GBIF Data Discovery

Team Members

- Team Lead & Data Visualization Expert - **Khyati Majmudar**
- Team Member & Developer - **Ninad Majmudar**

Objective

To propose a **Robust Data Enrichment Process** for GBIF mediated data using **Software based Analytics Framework**. The proposed Process includes the Analytics & Data Gap Visualization Framework -**GBIF Data Discovery**, to analyse all sorts of data gaps, for any region, for any period and at every taxonomic hierarchy level.

The framework consists of built-in *Data Import scripts*, *Intelligent Algorithms* to calculate *Ignorance Score* and *Inventory Completeness* and has *Rich Visualizations* to analyse trends and gaps through *Interactive Dashboards* and *Associative Filtering* of Data. It can also be easily extended to incorporate other complex algorithms to analyse gaps, create new dashboards with in-built dimensions & features as well as make changes to data-import level logic, even by data users/publishers themselves.

Targeted Audience

- **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

Gaps Considered

- **Spatial -** Determine Geographical Biases on occurrence data collection and its quality, at every taxonomic hierarchy level
- **Temporal -** Determine which time of the year is difficult to collect data, for any region, at every taxonomic hierarchy level
- **Taxonomic -** Determine which taxonomic group have bias in collection of data, for any region
- **Data Sets -** Find out which contributing institutions can focus more on improving the data quality, and how
- **Inventory Completeness -** Determine the completeness of geographical knowledge we have for species
- **Ignorance Factor -** Identify Which Geographical Grid has Species Ignorance to understand the under-sampled areas

Proposed Data Enrichment Process

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.

The steps are:

1. Download Data from GBIF Website or API and Import into the framework
2. Visualize Trends & Gaps through the User-friendly & Interactive visualizations which can provide analytics for millions of occurrence data rows in seconds
3. Determine Biases and respective Reasons by analysing the data using charts & maps and study the gaps using in-built algorithms at Every Level of Taxonomy & Region
4. 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. Post Implementation of Plan, re-gather occurrence data through the new surveys
6. Based on existing methods, GBIF can sanitize and store the occurrence data and make it available over API. And then, re-do the analysis till the desired level of data completeness is achieved by restarting from Step 1.

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

Inspiration

GBIF is an open-data research infrastructure which provides free and open online access to species occurrence (primary biodiversity) data through their Website & APIs. However, GBIF mediated occurrence data has certain limitations and gaps, due to various reasons like sampling biasness, lack of efforts or lack of coverage of distribution of organisms, etc.

Through data gap analysis, Data users can determine if available data is suitable and sufficient to address their research investigations. Data holders & Publishers can prioritize mobilization and digitization efforts. And so, we feel that there is a need for a flexible Data Analytics Framework which can be an integral part of the Data Enrichment Process. This framework should analyse any GBIF data set without any restrictions to geography, taxonomy and temporal boundaries.

Generally, we feel that data sets are very lengthy, bulky and tedious to obtain the right information. This brings in the idea to get a software driven data analytics in use and get right visualizations for all the data sets in an interactive framework which can help get insights through algorithms and analytics tools, all in the same platform. And so, we present GBIF Data Discovery to help Data Users and Data Publishers get insights in an effective manner.

How We Built It

GBIF Data Discovery is built using the widely used QlikSense platform. As a desktop App, it can be used directly out of the box by installing the QlikSense Desktop - which is a Free Data Visualization Tool

QlikSense Desktop also will allow Data Users & Publishers to create their own Visualizations to gain further insights.

Using QlikSense Enterprise or Cloud, GBIF Data Discovery can also be hosted on Servers/Cloud, wherein Administrators can control the Data sets and Dashboards. But this will increase the reach of the dashboards since they would be available through Web Apps, thereby accessible over Desktops, Laptops, Tablets, Smart Phones, or any Device which can connect to Internet and has a supported browser.

QlikSense Desktop version can be downloaded from the GIT, and the Cloud version (limited data - Animalia in India, 2011-2014, accessible on Web) is available to limited users, shared with the DevPost Team. Contact us, if you wish to have a go on the Web Application!

