# GBIF in the cloud workshop framework

## Aims

This document aims to describe the EBRII Platform workshop for the participants, explaining what we try to achieve and what they will do during the workshop. This workshop contains a small data part and a big data part accessing GBIF occurrences snapshots in the cloud.

## Part I : Small Data

### Introduction

During the morning session, you will use Jupyter Lab Framework to explore the specimens data of Ghent University (see `https://doi.org/10.15468/dl.axpas3`)

#### SD-Ex0

#### Aims

To download data for the Small Data exercises.

#### Steps

- Login to GBIF.org
- Select occurrences with
    - basisOfRecord=’preserved specimen’
    - dataPublisher= ‘Ghent  University’
- Download data in simple CSV format
- Explore the DarwinCore fields

#### Questions

1. What records would you exclude? Why?

## Python Exercises

#### SDP-Ex1

#### Aims

Participants will get acquainted with Jupyter Lab, Python and Pandas data frame.

#### Pre-requisites

1. SD-EX0 finished

2. Have access to JupyterLab
3. Create your own Jupyter notebook

## Steps

- Load CSV into a panda dataframe
- Subset some columns 'gbifID', 'individualCount','countryCode', 'year', 'class', 'speciesKey'
- Drop rows with null values
- Save this dataframe in a EX1.CSV file
- Describe the resulting data

## Questions

1. How many rows do you have in the original file?
2. How many rows remain after dropping null values?

# SDP-Ex2

## Aims

Participants will do simple manipulations on Pandas data frame.

## Pre-requisites

1. SDP-EX1 finished

## Steps

- Load EX1.csv into panda dataframe
- Group data by class
- Group data by countryCode
- Group data by year
- Create a dataframe with all Belgian Insecta specimens, discard the others
- Describe the resulting data

## Questions

1. How many rows do you have for class Insecta?
2. How many rows do you have for Belgium?
3. How many rows do you have for 1961?
4. How many Belgian Insecta in the 1960's rows do you have?

## SDP-Ex3

### Aims

Participants will create a data cube with 3 dimensions : taxonomy, geography and time.

### Pre-requisites

1. SDP-EX2 finished

### Steps

- Load EX1.csv into panda dataframe
- Add a ‘decade’ field
- Count the number of records, numbers of specimens and species for this cell : Insecta, Belgium, 1960s
- Create a 3D data frame based on
    - Taxonomy by class
    - Geography by countryCode
    - Time by decade
- In addition to these fields, each cell of your cube will contain:
    - Number of records (count(*))
    - Number of specimens (sum of ‘individual_count’)
    - Number of species (distinct (species_key))
- Save your data cube in EX3.csv file

### Questions

Inspect three cells of your data cube:

- A (Aves, Belgium, 1950s)
- B (Insecta, Belgium, 1900s)
- C (Insecta, Belgium, 1950s)

For each cell, answer the following questions:

1. How many records do you have in these cells?
2. How many specimens do you have in these cells?
3. How many species do you have in these cells?

## SDP-Ex4

### Aims

Participants will combine their data cube with countries information

### Pre-requisites

1. SDP-EX3 finished

### Steps

- Load your Data Cube(EX3.csv) in a dataframe
- Filter the Insecta cells, discard the others
- Load Countries.csv in a dataframe
- Join the two dataframe on countryCode add country region and  sub-region
- Save your results in EX4.csv file

### Questions

1. How many non empty cells do you have?

### SDP-Ex5

### Aims

Participants will use their data cube to plot specimens provenance over time per continent and subregion.

### Pre-requisites

1. SDP-EX4 finished

### Steps

- Goto RawGraphs website
- Load your enriched Data Cube(EX4.csv)
- Plot specimens provenance over time per region and sub-region

### Questions

- Are there differences between the regions?
- Are there gaps in time(decade) or geography(country, region, sub-region?
- Which regions/subregions offer the highest biodiversity(number of species)?
- Where would you recommend to collect specimens today?

