

Разнообразие почвенных животных России: публикация и эффективное использование исходных данных

Материалы для практических занятий на курсе повышения квалификации
29–30 августа 2019 г., ИПЭЭ им. А.Н. Северцова РАН



Cicindela hybrida фото: Наталья Иванова, CC-BY
<https://www.inaturalist.org/observations/6826234>

Материалы упражнений основаны на реальных данных, измененных в учебных целях. Концепция позаимствована из курсов повышения квалификации по проекту BioDATA.

Упражнение 1. Стандартизация данных о находках видов – пример из литературы

В Приокско-Тerrasном заповеднике исследовали видовое разнообразие комплекса членистоногих (Arthropoda). Сборы проводили в период с 30 августа по 18 сентября 2014 г. Полевой материал собирали в девяти биотопах с различным гидрорежимом, географические координаты точек отбора представлены в табл. 1.1. Для сбора материала в каждом биотопе устанавливали ловчие банки с фиксирующей жидкостью.

Все учтенные особи были определены до отряда, в некоторых случаях – до семейства, рода или вида. В табл. 1.2 частично представлены полученные результаты. Подробный анализ всех собранных данных опубликован в сборнике трудов заповедника: Сейфулина Р.Р. Комплекс членистоногих (Arthropoda) Приокско-Тerrasного заповедника // Труды Приокско-Тerrasного заповедника, вып. 6. Тула: Аквариус, 2015. С. 178–203.

Таблица 1.1 Географические координаты исследованных биотопов

Биотопы	Северная широта	Восточная долгота
Широколиственный лес	54°51.274′	37°40.242′
Мелколиственный лес	54°64.416′	37°34.350′
Смешанный лес	54°53.807′	37°34.355′
Сосновый бор	54°51.433′	37°36.215′
Берег оз. Протского	54°51.541′	37°35.712′
Сфагновое болото	54°54.846′	37°36.129′
Надпойменный луг	54°51.287′	37°39.125′
Суходольный луг	54°53.668′	37°38.623′
Дол (остепненный участок)	54°51.327′	37°36.260′

Таблица 1.2 Суммарная численность и биомасса учтенных членистоногих

Биотоп	Таксон	Число экз.	Биомасса, г
Широколиственный лес	<i>Myrmica rubra</i>	154	0.7
Сосновый бор	<i>Carabus nemoralis</i>	79	43.8
Сфагновое болото	<i>Trachelipus rathkii</i>	34	1.2
Мелколиственный лес	<i>Trachelipus rathkii</i>	18	0.7
Суходольный луг	<i>Carabus nemoralis</i>	4	1.9
Надпойменный луг	<i>Carabus cansellatus</i>	19	6.02
Дол	<i>Onthopagus sp.</i>	1	0.09
Смешанный лес	<i>Geotrupes stercorarius</i>	315	207.9

Представьте, что вы участвуете в проекте по мобилизации данных о распространении беспозвоночных на юге Московской области. Используя шаблон в файле **TemplateOccurrence.xlsx**, таблицы 1.1 и 1.2 и описание ситуации, представьте данные о находках беспозвоночных стандарте Darwin Core. При необходимости воспользуйтесь подсказками в таблице 1.3. Какие поля DwC вы могли бы добавить для более полного представления информации?

Таблица 1.3 Подсказки для выполнения упражнения 1

Обязательные для заполнения поля	occurrenceID, basisOfRecord, scientificName, eventDate
Что такое basisOfRecord?	Это то, что послужило основанием для появления этой записи. Для заполнения поля предусмотрен фиксированный набор значений: PreservedSpecimen, FossilSpecimen, LivingSpecimen, MaterialSample, Event, HumanObservation или MachineObservation.
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. 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. Временной период задается следующим образом: 2019-06-27/07-01 (27 июня – 1 июля 2019 года)
countryCode	Код страны согласно стандарту ISO 3166-1-alpha-2. Его необходимо приводить для однозначного указания страны. Код России RU.

Географические координаты	Координаты должны быть представлены в десятичных градусах (ГГ.ГГГГГ). Если ваши координаты представлены в формате ГГ ММ СС, их можно легко пересчитать в десятичные градуса, используя формулу: $\text{ГГ.ГГГГГ} = \text{ГГ} + \text{ММ}/60 + \text{СС}/3600$ Также можно воспользоваться готовым веб-инструментом национального портала о биоразнообразии Канады Canadensys http://data.canadensys.net/tools/coordinates?lang=en .
Спецификация Darwin Core	Оригинальная (на английском) http://rs.tdwg.org/dwc/terms/ Основные термины на русском http://gbif.ru/DwC_spec

Упражнение 2. Sampling events – простой способ представления данных, собранных методом отбора почвенных проб

В окрестностях г. Пущино, расположенного на юге Московской области, изучали разнообразие дождевых червей на бывших сельскохозяйственных землях. Было исследовано три биотопа: луг (залежь), молодой березняк на месте пашни и старый смешанный лес с участием широколиственных деревьев. Исследования проводили методом разбора почвенных проб. В каждом биотопе было разобрано по 8 почвенных монолитов размером 25×25×25 см. Червей фиксировали в 4%-м растворе формалина и определяли в лабораторных условиях. Собранные данные представлены в файле **SamplingEventDataExample.xlsx**. В таблице приводятся значения суммарной численности и биомассы для каждого вида в каждом сборе.