How to Use

3 easy steps to setup the framework:

1. 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. 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. 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.

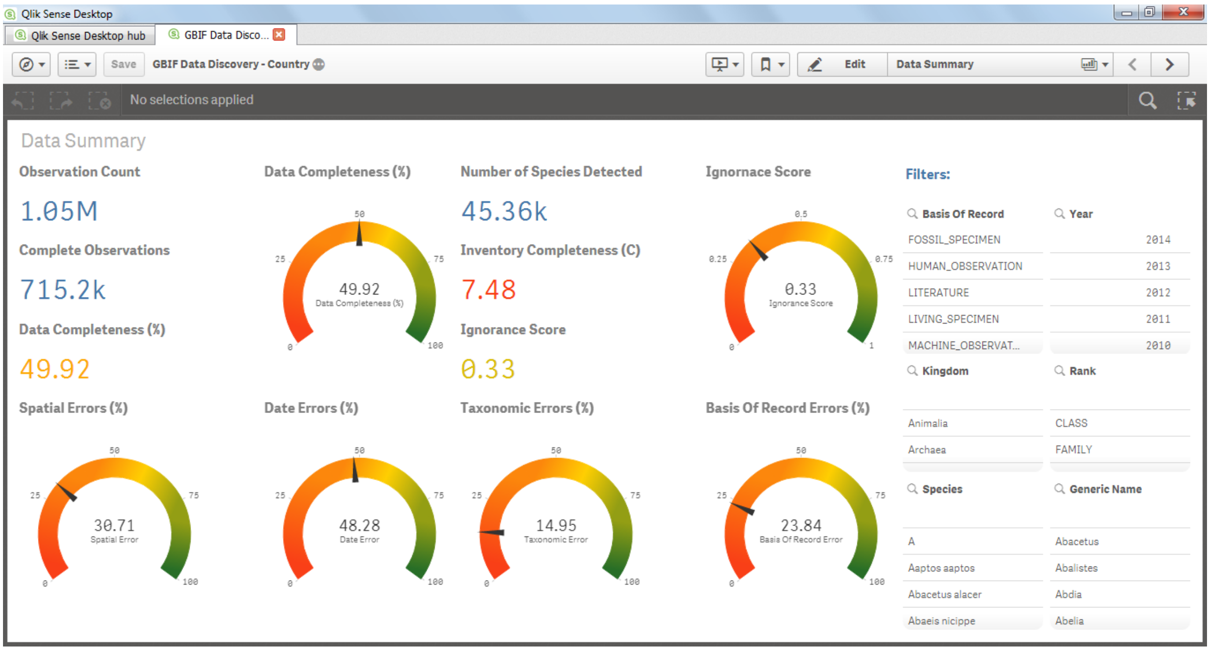
Case Study

To demonstrate the capabilities of **GBIF Data Discovery**, data for **Country-India** for **Years 2008 to 2014** is analysed. And, Spatial, Temporal, Taxonomic, Data Set Information, Inventory Completeness and Ignorance Score based Gaps are considered and recommended actions are provided to close the gaps.

However, GBIF Data Discovery Framework can be used for any data set, for any continent/country/region, for any taxonomy group. Please download appropriate data from GBIF and use the files within GBIF Data Discovery, as per the instructions provided in GitHub Repository Read Me.

The details of the findings are as below:

1. Dashboard: Data Summary



Data Summary

Observation Count 1.05M

Complete Observations 715.2k

Data Completeness (%) 49.92

Data Completeness (%)

Number of Species Detected 45.36k

Inventory Completeness (C) 7.48

Ignorance Score 0.33

Ignornace Score

Spatial Errors (%)

Date Errors (%)

Taxonomic Errors (%)

Basis Of Record Errors (%)

Filters:

Basis Of Record: FOSSIL_SPECIMEN, HUMAN_OBSERVATION, LITERATURE, LIVING_SPECIMEN, MACHINE_OBSERVAT...

