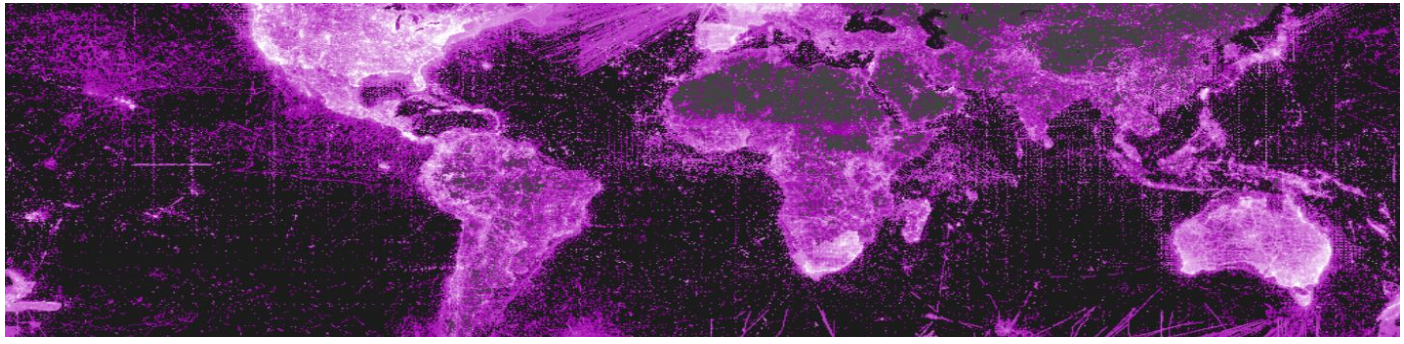
Разнообразие почвенных животных России: публикация и эффективное использование исходных данных
29–30 августа 2019 г., ИПЭЭ им. А.Н. Северцова РАН

Сессия 5.

Стандарты данных: Darwin Code

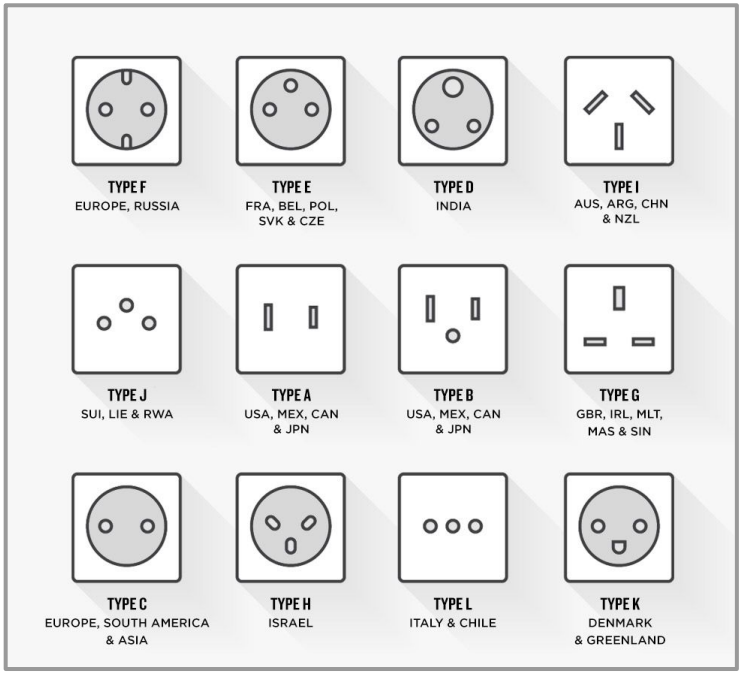
Шашков Максим

Институт математических проблем биологии РАН – филиал ИПМ им. М.В. Келдыша РАН
г. Пущино Московской области

