлению ошибок в данных

Столбец	Подсказка	Ошибка	Какое исправление внесено
<b>Country col.</b>	Проверьте наличие пустых значений		
<b>YE</b>	Данные собраны с 1931 по 2015 годы		
<b>countryCode</b>	Код страны должен включать только буквы		
<b>MO</b>	JAN - 1, FEB - 2 и т.д.		
<b>taxonRank</b>	Проверьте наличие пустых значений		
<b>Elevation</b>	Пик Сомони 7495 м (самая высокая точка Таджикистана)		
<b>phylum</b>	Не появился ли новый Отдел растений?		
<b>coordinate Uncertainty</b>	Значение следует указывать в метрах		
<b>lat/lon</b>	Все ли координаты согласованы и представлены в десятичных градусах?		
<b>eventDate</b>	Существует ли такой столбец?		

Если ваши координаты представлены в формате ГГ ММ СС, их можно легко пересчитать в десятичные градусы, используя формулу:

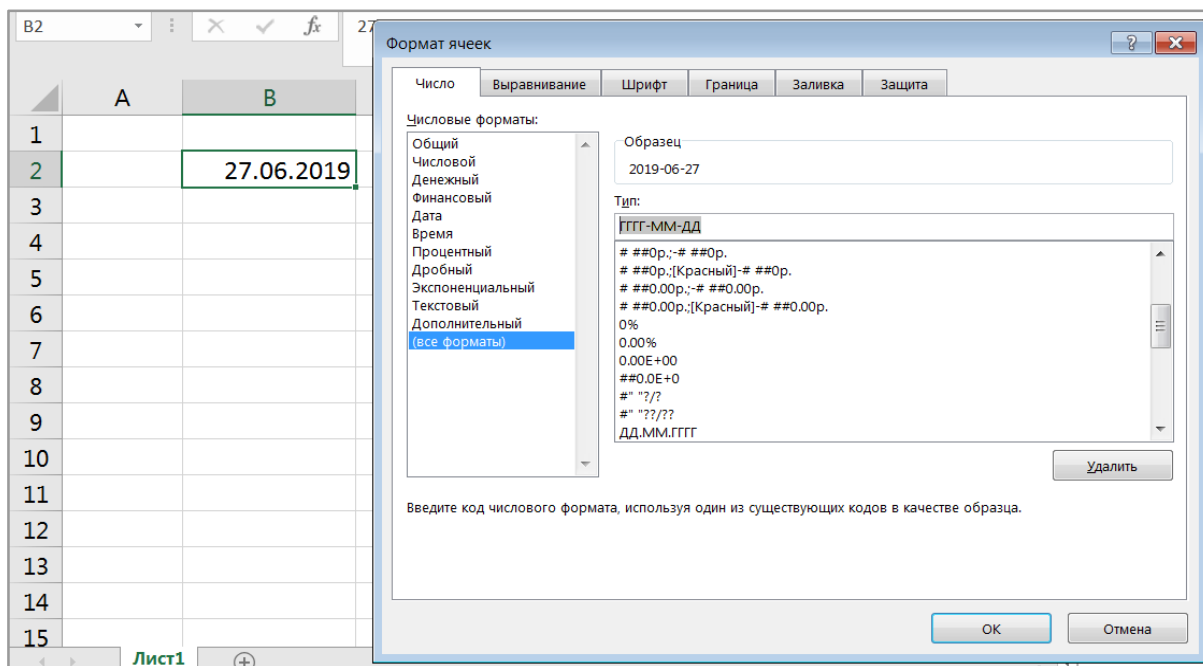
$$\text{ГГ.ГГГГГ} = \text{ГГ} + \text{ММ}/60 + \text{СС}/3600$$

Пересчет можно выполнить, используя встроенные формулы в электронных таблицах (рис. ниже), но важно понимать, что в этом случае значения должны храниться в отдельных полях (столбцах), без знаков .°, ' и ".

G2							=A2+B2/600+C2/3600		
	A	B	C	D	E	F	G	H	I
1	DD_lat	MM_lat	SS_lat	DD_lon	MM_lon	SS_lon	DD.DDDDD_lat	DD.DDDDD_lon	
2	50	40	45	40	50	30.3	50.07917	40.84175	
3									
4									
5									

Если в исходных данных координаты хранятся в ячейках целиком, для конвертации их в десятичные градусы удобнее воспользоваться веб-инструментом национального портала о биоразнообразии Канады Canadensys <http://data.canadensys.net/tools/coordinates?lang=fr>. Для расчета используется та же формула, что и в предыдущем примере, но входные данные могут содержать знаки °, ' и ''.

Чтобы привести даты к требуемому формату, необходимо в таблице Excel выделить соответствующую ячейку (или весь столбец), кликнув правой кнопкой мыши, выбрать пункт меню Формат ячеек -> Все форматы, задать необходимый формат даты вручную, нажать ОК (рис. ниже).