Year: 2014, 2013, 2012, 2011, 2010

Kingdom: Animalia, Archaea

Rank: CLASS, FAMILY

Species: A, Aaptos aaptos, Abacetus alacer, Abaeis nicippe

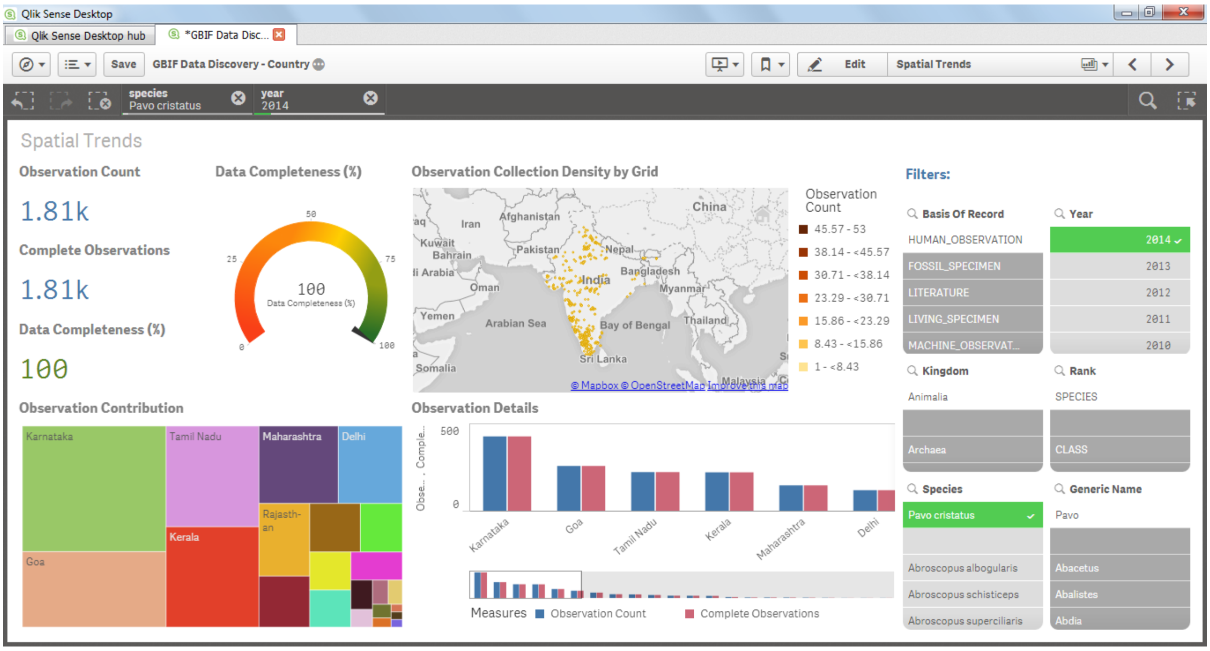
Generic Name: Abacetus, Abalistes, Abdia, Abelia

Data Analysed: Country-India; Years-2008 to 2014

Insights: Overall Data Completeness is at 49.92%. Majority of the data issues is due to Incorrect Date. 45K species have been identified with over 1 million occurrence recorded in 7 years.

Recommended Actions: Surveys should ensure correct and entire date to be recorded

2. Dashboard: Spatial Trends



species Pavo cristatus

year 2014

Spatial Trends

Observation Count 1.81k

Complete Observations 1.81k

Data Completeness (%) 100

Data Completeness (%)

Observation Collection Density by Grid

Observation Count

- ■ 45.57 - 53
- ■ 38.14 - <45.57
- ■ 30.71 - <38.14
- ■ 23.29 - <30.71
- ■ 15.86 - <23.29
- ■ 8.43 - <15.86
- ■ 1 - <8.43

Observation Contribution

Observation Details

Measures ■ Observation Count ■ Complete Observations

Filters:

Basis Of Record: HUMAN_OBSERVATION, FOSSIL_SPECIMEN, LITERATURE, LIVING_SPECIMEN, MACHINE_OBSERVAT...