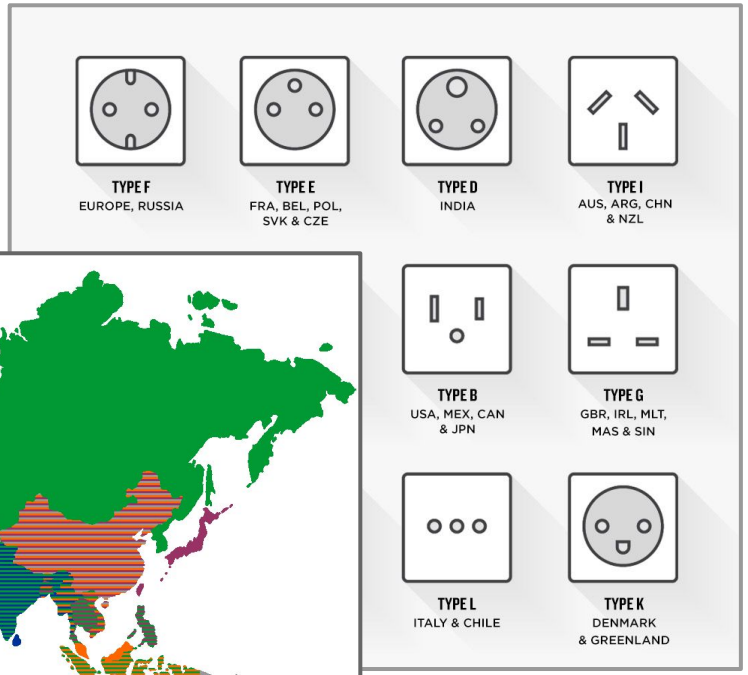


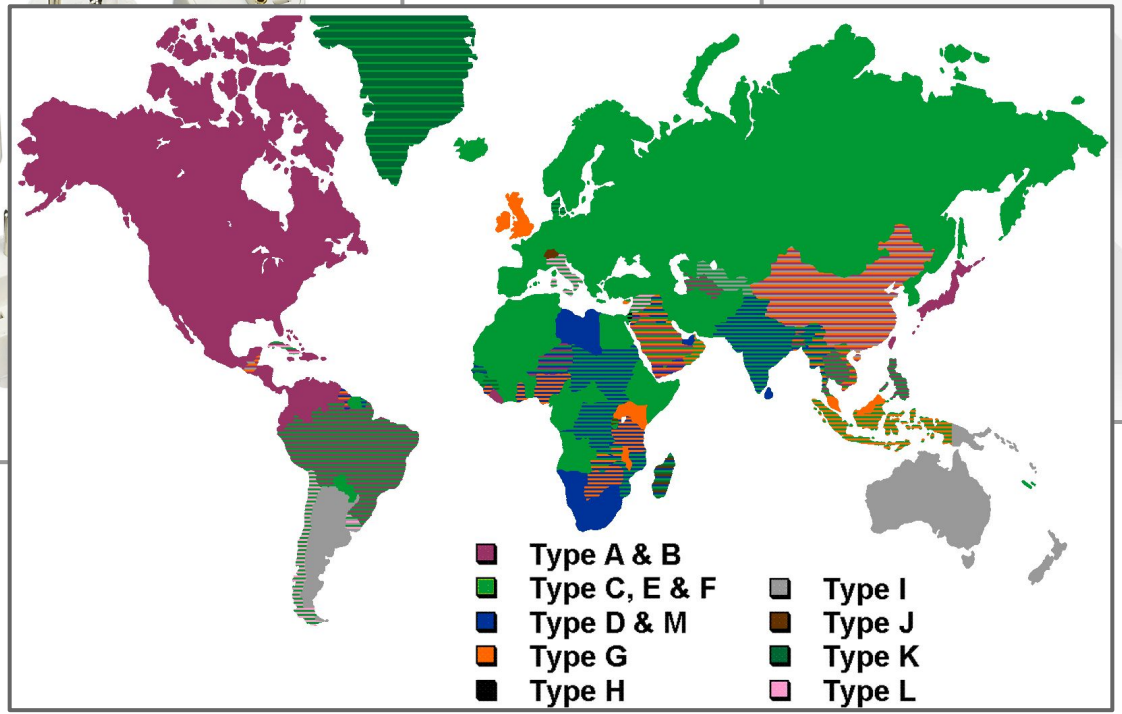
Слайды CC BY:
Paula Zermoglio,
John Wieczorek и
Максим Шашков

Type A
Type B
Type C
Type D
Type E/F
Type G
Type H
Type I
Type J
Type L
Type N


TYPE F
EUROPE, RUSSIA
TYPE E
FRA, BEL, POL, SVK & CZE
TYPE D
INDIA
TYPE I
AUS, ARG, CHN & NZL
TYPE J
SUI, LIE & RWA
TYPE A
USA, MEX, CAN & JPN
TYPE B
USA, MEX, CAN & JPN
TYPE G
GBR, IRL, MLT, MAS & SIN
TYPE C
EUROPE, SOUTH AMERICA & ASIA
TYPE H
ISRAEL
TYPE L
ITALY & CHILE
TYPE K
DENMARK & GREENLAND

