GBIF IN THE CLOUD

Belgian Biodiversity Platform
29 June 2022 André Heughebaert,
Sebastien Ronveaux & Maxime Coupremanne

GBIF

EMPOWERING BIODIVERSITY RESEARCH

.be
Biodiversity.be

SUMMARY

- Introduction
- Practicalities
- Part I: Small Data
- Part II: Big Data
- Conclusion

INTRODUCTION

You will have the opportunity to elaborate simple IT solutions similar to those needed by real scientific analysis. The workshop will cover the **discovery, query, filtering of data** but also the **plotting** into visually effective diagrams.

INTRODUCTION

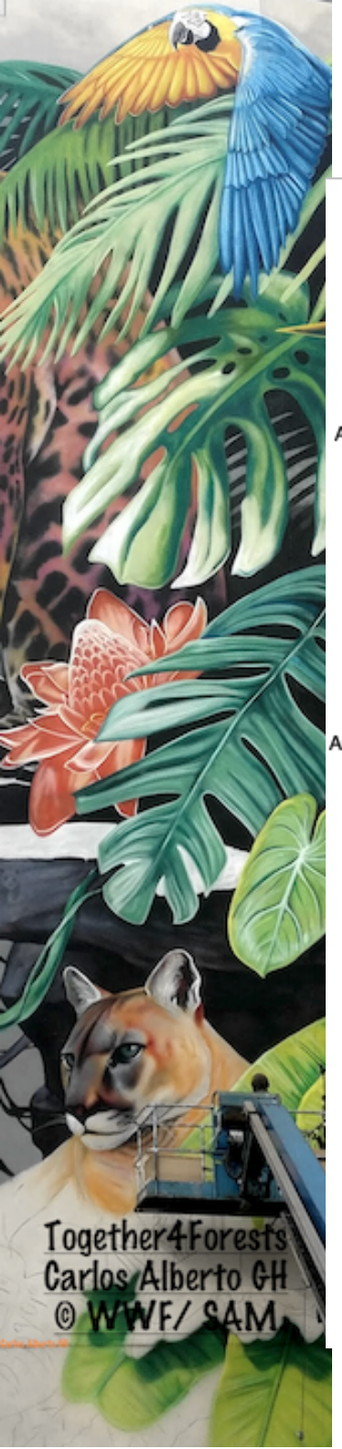
You will work at two levels(small and big data), you will be able to develop and run solutions both locally on their laptop and on remote servers in the cloud. You will have the opportunity to **create, write and run** their code using two excellent Data Science tools : **Jupyter and Databricks** notebooks.

INTRODUCTION

Data science tools of the day:

- Jupyter
- Azure/Databricks
- RawGraphs

What will you do?
***Discover, query, filter, plot and visualise** GBIF occurrences*



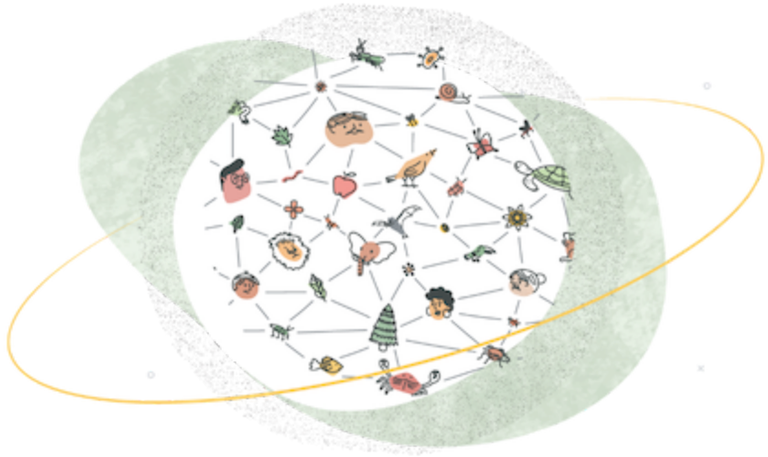
When/where were the Bivalvia specimens collected?