Year: 2014 ✓, 2013, 2012, 2011, 2010

Kingdom: Animalia, Archaea

Rank: SPECIES, CLASS

Species: Pavo cristatus ✓, Abroscopus albogularis, Abroscopus schisticeps, Abroscopus superciliaris

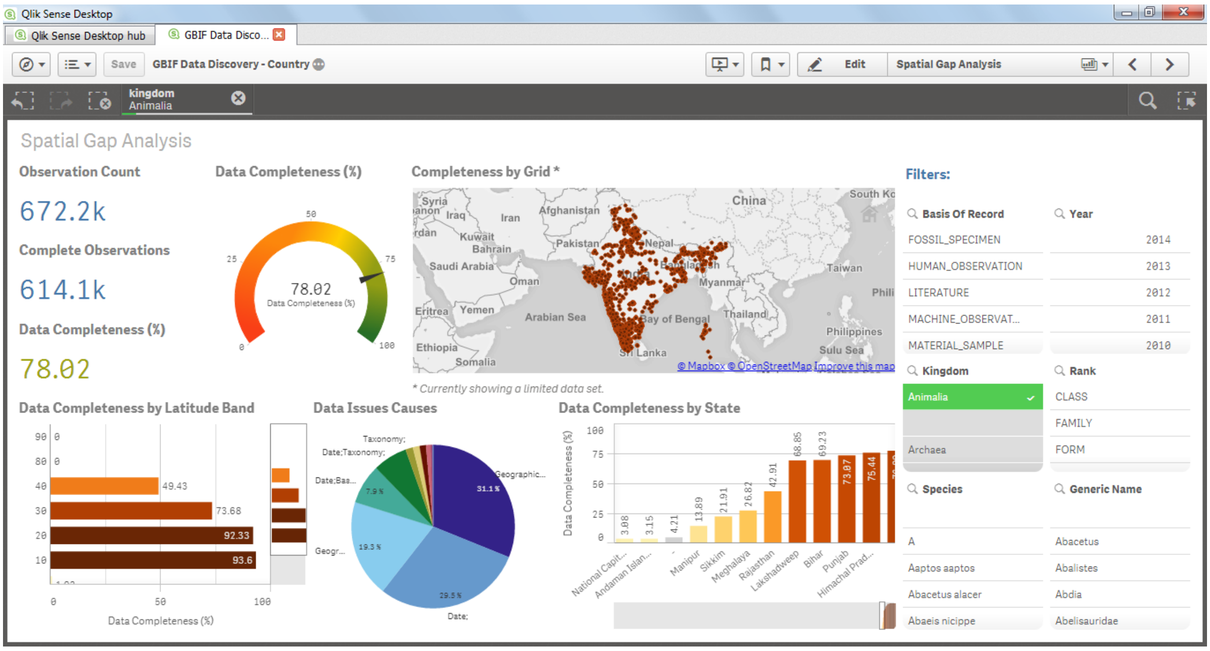
Generic Name: Pavo, Abacetus, Abalistes, Abdia

Data Analysed: Species-Pavo Cristatus (Indian Peacock); Country-India; Year-2014

Insights: States - Karnataka, Goa and Tamil Nadu have the maximum number of occurrences and the data completeness for 2014 is 100%. For other years, it is lesser, which shows improvement.

Recommended Actions: Many states do not have any occurrences recorded. Mobilize Surveys since Indian Peacock is found in majority of states.