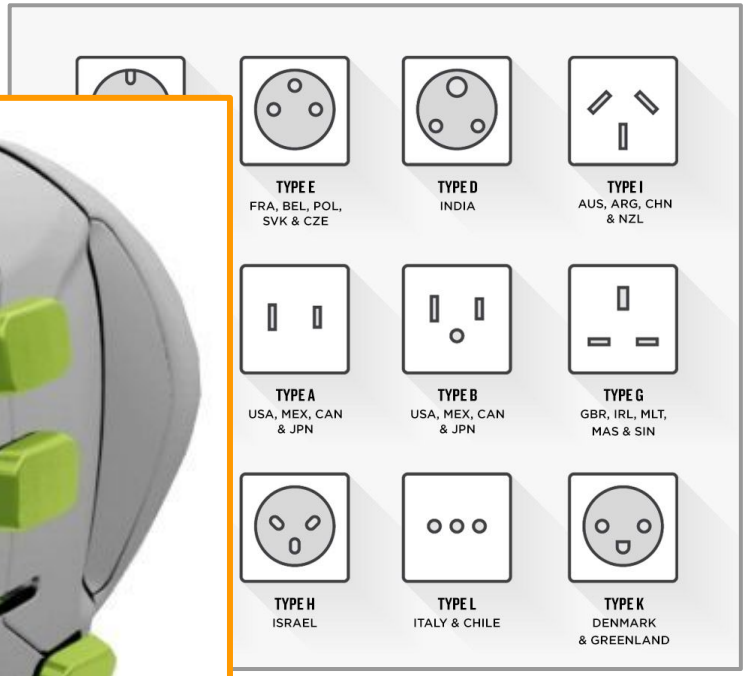
Type A
Type B
Type C
Type D
Type E/F
Type G
Type J


TYPE F
EUROPE, RUSSIA
TYPE E
FRA, BEL, POL, SVK & CZE
TYPE D
INDIA
TYPE I
AUS, ARG, CHN & NZL
TYPE B
USA, MEX, CAN & JPN
TYPE G
GBR, IRL, MLT, MAS & SIN
TYPE L
ITALY & CHILE
TYPE K
DENMARK & GREENLAND


Type A & B
Type C, E & F
Type D & M
Type G
Type H
Type I
Type J
Type K
Type L

Type A
Type B
Type C
Type E/F
Type G
Type L
Type J

L
N
TÜV
CE


TYPE E
FRA, BEL, POL, SVK & CZE
TYPE D
INDIA
TYPE I
AUS, ARG, CHN & NZL
TYPE A
USA, MEX, CAN & JPN
TYPE B
USA, MEX, CAN & JPN
TYPE G
GBR, IRL, MLT, MAS & SIN
TYPE H
ISRAEL
TYPE L
ITALY & CHILE
TYPE K
DENMARK & GREENLAND

Естественнонаучные стандарты

“Стандарты данных – это правила, по которым данные описываются и регистрируются. Для того, чтобы данные распространять, обмениваться ими и понимать их, мы должны стандартизировать как формат, так и смысл понятий.” (Геологическая служба США)

Некоторые существующие стандарты данных:

- Язык стандарта экологических метаданных (EML),
- Описание мультемидейных материалов Audubon (Audubon Core),
- Global Глобальная сеть по биоразнообразию генома (GGBN),
- Проект передового опыта и стандартов об океанических данных (ODSBP),
- **Darwin Core**

Стандарты данных

TDWG - Biodiversity Information Standards

Организация основана в 1985 году под названием “Рабочая группа по таксономическим базам данных”

Первая конференция 1985 - Geneva, Switzerland

2005 - Россия, Санкт-Петербург, Зоологический институт РАН

2019 - Leiden, The Netherlands, 20-25 октября

Biodiversity Information Standards TDWG

Audiobone Core
Access Protocol for Information Retrieval (TAPIR)
Standards Documentation Standard (SDS)
…
Darwin Core (DwC)

GBIF

Что такое Darwin Core?

“Список полей и их определения, как они связаны с данными о биоразнообразии.”

