GBIF Data Discovery

GBIF Data on your Fingertips –
Visualize Data & Gaps for Your Research

Inspiration

GBIF

- Is an open-data research infrastructure
- Provides free and open online access to species occurrence (primary biodiversity) data through their Website & APIs

Data Prone to Gaps

- Have certain limitations and gaps in occurrence data
- Due to various reasons: sampling biasness, lack of efforts or lack of coverage of distribution of organisms, etc.

Benefits of Data Gap Analysis

- Data users – Determine if available data is suitable and sufficient to address their research investigations
- Data holders & Publishers - Prioritize mobilization and digitization efforts

Hence, the need of a Flexible Data Analytics Framework which can analyse any GBIF Data Set, not restricted to any geographical/taxonomy/temporal boundaries

Data Enrichment Process Cycle

- And so, we propose a Software Analytics driven Data Enrichment Process to identify gaps or biases in GBIF mediated data and which can help set priorities for data mobilization or enhancements to GBIF. The process is a 6 step cycle, which repeats until the identified gaps are closed. In order to achieve the purpose, a software based analytics framework - GBIF Data Discovery is presented, which forms an integral part of the entire cycle.

1. Download Data from *GBIF Website* or *API*

- Export GBIF Data in Tab Separated Files either through download API or through query explorer from GBIF website

2. Visualize Trends & Gaps using *GBIF Data Discovery*

- User-friendly & Interactive visual analytics framework for analysing millions of occurrence data in seconds.

3. Determine Biases and Reason

- Analyse the data using charts & maps and study the gaps using in–built algorithms at Every Level of Taxonomy & Region.

4. Formulate & Implement Plan to Reduce Bias

- Determine the biases and reasons for data gaps and plan for reducing the gaps based on the reasons, be it Spatial, Temporal or Taxonomic

5. Collect Occurrence Data from Regions

- Post Implementation of Plan, re-gather occurrence data through the new surveys

6. Sanitize and Store Data in GBIF

- Based on existing methods, GBIF can sanitize and store the occurrence data and make it available over API

GBIF Data Discovery

- **GBIF Data Discovery**
  - A Robust Data Analytics Framework with in-built scripts to extract data, run algorithms & generate dashboards
  - Data Gap Analysis can be easily done across Spatial, Temporal, Taxonomic and Data Sets
  - In built **Dashboards** with Trends, KPIs and Data Gap Analysis, which can by extended by SME with ease
- **Targeted Audience & How will it Help**
  - **Data holders:** Understand where does the gaps lie and where should the efforts be made to close them
  - **Biological knowledge experts:** Helps validate the GBIF Data Sets through a user friendly Interface
  - **Data users:** Helps in assessing whether available data is suitable and sufficient to address their research investigations

Key Features

Information

- **Dashboards**
- **KPIs**
- **Trends**
- **Associative Filters**

Gaps

- **Spatial**
- **Temporal**
- **Taxonomic**
- **Data Set**

Ignorance

- **Inventory Completeness**
- **Ignorance Score**

GBIF Data Discovery – Features

Data Enrichment Process Driver

- A **Robust Data Analytics Framework** with in-built scripts to extract data, run algorithms & generate dashboards

Data Gap Analysis Framework

- Data Gap Analysis can be easily done across **Spatial, Temporal** and **Taxonomic** biases

Highly Extensible Data Gap Analysis

- Data gap Analysis can be done for any **Region**, any **Time Period** and at Every **Taxonomic** Hierarchy Level

Rich Visualizations

- In built **Dashboards** with Trends, KPIs and Data Gap Analysis using Charts & Maps - Provides more Insight than Tabular data

Intelligent Algorithms

- **Ignorance Score** and **Inventory Completeness** calculated on the fly for the current data filters

Associative Filters

- On the fly filters to view trends and data gap analysis at a **granular level** along with multi-level drill downs

Easy to Use

- Can be **configured easily** with 4-5 lines for Spatial granularity, Time period and Half Ignorance Factor