3. Dashboard: Spatial Gap Analysis



kingdom
Animalia

Spatial Gap Analysis

Observation Count
672.2k

Complete Observations
614.1k

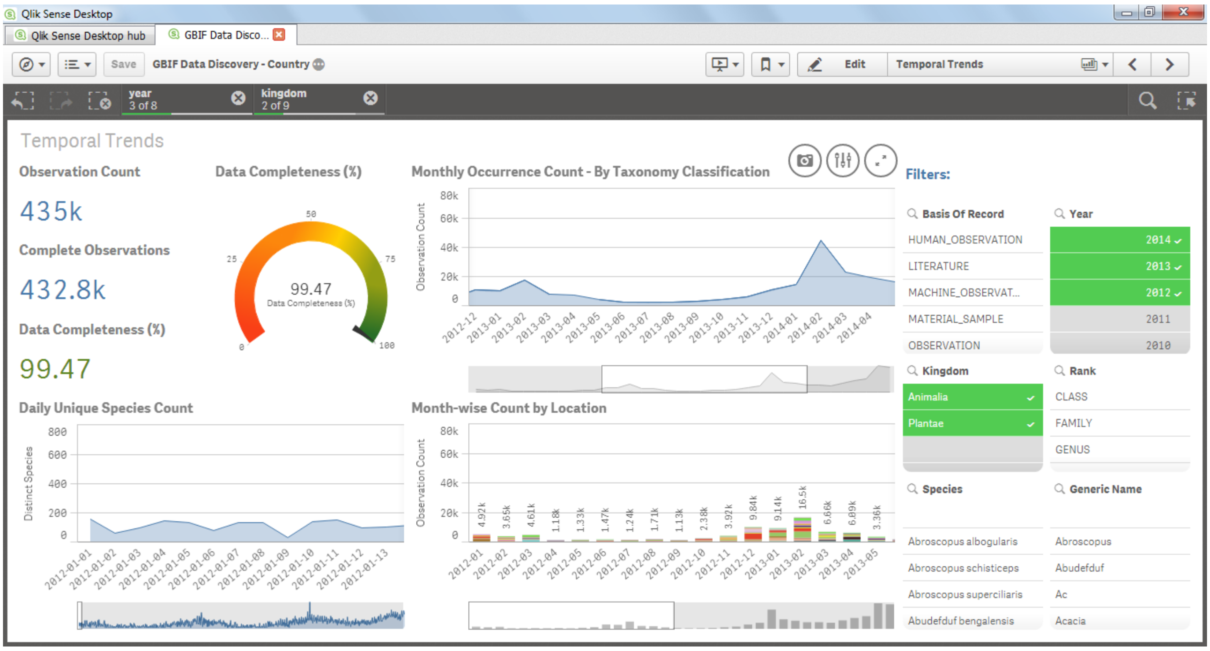
Data Completeness (%)
78.02

Data Analysed: Kingdom-Animalia

Insights: Eastern states have lesser data completeness %. Also, data completeness % reduces as we go up the Latitude band. Also, incorrect co-ordinates & dates are the major issues

Recommended Actions: More efforts should be put into collect efforts correctly in Eastern and Northern parts of India.