region
Asia
America
Africa
Antarctica
Oceania
Europe
decade
1,800
1,820
1,840
1,860
1,880
1,900
1,920
1,940
1,960
1,980
2,000
2,020

numberSpecies
1.0
858.0

sub-region
Antarctica
Australia and New Zealand
Central Asia
Eastern Asia
Eastern Europe
Latin America and the Caribbean
Melanesia
Micronesia
Northern Africa
Northern America
Northern Europe
Polynesia
South-eastern Asia
Southern Asia
Southern Europe
Sub-Saharan Africa
Western Asia
Western Europe

ABOUT GBIF



GBIF.org, assessed on 16 May 2022

2,167,543,437
Occurrence records

69,263
Datasets

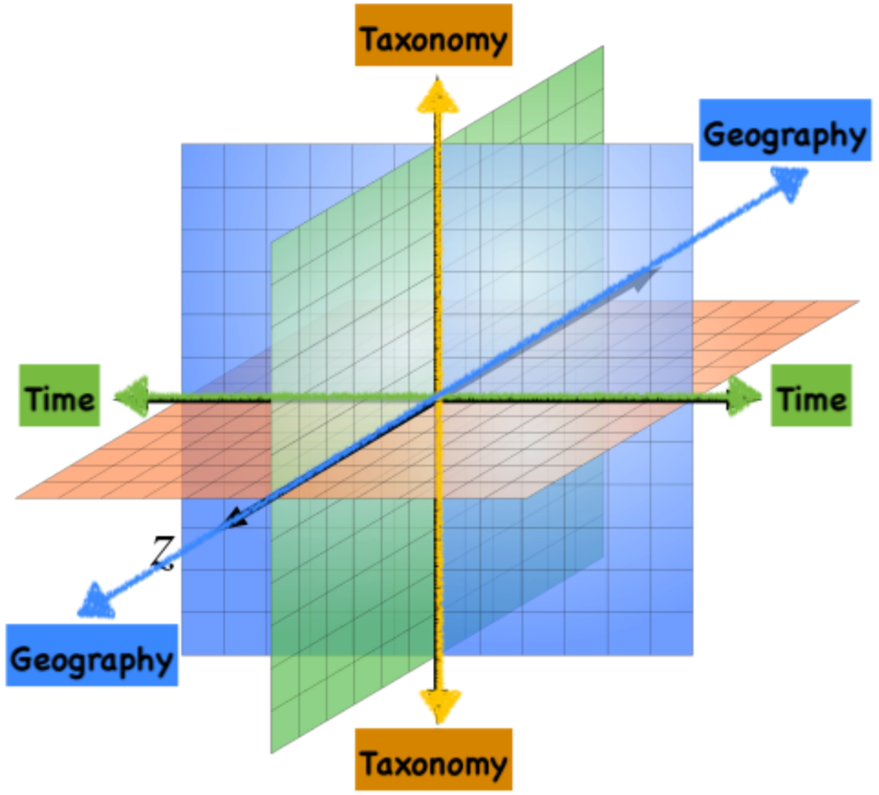
1,831
Publishing institutions

7,196
Peer-reviewed papers using data

ABOUT GBIF

Occurrences snapshots are available on the cloud: Amazon, Microsoft Planatary Computer (Azure), Google
see GBIF news, 21 May 2021

WHAT? ⇄ WHERE? ⇄ WHEN?



Of course, additional information are attached to these points such as **events, methods, measurements or environmental** data.
But, let's have a look at the DarwinCore fields that correspond to our 3 dimensions (what? where? when?).

1. TAXONOMY(=WHAT?)

kingdom, phylum, class, order,
family, genus, species, infraspecificEpithet,
taxonRank, taxonKey, **speciesKey**, scientificName,
verbatimScientificName,
verbatimScientificNameAuthorship