### Упражнение 3. Использование веб-инструмента Species matching для проверки таксономических списков.

Во время поиска и исправления технических и географических ошибок выяснилось, что не все названия на этикетках указаны корректно и соответствуют современной номенклатуре. В этом упражнении мы выполним проверку соответствия таксономических списков таксономическому справочнику GBIF Backbone. Для этого будем использовать инструмент Species matching, который доступен на портале gbif.org в разделе Tools (<https://www.gbif.org/tools/species-lookup>).

Используйте файл **ChecklistHerbariumTJ.csv**.

1. Откройте файл и ознакомьтесь с его структурой. Species matching работает только с файлами CSV. Название столбца с таксономическим списком должно быть scientificName.
2. Выполните проверку таксономии при помощи Species matching, при необходимости исправьте названия таксонов онлайн, загрузите результаты в виде CSV файла на свой компьютер.

Файл CSV нужно загрузить через диалоговое окно инструмента Species matching - выбрать его из папки (SELECT FILE), или перетащить (DROP HERE). В открывшемся окне вы увидите список ваших таксонов. Если возможно, укажите Царство, к которому они относятся (кликнув соответствующую картинку), и нажмите кнопку MATCH TO GBIF BACKBONE.

[Get data](#)
[Share](#)
[Tools](#)
[Inside GBIF](#)

TOOLS | LOOK UP

If no kingdom specified then prefer

animalia
 plantae
 fungi
 chromista
 bacteria
 protozoa
 viruses
 archaea

MATCH TO GBIF BACKBONE

scientificName	preferred kingdom
<i>Ablabesmyia longistyla</i> Fittkau 1962	
<i>Ablabesmyia monillis</i> (Linnaeus, 1758)	
<i>Ablabesmyia phatta</i> (Egger, 1863)	
<i>Ablabesmyia</i> sp.	
<i>Abra segmentum</i> (R?cluz, 1843)	

Результат сопоставления доступен для просмотра в браузере (рис. ниже), также его можно сохранить в CSV файл, содержащий следующие поля:

***occurrenceId*** - идентификатор. Если вы его не указывали, поле будет пустым.

**verbatimScientificName** - предоставленные вами названия таксонов.

**scientificName** - название в соответствии с таксономической системой GBIF

**key** - идентификатор таксона в GBIF. При добавлении значения key к URL <https://www.gbif.org/species/> вы получите доступ к странице соответствующего таксона. Например, для Bufo bufo (Linnaeus, 1758) key=5217160, следовательно URL для страницы этого вида на портале GBIF <https://www.gbif.org/species/5217160>.

***matchType*** - результат сопоставления. В этом поле возможны три значения: EXACT - название полностью соответствует GBIF Backbone, FUZZY - т.н. подозрительные, название не полностью соответствует GBIF Backbone, есть расхождения в нескольких буквах, возможно вследствие опечатки в исходных данных, HIGHERRANK - таксон не найден в GBIF Backbone, но есть родительский таксон более высокого ранга, например, вид отсутствует в GBIF, но есть род.

**confidence** - количественная характеристика степени совпадения

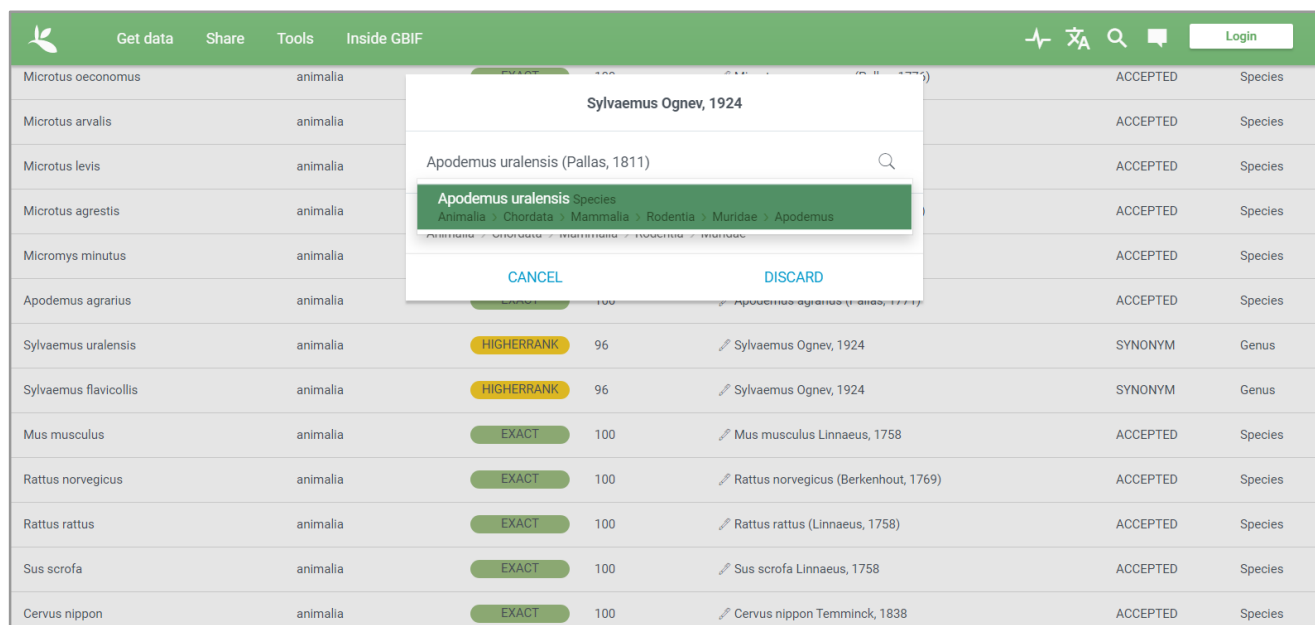
**status** - статус таксона в GBIF Backbone. Возможные значения: ACCEPTED - принятое в GBIF Backbone название, SYNONYM - синоним, DOUBTFUL - сомнительное.