4. Dashboard: Temporal Trends



year
3 of 8

kingdom
2 of 9

Temporal Trends

Observation Count
435k

Complete Observations
432.8k

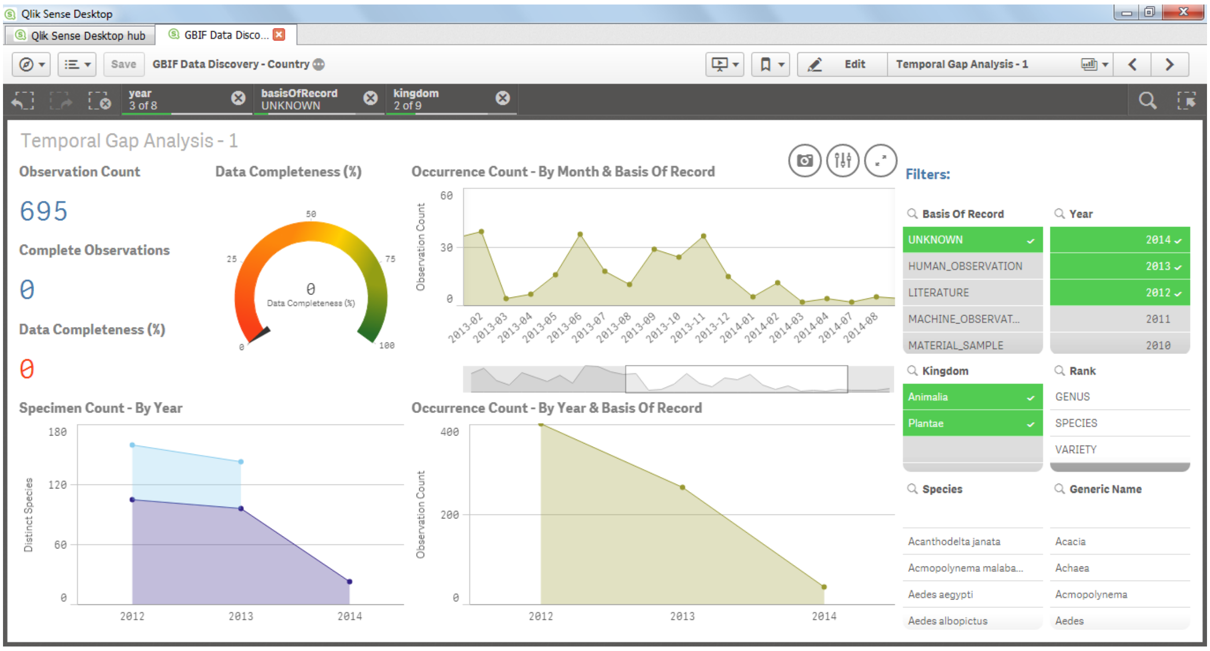
Data Completeness (%)
99.47

Data Analysed: Kingdom-Animalia&Plantae; Years-2012 to 2014

Insights: Occurrence and Daily detected species count have increased over the years. Majority of Records are found in the year 2014

Recommended Actions: Find out what has driven the increase in counts and implement and encourage such techniques going forward