Рассмотрите таблицу SamplingEventDataExample.xlsx и определите, какие поля относятся к описанию сборов, а какие - к описанию находок видов в сборах. Заполните таблицу 2.1

Таблица 2.1

Поля для характеристики сборов	Поля для характеристики видов в сборах

Используя шаблон в файле **TemplateSamplingEvent.xlsx** и таблицу 2.2, представьте эти данные в стандарте Darwin Core как отдельные сборы. Вся ли необходимая информация представлена в таблице с исходными данными?

Таблица 2.2 Подсказки для выполнения упражнения 2

Обязательные для заполнения поля	Для описания сборов (Sampling Events): eventID, eventDate, samplingProtocol
Обязательные для заполнения поля	Для описания находок в сборах (Associated Occurrences): eventID, occurrenceID, basisOfRecord, eventDate, scientificName

Упражнение 2а. Sampling events – продвинутый способ представления данных, собранных методом отбора почвенных проб (домашнее задание для желающих)

Возможности Darwin Core позволяют публиковать не только данные по сборам в целом, но и представить каждую отдельную пробу внутри сбора. В этом случае в данных появляется еще один иерархический уровень

СБОР
! Почвенный монолит (проба) внутри сбора
Находки видов в определенном монолите в определенном сборе

Для представления данных с такой структурой нужен дополнительный идентификатор parentEventID, позволяющий показать принадлежность находки не только к сбору, но и одному из монолитов (проб) внутри сбора. Важно не ошибиться в присвоении идентификаторов parentEventID, eventID и occurrenceID. В предыдущем упражнении использована следующая схема присвоения ID:

СБОР	eventID	
Находки видов в определенном монолите в определенном сборе	eventID	occurrenceID

При добавлении отдельных монолитов (проб) внутри сборов структура данных усложняется и порядок идентификаторов будет следующим:

СБОР	parentEventID		
Почвенный монолит внутри сбора	parentEventID	eventID	
Находки видов в определенном монолите в определенном сборе	parentEventID	eventID	occurrenceID

Для упражнения в стандартизации таких данных вы можете использовать файл **SamplingEventDataExample-homework.xlsx**, в котором представлены расширенные данные из упражнения 2. Используйте шаблон **TemplateSamplingEvent-homework.xlsx**. Вы можете прислать заполненные таблицы на проверку преподавателям курса gbif.ru@yandex.ru.

Упражнение 3. Проверка и верификация данных

В окрестностях Полистовского заповедника в июне 2010 г. проводили исследования разнообразия панцирных клещей. Данные были перенесены из бумажных носителей в электронную таблицу *DataCleaningExample.xlsx*. Используя контрольный список для проверки и таблицу 3.1 найдите и исправьте ошибки в данных. Задokumentируйте свои действия в таблице.

Контрольный список для проверки данных:

- Все названия таксонов написаны правильно
- Ранги всех таксонов указаны корректно
- Все записи имеют координаты
- Все точки находок находятся в районе Полистовского заповедника
- Все координаты преобразованы в десятичные градусы
- Все даты находятся в соответствующем столбце и представлены в формате ГГГГ-ММ-ДД

Таблица 3.1

Предлагаемые действия по поиску и исправлению ошибок в данных				
Столбец	Подсказка	Ошибка	Какое исправление сделано	Кто внес исправление
eventDate	Проверьте наличие пустых значений			
YE	Данные собраны в 2010-м году			
countryCode	Проверьте отсутствие пустых значений			
decimalLatitude	Все ли точки сборов находятся в северном полушарии?			
decimalLongitude	Все ли координаты указаны в десятичных градусах?			
coordinateUncertainty	Значение следует указывать в метрах			
taxonRank	Для всех ли таксонов ранги указаны корректно?			
species	Все ли названия соответствуют базовой таксономии GBIF? Используйте онлайн инструмент https://www.gbif.org/tools/species-lookup			