**rank** - ранг таксона в GBIF Backbone на английском языке

Далее представлено положение таксона в иерархическом порядке, начиная с царства.

Tools   Look up							
verbatimScientificName	preferredKingdom	matchType	confidence	scientificName <small>(editable)</small>	status	rank	kingdom
Ablabesmyia longistyla Fittkau 1962	animalia	EXACT	100	<a href="#">Ablabesmyia longistyla Fittkau, 1962</a>	ACCEPTED	Species	<a href="#">Animalia</a>
Ablabesmyia monilis (Linnaeus, 1758)	animalia	EXACT	100	<a href="#">Ablabesmyia monilis (Linnaeus, 1758)</a>	ACCEPTED	Species	<a href="#">Animalia</a>
Ablabesmyia phatta (Egger, 1863)	animalia	EXACT	100	<a href="#">Ablabesmyia phatta (Egger, 1863)</a>	ACCEPTED	Species	<a href="#">Animalia</a>
Ablabesmyia sp.	animalia	EXACT	85	<a href="#">Ablabesmyia Johannsen, 1905</a>	ACCEPTED	Genus	<a href="#">Animalia</a>
Abra segmentum (R�cluz, 1843)	animalia	EXACT	100	<a href="#">Abra segmentum (R�cluz, 1843)</a>	ACCEPTED	Species	<a href="#">Animalia</a>
Acarina sp.	animalia	EXACT	84	<a href="#">Acarina Baraud, 1965</a>	SYNONYM	Genus	<a href="#">Animalia</a>
Acilius sp.	animalia	EXACT	80	<a href="#">Acilius Leach, 1817</a>	ACCEPTED	Genus	<a href="#">Animalia</a>
Acricotopus lucens (Zetterstedt, 1850)	animalia	EXACT	100	<a href="#">Acricotopus lucens (Zetterstedt, 1850)</a>	ACCEPTED	Species	<a href="#">Animalia</a>
Acricotopus sp.	animalia	EXACT	85	<a href="#">Acricotopus Kieffer, 1921</a>	ACCEPTED	Genus	<a href="#">Animalia</a>

Названия таксонов можно отредактировать в выгружаемом файле CSV, или непосредственно в браузере. Для редактирования онлайн нужно нажать на значок редактирования (“карандаш”) и в открывшемся диалоговом окне ввести корректное название (рис.), или выбрать название, предлагаемое автоматически. После нажатия клавиши ENTER изменения будут сохранены.



Species	Kingdom	Phylum	Class	Order	Family	Genus	Species	Status	Rank
Microtus oeconomus	animalia							ACCEPTED	Species
Microtus arvalis	animalia							ACCEPTED	Species
Microtus levis	animalia							ACCEPTED	Species
Microtus agrestis	animalia							ACCEPTED	Species
Micromys minutus	animalia							ACCEPTED	Species
Apodemus agrarius	animalia							ACCEPTED	Species
Sylviaemus uralensis	animalia							SYNONYM	Genus
Sylviaemus flavicollis	animalia							SYNONYM	Genus
Mus musculus	animalia							ACCEPTED	Species
Rattus norvegicus	animalia							ACCEPTED	Species
Rattus rattus	animalia							ACCEPTED	Species
Sus scrofa	animalia							ACCEPTED	Species
Cervus nippon	animalia							ACCEPTED	Species

### Упражнение 3. Поиск и исправление ошибок в данных с помощью OpenRefine.

В этом упражнении мы используем OpenRefine для улучшения качества набора данных, используя стандартные функции, существующие веб-сервисы и регулярные выражения.

1. Выполните упражнения, представленные в соответствующем документе OpenRefine (‘**Data Cleaning OpenRefine - Exercise\_TJ RU.pdf**’).
2. Используйте файл ‘**Data Cleaning OpenRefine\_DATA EXAMPLE\_TJ.csv**’.

Материалы разработаны для практических занятий на практическом семинаре 10 октября 2020 г., в Уральском федеральном университете. Упражнения основаны на реальных данных, измененных в учебных целях.

Содержание: Наталья Иванова, Максим Шашков. Концепция: Alberto González-Talaván, Néstor Beltrán, Nicolas Noé, Sharon Grant. CC-BY.

Справочные материалы из пособия Буйволов Ю.А., Иванова Н.В., Шашков М.П. Оцифровка данных Летописей природы и научных биологических коллекций особо охраняемых природных территорий. Учебное пособие. ФГБУ Приокско-Террасный государственный природный биосферный заповедник, 2019.