5. Dashboard: Temporal Gap Analysis - 1

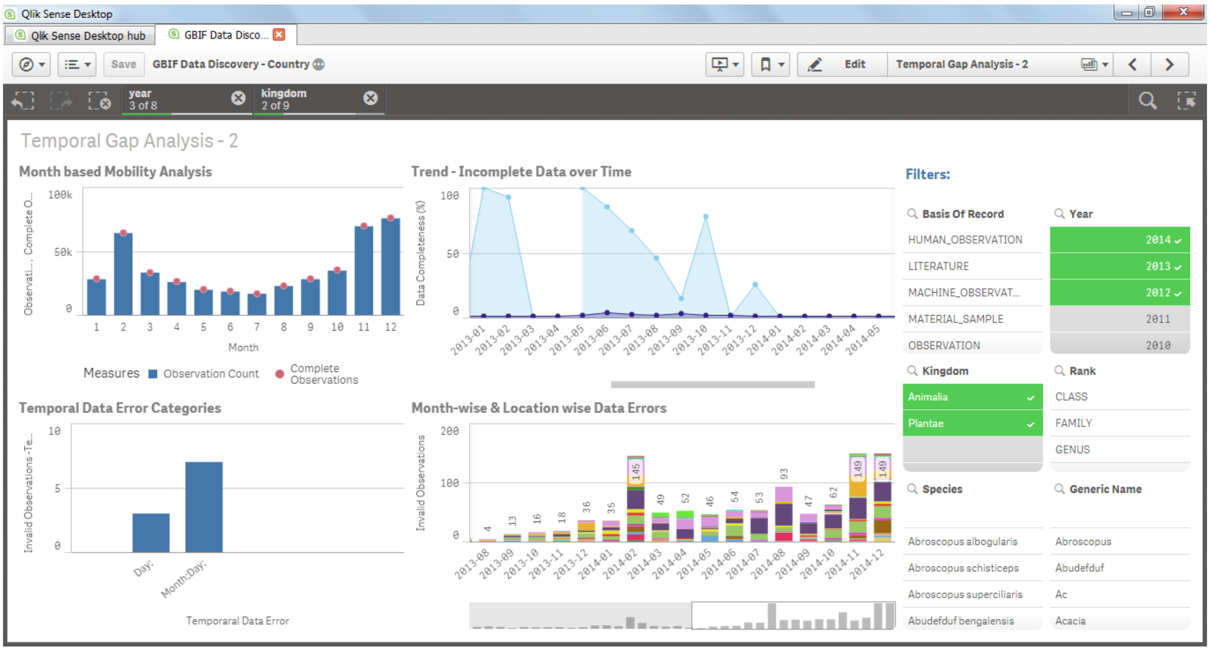

year 3 of 8
basisOfRecord UNKNOWN
kingdom 2 of 9

Data Analysed: Kingdom-Animalia&Plantae; Years-2012 to 2014; Basis of Record-Unknown

Insights: Occurrence details with Basis of Records have reduced over the months and Years in the past 3 years.

Recommended Actions: Find out what has driven the increase in data completeness and implement and encourage such techniques going forward

6. Dashboard: Temporal Gap Analysis - 2

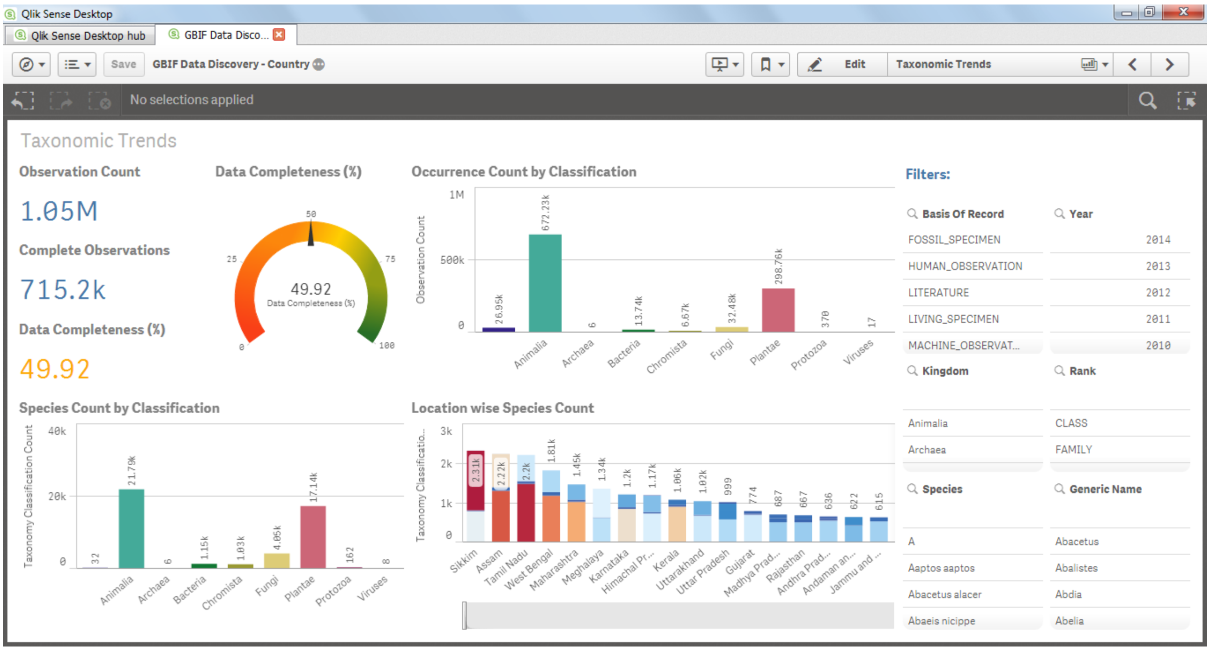

year 3 of 8
kingdom 2 of 9

Data Analysed: Kingdom-Animalia&Plantae; Years-2012 to 2014

Insights: Occurrence Records are lesser during May to Sept, which is Monsoon in India. Also, major cause of data issues are due to incorrectly recorded date and month

Recommended Actions: Mobilize surveys during monsoon to capture species found abundantly during monsoon. Also, make efforts to records month and days of event.