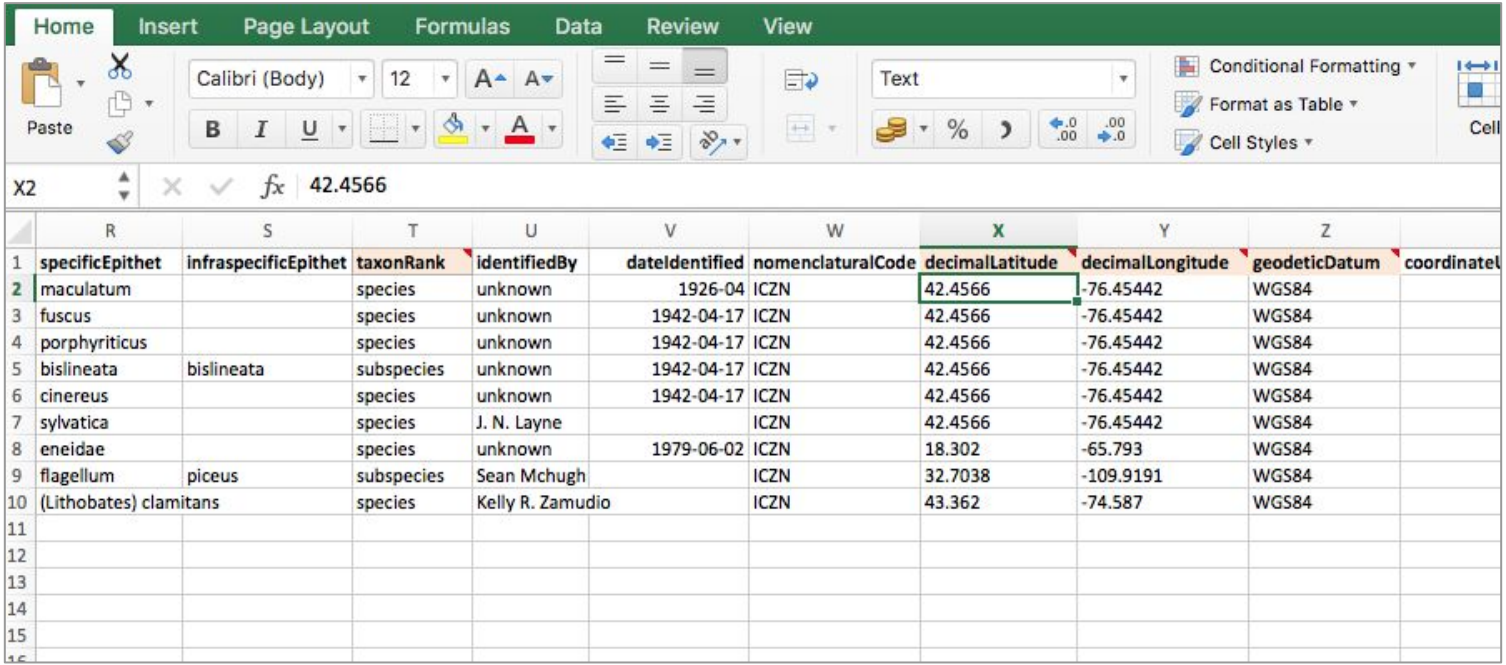
Darwin Core

Darwin Core - это стандарт, поддерживаемый группой Darwin Core из TDWG. Он включает в себя словарь терминов (в разных ситуациях их также могут назвать свойствами, элементами, полями, столбцами, атрибутами или понятиями), предназначенных для облегчения обмена информацией о биологическом разнообразии путем предоставления идентификаторов, меток и определений. Darwin Core основан на таксонах, их появлении в природе, подтвержденном наблюдениями, образцами, сборами и сопутствующей информацией.

- Краткое руководство
- Руководства по использованию:
  как использовать Darwin Core как Simple Darwin Core, text, XML или RDF
- GitHub репозиторий: страница Darwin Core
- Normative document: CSV файл с полной версией истории терминов Darwin Core
- Distribution files: шаблоны для начала использования Darwin Core

Darwin core - пример таблицы

Darwin Core



X2 | 42.4566

	R	S	T	U	V	W	X	Y	Z
1	specificEpithet	infraspecificEpithet	taxonRank	identifiedBy	dateIdentified	nomenclaturalCode	decimalLatitude	decimalLongitude	geodeticDatum
2	maculatum		species	unknown	1926-04	ICZN	42.4566	-76.45442	WGS84
3	fuscus		species	unknown	1942-04-17	ICZN	42.4566	-76.45442	WGS84
4	porphyriticus		species	unknown	1942-04-17	ICZN	42.4566	-76.45442	WGS84
5	bislineata	bislineata	subspecies	unknown	1942-04-17	ICZN	42.4566	-76.45442	WGS84
6	cinereus		species	unknown	1942-04-17	ICZN	42.4566	-76.45442	WGS84
7	sylvatica		species	J. N. Layne		ICZN	42.4566	-76.45442	WGS84
8	eneidae		species	unknown	1979-06-02	ICZN	18.302	-65.793	WGS84
9	flagellum	piceus	subspecies	Sean Mchugh		ICZN	32.7038	-109.9191	WGS84
10	(Lithobates) clamitans		species	Kelly R. Zamudio		ICZN	43.362	-74.587	WGS84
11									
12									
13									
14									
15									