Easy to Extend

- Users can easily create their own dashboards using built in **Dimensions** and **Measures**

High Performance and Scalability

- Analyse **millions of records** in seconds from within the same framework

How it Works

- **GBIF Data Discovery** is a flexible framework built using *Qlik Sense Desktop*, where in you can import Occurrence Data and it will generate the Dashboards & Data Gap Analytics on its own through it's *inbuilt scripts* & *algorithms*. **In 3 easy Steps!**
- All you need to do is, to place the occurrence.txt files (1 or many) in the defined folder.
- The dashboards are interactive and user friendly and supports self discovery.
- The Analysis is available for the entire data sets. However, granular analysis is also supported through associative filters like Event Year, Taxonomy, Basis of Record, etc.
- You can also create your own Dashboards, with an easy drag and drop interface. All the dimensions & Measures are in-built

1. Export GBIF Data

Export GBIF Data in Tab Separated Files either through download API or through query explorer from GBIF website

Place 1 or more occurrence.txt files in Folder E:\GBIF\occurrence (Path is configurable)

2. Import Data into GBIF Data Discovery

Open Qlik Sense Desktop and Open the App 'GBIF Data Discovery'

Open the Data Load Editor and click on 'Load Data'. The App will automatically generate all the Dashboards.

3. Visualizations and Data Gap Analysis

Open the 'App Overview' and View the Trends, using flexible filters and visualize the available data through interactive charts and maps

Obtain insights on Data Gap Analysis, either through the in-built KPIs or through self discovery

Works On All Screens

Desktop / Laptop (Zero Cost)

- Based on Qlik Sense Desktop Platform – Install once and use ‘GBIF Data Discovery' out-of-the-box. No Cost required
- Easy to analyse any occurrence data set through simple configuration.

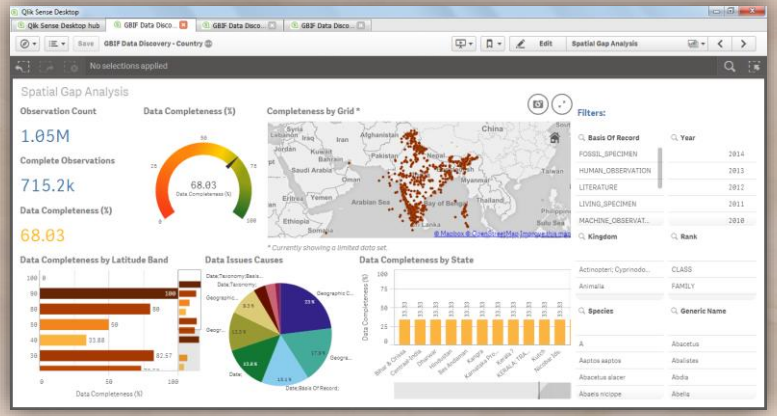
Web Application (Licensed)

- ‘GBIF Data Discovery' can also be made available under Server-Client Architecture through Qlik Sense Server Installation
- Deploy on Server and control the import & sharing of datasets
- Accessible over any device connected to the Internet.
- i.e. Desktop, Laptop, Tablets and Smart Phones
- Easy to share insights with fellow researchers and publishers

Spatial & Temporal Analytics

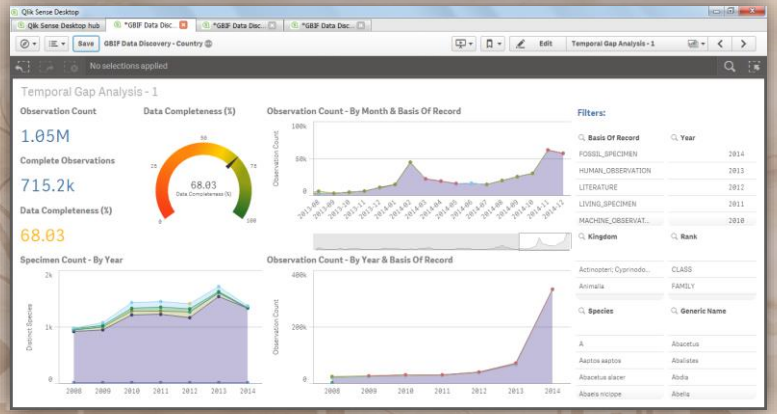
Spatial:

- **Trends:**
  - Occurrence Count by Location (State, Locality)
  - Map Density of Occurrence
  - Major Contributing States for Occurrence Data
- **Gap Analysis:**
  - Data Completeness by:
    - Grid of 1˚ / 0.1˚ / 0.01˚
    - Latitude Bands
    - State / Locality
  - Core Data Issue Reason Distribution



Temporal:

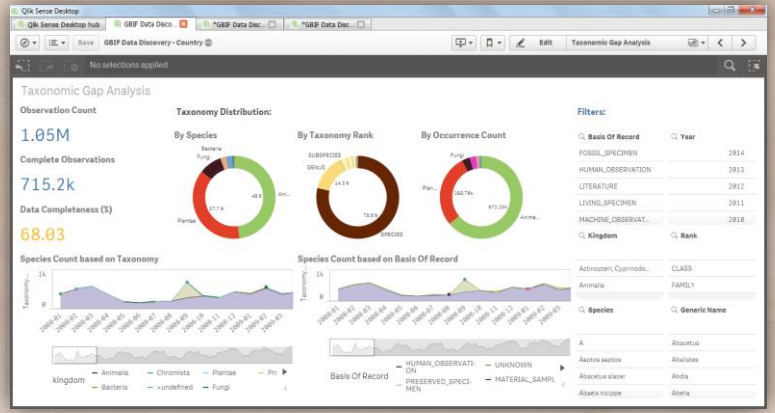
- **Trends:**
  - Monthly Occurrence Count by Taxonomy Classification & Location
  - Daily Species Count
- **Gap Analysis:**
  - Data Completeness by:
    - Basis of Record
    - Yearly / Monthly
  - Mobility Analysis by Month
  - Data Issue Reason Distribution



Taxonomy & Data Sets Analytics

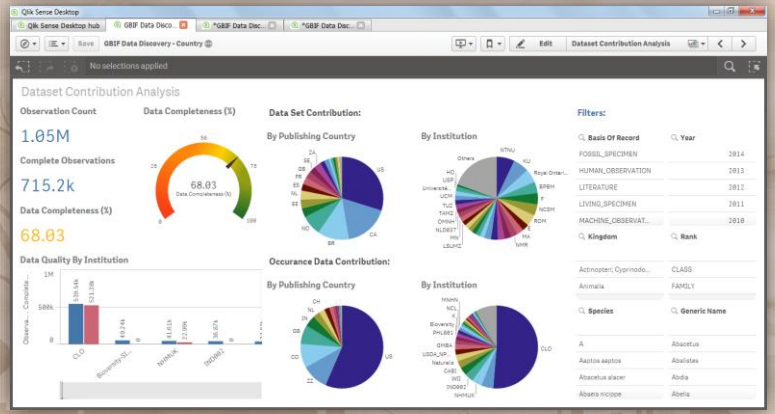
Taxonomy:

- **Trends:**
  - Occurrence Count by Taxonomy Classification
  - Species Count by Taxonomy Classification
  - Location wise Species Count
- **Gap Analysis:**
  - Data Completeness by:
    - Species
    - Taxonomy Rank
    - Occurrence Count
  - Species Count by Taxonomy, Location, Basis Of Record



Data Sets:

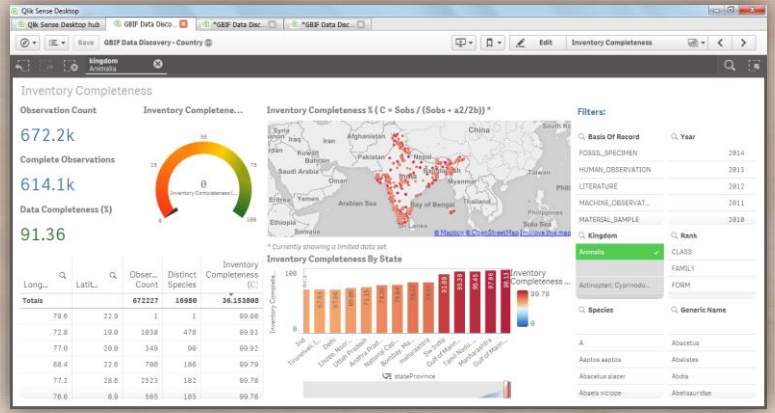
- **Trends:**
  - Data Set Contribution by Publishing Country
  - Data Set Contribution by Institution
- **Gap Analysis:**
  - Data Completeness by:
    - Institution



Inventory Completeness & Ignorance Score Algorithms

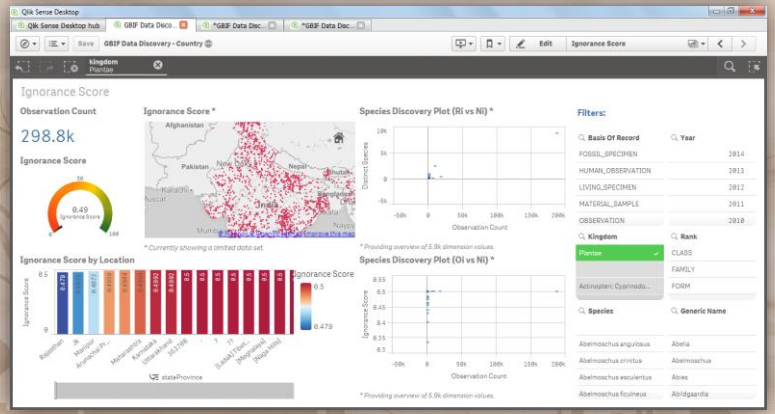
Inventory Completeness:

- Inventory Completeness (C) is calculated for each Grid (1˚ / 0.1˚ / 0.01˚) for each Species
- This is based on **Sousa-Baena et al (2013)** method
- Determines the completeness of geographical knowledge we have for species; Identifies which Grid, for what classification needs more efforts to complete the species inventory
- Highly flexible, you can filter at any level of taxonomy (Kingdom up to species) to get the Inventory Completeness



Ignorance Score:

- Ignorance Score is calculated for each Grid (1˚ / 0.1˚ / 0.01˚) for each Species
- This is based on **Mair L & Ruete A (2016)** method
- Ignorance Score is spatially explicit indices of sampling bias
- Identified which Grid has Species Ignorance, which could therefore inform users of under-sampled areas to be targeted for surveys
- Highly flexible, you can filter at any level of taxonomy (Kingdom up to species) to get the Ignorance Score



Demo

Data Set:

Data downloaded using GBIF API (Refer: http://www.gbif.org/developer/occurrence)

Data downloaded with Filter: country=IN (India) and Year Between 2008 and 2014
Can be replicated with any filter – Geographic/Temporal/Taxonomic

References:

Based on the Suggested Readings provided by GBIF

1. Inventory Completeness –

a. Sousa-Baena MS, Couto Garcia L & Peterson AT (2013) Completeness of digital accessible knowledge of the plants of Brazil and priorities for survey and inventory. Diversity and Distributions 20(4): 369-381. http://doi:10.1111/ddi.12136

b. Stropp J, Ladle RJ, Malhado ACM, Hortal J, Gaffuri J, Temperley WH, Olav Skøien J & Mayaux P (2016), Mapping ignorance: 300 years of collecting flowering plants in Africa. Global Ecology and Biogeography. http://doi:10.1111/geb.12468

2. Ignorance Score –

Mair L & Ruete A (2016) Explaining Spatial Variation in the Recording Effort of Citizen Science Data across Multiple Taxa. PLoS ONE 11(1): e0147796. http://doi:10.1371/journal.pone.0147796

Thank You