## Useful links

Getting started with pandas
Data Wrangling with pandas Cheat Sheet
RawGraphs

## Further readings

The easiest way to plot data from Pandas on a world map
An intro to Python GIS

## SQL Exercises

If time allows, re-implement the python exercises SDP-Ex1 - SDP-Ex4 in SQL.

# Part II : Big Data

## Introduction

During the afternoon session, you will use Databricks Framework to explore all specimens data available on GBIF.org

## Exercices

### BD-Ex0

- Get your Databricks account ready, connect to BigData workspace.
- Connect to Databricks Workspace

## SQL Exercises

### BDS-Ex1

### Aims

Participants will get acquainted with Databricks.

### Pre-requisites

1. BD-EX0 finished
2. Have access to Databricks

### Steps

- Create a new notebook
- Discover the default.occurrence_20220601 table
- Create a view on preserved specimen records with year, class and countryCode not null
- Explore further the data

### Questions

- How many occurrences are recorded in the GBIF snapshot?
- How many specimens are recorded?
- How many specimens are recorded with year, class and countryCode?
- What other records would you exclude from your analysis? Why?
  See individualCount, decimalLat/long, speciesKey, issues…

## BDS-Ex2

### Aims

Participants to discover GBIF mediated specimens data.

### Pre-requisites

1. BD-EX1 finished
2. Have access to Datrabricks

### Steps

- Create a table on specimen records with year, class and countryCode not null
- Subset some columns 'gbifID', 'individualCount','countryCode', 'year', 'class', 'speciesKey'
- Drop rows with null values
- Describe the resulting data

### Questions

- Which ‘class’ has more **specimens records**?
- Which ‘class’ has more **distinct species of specimens**?
- Which ‘class’ has more **individual specimens**?

## BDS-Ex3

### Aims

Participants will create a data cube with 3 dimensions : taxonomy, geography and time.

### Pre-requisites

1. BDS-EX2 finished

### Steps

- Create a DataCube view based on
    - Taxonomy by class
    - Geography by countryCode
    - Time by decade
- In addition to these fields, each cell of your cube will contain:
    - Number of records (count(*))
    - Number of specimens (sum of ‘individual_count’)
    - Number of species (distinct (species_key))

### Questions

Inspect three cells of your data cube:

- A (Aves, Australia, 1920s)

- B (Insecta, Australia, 1900s)
- C (Insecta, Australia, 1950s)

For each cell, answer the following questions:

1. How many records do you have in these cells?
2. How many specimens do you have in these cells?
3. How many species do you have in these cells?

## BDS-Ex4

### Aims

Participants will combine their data cube with countries information

### Pre-requisites

1. BDS-EX3 finished

### Steps

- Explore your datacube table
- Explore Countries table
- Join the two on countryCode, add country's region and  sub-region
- Select the class you want to analyse
- Export your datacube for that class in a CSV file (EX4.csv)

### Questions

1. Which region, subregion have the most records of your class?
2. Which region, subregion have the most specimens of your class?
3. Which region, subregion have the most species of your class?

## BDS-Ex5

### Aims

Participants will use their data cube to plot specimens provenance over time per continent and subregion.

### Pre-requisites

1. BDS-EX4 finished

### Steps

- Goto RawGraphs website
- Load your enriched Data Cube(EX4.csv)
- Plot specimens provenance over time per region and sub-region
    - Number of records

    - Number of specimens
    - Number of species

### Questions

- Are there differences between the regions?
- Are there gaps in time(decade) or geography(country, region, sub-region?
- Which regions/subregions offer the highest biodiversity(number of species)?
- Where would you recommend to collect specimens today?

## Useful links

Azure Databricks documentation
Quickstart: run and visualize a query
RawGraphs

## Participants survey

Please fill this short survey regarding the workshop. Answers are anonymous and it should not take more than 5 minutes!