Darwin Core - это список заголовков столбцов вашей таблицы и правила как это таблицу заполнять

GBIF

Стандарт Darwin Core (DwC) - спецификация

Darwin Core

TDWG Home Terms Guides Namespace policy

Darwin Core quick reference guide

This page provides a list of all currently recommended terms of the Darwin Core standard. Categories such as `Occurrence` or `Event` correspond to Darwin Core classes which group other terms. Convenient files of these terms and their full history can be found in the Darwin Core repository.

Record-level

type | modified | language | license | rightsHolder | accessRights | bibliographicCitation | references
institutionID | collectionID | datasetID | institutionCode | collectionCode | datasetName
ownerInstitutionCode | basisOfRecord | informationWithheld | dataGeneralizations | dynamicProperties

type	Property
Identifier	http://purl.org/dc/terms/type
Definition	The nature or genre of the resource.
Comments	Must be populated with a value from the DCMI type vocabulary (http://dublincore.org/documents/2010/10/11/dcmi-type-vocabulary/).
Examples	`StillImage`, `MovingImage`, `Sound`, `PhysicalObject`, `Event`, `Text`

Record-level
Occurrence
Organism
MaterialSample
Event
Location
GeologicalContext
Identification
Taxon
MeasurementOrFact
ResourceRelationship
UseWithIRI

LivingSpecimen
PreservedSpecimen
FossilSpecimen
HumanObservation
MachineObservation

Стандарт Darwin Core - basisOfRecord

Darwin Core

TDWG Home Terms Guides Namespace policy

Record-level

type | modified | language | license | rightsHolder | accessRights | bibliographicCitation | references
institutionID | collectionID | datasetID | institutionCode | collectionCode | datasetName
ownerInstitutionCode | basisOfRecord | informationWithheld | dataGeneralizations | dynamicProperties

basisOfRecord	Property
Identifier	http://rs.tdwg.org/dwc/terms/basisOfRecord
Определение	То, что послужило основанием для данной записи.
Комментарии	Рекомендуется использовать значение из словаря (перечислены ниже)
Примеры	`PreservedSpecimen`, `FossilSpecimen`, `LivingSpecimen`, `MaterialSample`, `Event`, `HumanObservation`, `MachineObservation`, `Taxon`, `Occurrence`

Record-level
- Occurrence
- Organism
- MaterialSample
- Event
- Location
- GeologicalContext
- Identification
- Taxon
- MeasurementOrFact
- ResourceRelationship
- UseWithIRI

- LivingSpecimen
- PreservedSpecimen
- FossilSpecimen
- HumanObservation
- MachineObservation

Basis of Record среди обязательных для заполнения терминов - на основании чего сделдана запись в таблице

Стандарт Darwin Core (DwC)

Darwin Core

Классы полей

в настоящее время в DwC описано использование использование **169 терминов** в варианте Simple DwC,
всего терминов **более 200** в основном варианте и еще **32 расширения** со своим небольшим набором терминов

Расширения

Геоботанические описания
История определений
Местные названия
…..

Простой Darwin Core

- Запись & Набор данных
- Находка
- Организм
- Образец материала
- Событие
- Место нахождения
- Геологический контекст
- Определение
- Таксон

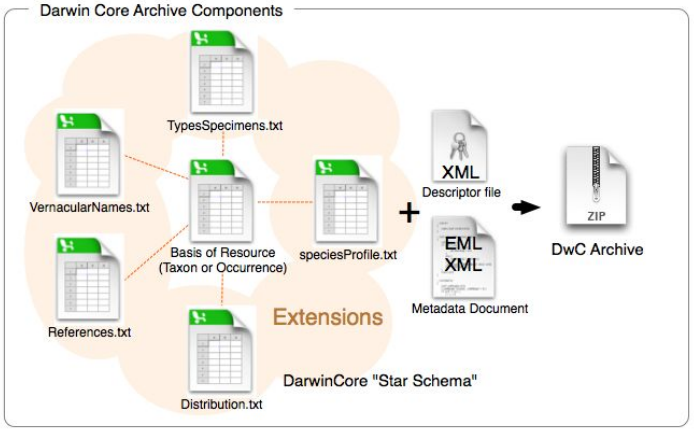
Дополнительные классы

- Взаимосвязь ресурсов
- Измерения и факты

Архив Darwin Core (DwC)

Это архив, которые содержит как минимум три других файла. Кодировка - UTF-8.