2. GEOGRAPHY(=WHERE?)

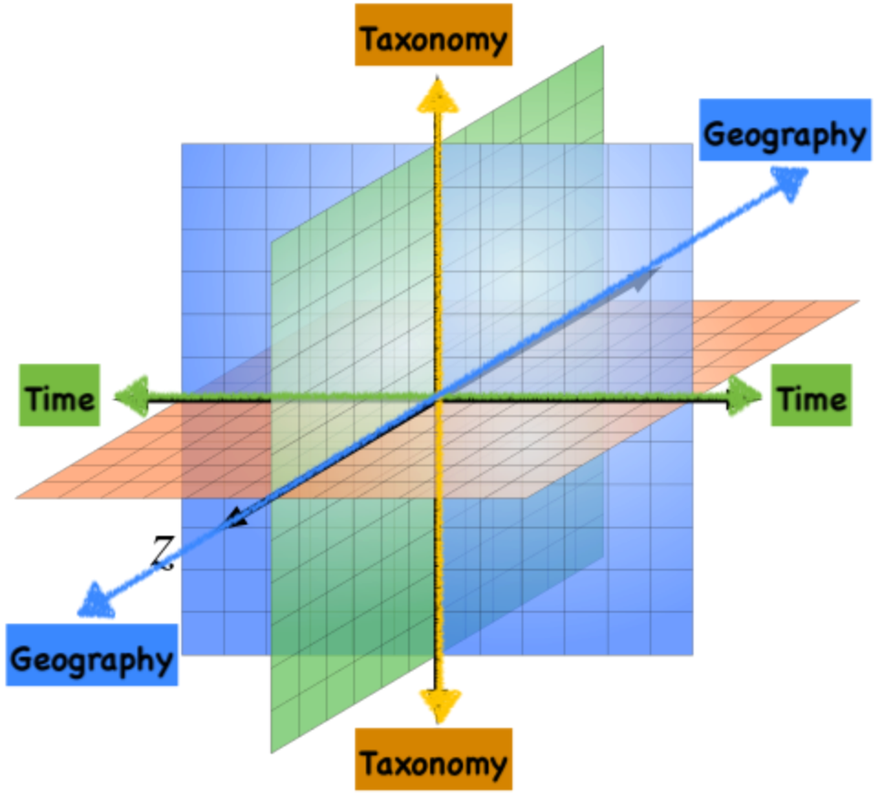
countryCode, locality, stateProvince,
decimalLatitude, **decimalLongitude**,
coordinateUncertaintyInMeters, coordinatePrecision,
elevation, elevationAccuracy, depth, depthAccuracy

3. TIME(=WHEN?)

eventDate, day, month, **year**

Others DarwinCore fields of interest:
gbifID, datasetKey, occurrenceID, occurrenceStatus,
individualCount, **publishingOrgKey**, **basisOfRecord**,
institutionCode, collectionCode, catalogNumber,
recordNumber, identifiedBy, dateIdentified,
license, rightsHolder, recordedBy, **typeStatus**,
establishmentMeans, lastInterpreted, mediaType,
issues

GRANULARITY



Granularity affects the number of cells in our 3DCube.
If we choose **years**, **species** and **Lat/long (1x1°) squares**, we'll have zillion of cells.

*25.920.000.000.000 cells (=200 years * 2.000.000 species * 64.800 squares)*

We take **decades**, **classes** and **countries**, the number of cells becomes more manageable.

*2.000.000 cells (= 20 decades * 400 classes * 250 countries)*

PRACTICALITIES

09:00- 09:30	Introduction
09:30- 12:30	PartI Small Data
12:30- 13:30	*---lunch break---*
13:30- 16:30	PartII Big Data
16:30- 17:00	Conclusion

PRACTICALITIES

Wi-Fi: ***WTC-Guest***
Password: ***WTC0nTh34ir!***

PART I: SMALL DATA

JupyterHub

https://sandbox.bebif.be/jupyter

login=user{x}

password='ebr'

PART II: BIG DATA

azureDatabricks

login=training{x}@gbifcloudtraining.onmicrosoft.com

password='PumaTraining2022'

RAW GRAPHS

https://sandbox.bebif.be/rg/

no login!

PART I : SMALL DATA

We will analyze the specimens of Ghent University,
about 28.000 records

- *Ex0 on GBIF.org*
- *Ex1-4 on Jupyter*
  *with **Python & pandas***
- *Ex5 on RawGraphs*

SP_EXO

On GBIF.org, download occurrences:

- preserved collection specimens (**basisOfRecord** = 'preserved specimen')
- from Ghent University Data provider (**publishingOrgKey**=05c249d0-dfa0-11d8-b22e-b8a03c50a862)

Observe the downloaded records
What records would you exclude? Why?