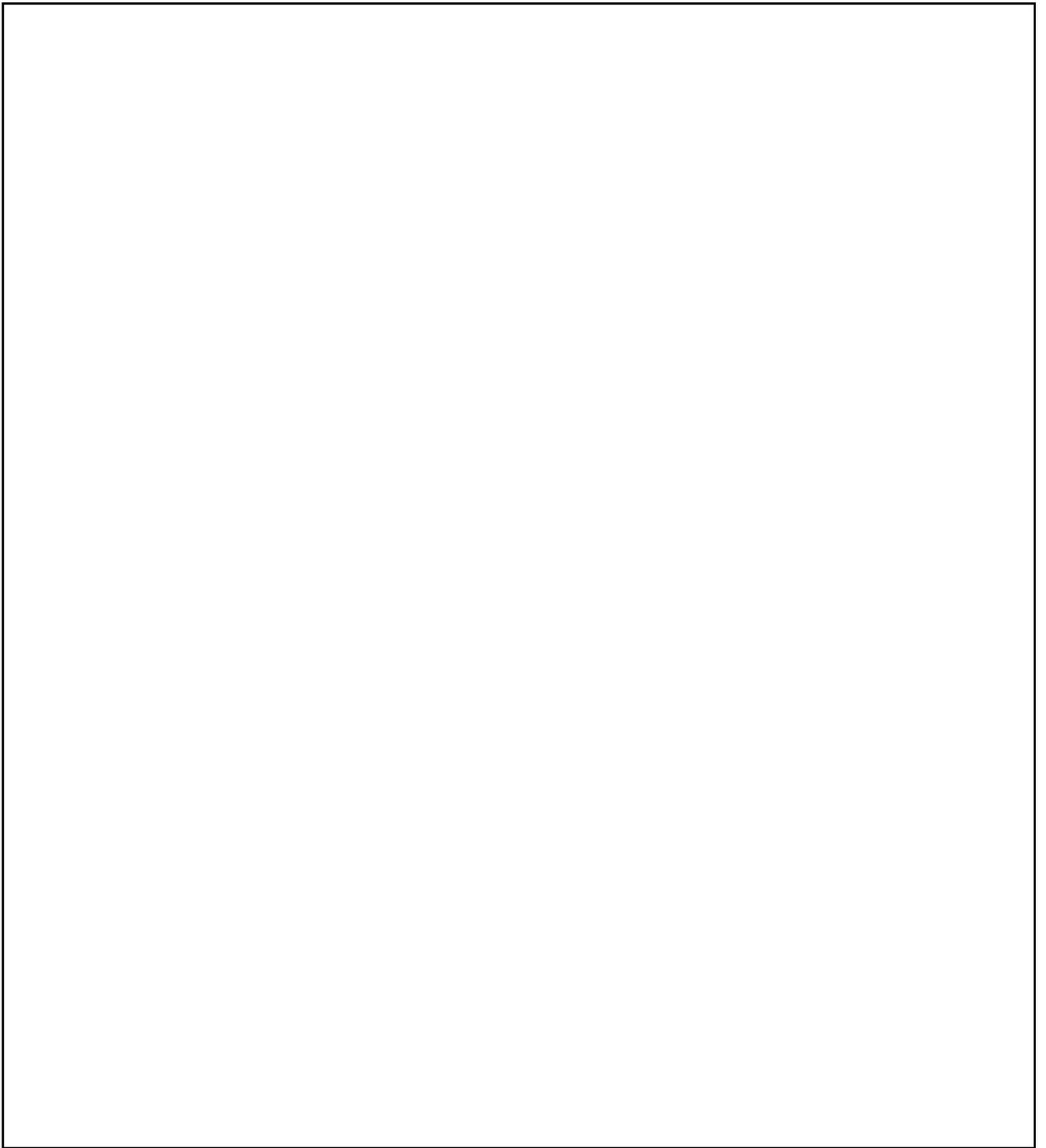
Упражнение 4. Публикация данных через IPT

После проверки и верификации данные о разнообразии панцирных клещей, собранные в окрестностях Полистовского заповедника, было решено опубликовать в GBIF.

Используйте файл **DataForPublishing.csv**. Загрузите файл в IPT, сопоставьте названия столбцов загруженной таблицы с терминами стандарта Darwin Core, добавьте необходимые метаданные, опубликуйте и зарегистрируйте набор данных. Опишите ваши действия в таблице 3.1

Таблица 3.1

Какое «ядро» (Core) DwC вы использовали для публикации? Почему вы выбрали именно его?
Опишите ваши действия при работе с IPT



Материалы разработаны для практических занятий на курсе повышения квалификации “Разнообразие почвенных животных России: публикация и эффективное использование исходных данных”, который состоялся 29–30 августа 2019 г., в ИПЭЭ им. А.Н. Северцова РАН.

Упражнения основаны на реальных данных, измененных в учебных целях.

Авторы: Наталья Иванова, Максим Шашков. CC-BY