1. Файл с данными (occurrence.txt) соответствует простому DwC в формате CSV, первая строка которого содержит названия стандартных терминов Darwin Core.
2. Мета-файл (meta.xml) в формате XML содержит технические подробности для указания компьютеру как использовать файл с данными.
3. Мета-файл (eml.xml) в формате XML содержит пояснительные сведения о записях, содержащихся в файле данных, чтобы проинструктировать пользователя о том, пригодны ли данные для использования.



Может также включать:

- ссылки на связанные медиа-файлы (фотографии)
- местные названия таксонов
-

http://rs.tdwg.org/dwc/terms/guides/text/index.htm

Стандарт Darwin Core (DwC)

Darwin Core

GBIF Global Biodiversity Information Facility

Occurrence-only datasets

Learn more about occurrence-only and other classes of datasets currently supported on GBIF.org

Darwin Core record details

Term	Status
occurrenceID	Required
basisOfRecord	Required
scientificName	Required
eventDate	Required
countryCode	Required
taxonRank	Strongly recommended
kingdom	Strongly recommended
decimalLatitude & decimalLongitude	Strongly recommended
geodeticDatum	Strongly recommended
coordinateUncertaintyInMeters	Strongly recommended
individualCount, organismQuantity & organismQuantityType	Strongly recommended
informationWithheld	Share if available
dataGeneralizations	Share if available
eventTime	Share if available
country	Share if available

Все термины (поля) стандарта Darwin Core для каждого набора данных и для каждой записи заполнить невозможно, но среди этого многообразия есть термины **обязательные**, а также **рекомендуемые** для заполнения.

Требования к качеству данных

https://www.gbif.org/data-quality-requirements

Несколько примеров существующих расширений

Расширения Darwin Core

Расширения (**Extensions**) позволяют опубликовать дополнительную, более подробную или специфическую информацию - *генетические данные*, *описание растительности*, *библиографию*, *историю определений*, *параметры среды*, ...

Extensions

Extensions provide the means to associate additional data with a record of the Core Type. Each Extension is meant to accommodate a different type of data. Please install all of the Extensions you require from the list below.

Darwin Core Resource Relationship [Remove]

Support for generic resource rleations according to Darwin Core terms.
See also http://rs.tdwg.org/dwc/terms/index.htm#relindex

Issued	Feb 13, 2015
Properties	7
Name	ResourceRelationship
Namespace	http://rs.tdwg.org/dwc/terms/
RowType	http://rs.tdwg.org/dwc/terms/ResourceRelationship

Germplasm (0.1) [Remove]

WARNING! THIS DRAFT GERMPLASM EXTENSION HAS BEEN DEPRECATED. SEE: http://rs.gbif.org/extension/germplasm/ FOR THE CURRENT GERMPLASM EXTENSIONS.
See also http://rs.nordgen.org/dwc/

Issued	Feb 13, 2015
Properties	50
Name	GermplasmSample
Namespace	http://rs.nordgen.org/dwc/germplasm/0.1/terms/
RowType	http://rs.nordgen.org/dwc/germplasm/0.1/terms/GermplasmSample
Keywords	dwc:Occurrence

Germplasm accession (v20140515) [Remove]

The Germplasm Vocabulary provides a set of terms (supplementing the Darwin Core terms) for describing genebank accessions. These terms are maintained by the thematic community of plant genetic resources for food and agriculture (PGRFA). Most of these terms are imported from the Multi-Crop Passport Descriptor List (MCPD) maintained by Bioversity International and the Food and Agriculture Organization of the United Nations (FAO). Some terms were also developed by the European Cooperative Programme for Plant Genetic Resources (ECPGR).
See also http://purl.org/germplasm/germplasmTerm

Issued	May 20, 2014
Properties	37
Name	GermplasmAccession
Namespace	http://purl.org/germplasm/germplasmTerm#
RowType	http://purl.org/germplasm/germplasmTerm#GermplasmAccession

Alternative Identifiers [Remove]

See also http://rs.gbif.org/extension/gbif/1.0/identifier.xml

Issued	Feb 13, 2015

Зачем использовать Darwin Core?

Darwin Core

Спасибо за внимание !!!

GBIF