SP_EX1

1. Load occurrences in a dataframe
2. Extract these columns: **gbifID, individualCount, countryCode, year, class, speciesKey**
3. Discard records with null() values
4. Save results in 'EX1.csv'

How many records do you have?

SP_EX2

1. Load EX1.csv in a dataframe
2. Save results in './data/SD_EX2.csv'
3. group occurences by class, country and decade

How many records for 'Insecta'?
How many records for 'Belgium'?
How many records for '1961'?
How many Belgian Insecta records for 1960s?

SP_EX3

1. Count the number of records, numbers of specimens and species
2. Our data cube consists of 3D : **decade, countries and classes**. For each cell in this cube, add **rcount**(#records), **icount**(#specimens), **scount** (#species)
3. Save your data cube in EX3.csv file

SP_EX3

Inspect three cells of your data cube:

- A= Aves, Belgium, 1920s
- B= Insecta, Belgium, 1900s
- C= Insecta, Belgium, 1950s

- *How many records ?*
- *How many specimens ?*
- *How many species ?*

SP_EX4

1. Load your Data Cube(EX3.csv)
2. Filter the Insecta cells, discard the others
3. Load Countries.csv in a dataframe
4. Join the two dataframe on countryCode
5. Add country region and sub-region
6. Save your results (export this data cube in a CSV file)

*How many **non empty** cells do you have?*
(rcount,icount,scount!=0)

SD_EX5

1. Upload your datacube on RawGraphs.io
2. Choose *Beeswarm plot* chart
3. Mapping
    3.1. XAxis= decade
    3.2. Size=#numberRecords/Specimens/Species
    3.3. Color=sub-region
    3.4. Groups=region
4. Customize your graph
    4.1. ShowLegend
    4.2 Chart/diameter= 1-40
    4.2 Color/ColorScheme= turbo discrete
5. Export the resulting graph in appropriate format

SD_EX5

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?

SQL EXERCISES 2-4

(If time allows, re-implement exercises 2-4 with SQL)

PART II : BIG DATA

We will now analyze
all collection specimens available on GBIF,
(about 200 million records).
from an occurrences snapshot
with **Databricks and SQL**

BD_EX1

1. Create a new notebook
2. Create a view on specimens records
3. Explore the data

How many occurrences and specimens?
How many with year, country and class?
What other records would you exclude? Why?

BD_EX2

1. Create a specimen view (year, class and countryCode not null)
2. Subset with 'gbifID', 'individualCount','countryCode', 'year', 'class', 'speciesKey'
3. Drop rows with null values
4. Describe the resulting data

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

BD_EX3

- Create a DataCube view based on **class, country and year**
- Add to each cell:
  - Number of records
  - Number of specimens
  - Number of species

BD_EX3

Inspect three cells of your data cube:

- cell A (Aves, Australia, 1920s)
- cell B (Insecta, Australia, 1900s)
- cell C (Insecta, Australia, 1950s)

How many records?
How many specimens?
How many species?

BD_EX4

1. Explore your datacube
2. Explore Countries table
3. Join the two on countryCode, add country's region and sub-region
4. Select the class you want to analyze
5. Export your datacube for that class

Which region, subregion have the most records?
Which region, subregion have the most specimens?
Which region, subregion have the most species?

BD_EX5

1. Upload your datacube on RawGraphs.io
2. Choose *Beeswarm plot* chart
3. Mapping
    3.1. XAxis= decade
    3.2. Size=#numberRecords/Specimens/Species
    3.3. Color=sub-region
    3.4. Groups=region
4. Customize your graph
    4.1. ShowLegend
    4.2 Chart/diameter= 1-40
    4.2 Color/ColorScheme= turbo discrete
5. Export the resulting graph in appropriate format

BD_EX5

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?

CONCLUSIONS

Today, you played with:

Jupyter (Python & Pandas)
Databricks (SQL)
RawGraphs.io
GBIF occurrences
Darwin Core

Discussion and Survey:

- Did you learn something?
- What did you like/dislike?
- Will you use these tools in the future?
- How would you improve the workshop?

THANK YOU!

✉ @andrejjh on twitter

GBIF

EMPOWERING BIODIVERSITY RESEARCH

.be Biodiversity.be

and thanks to **Carlos Alberto GH** Mexican street art painter,
WWF & Street Art for Mankind Together4Forests campaign