7. Dashboard: Taxonomic Trends



Taxonomic Trends

Observation Count
1.05M

Complete Observations
715.2k

Data Completeness (%)
49.92

Filters:

Basis Of Record	Year
FOSSIL_SPECIMEN	2014
HUMAN_OBSERVATION	2013
LITERATURE	2012
LIVING_SPECIMEN	2011
MACHINE_OBSERVAT...	2010

Kingdom	Rank
Animalia	CLASS
Archaea	FAMILY

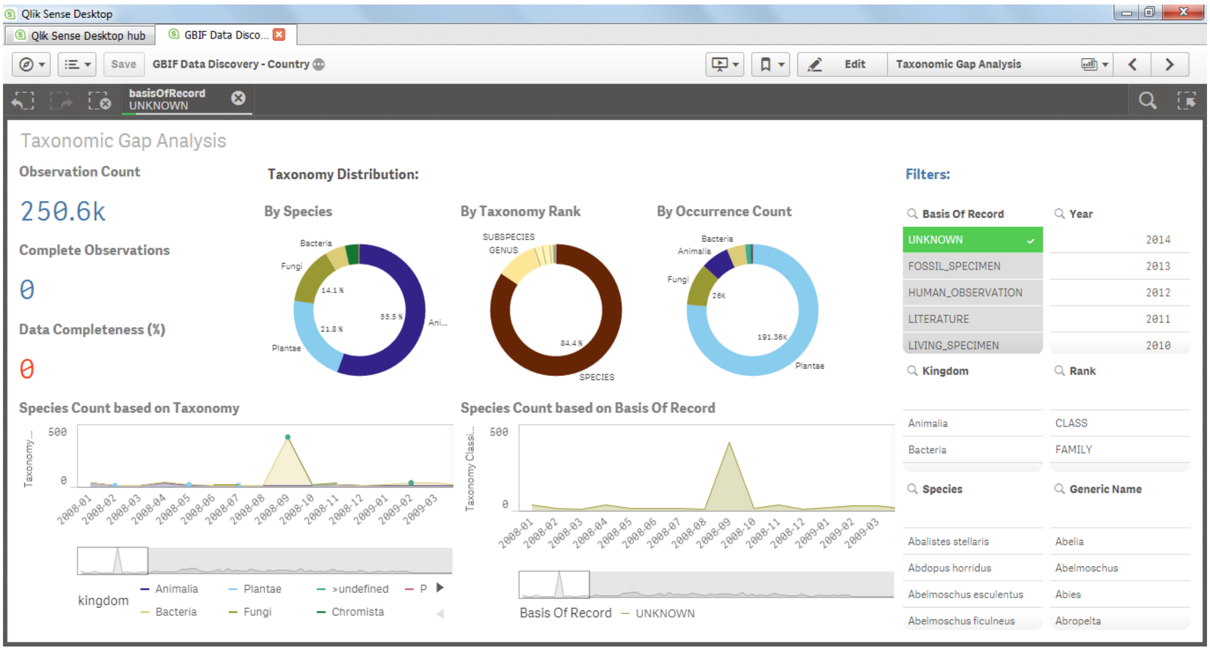
Species	Generic Name
A	Abacetus
Aaptos aaptos	Abalistes
Abacetus alacer	Abdia
Abaeis nicippe	Abelia

Data Analysed: Years-2008 to 2014

Insights: There is a bias on capturing more occurrences for Animalia and Plantae kingdoms. Especially more in Animalia, whereas species count is more or less same for Animalia & Plantae

Recommended Actions: Ensure better coverage of other kingdoms. There is a bias of kingdom collection at State level as well, which needs to be mitigated.

8. Dashboard: Taxonomic Gap Analysis



Taxonomic Gap Analysis

Observation Count
250.6k

Complete Observations
0

Data Completeness (%)
0

Filters:

Basis Of Record	Year
UNKNOWN ✓	2014
FOSSIL_SPECIMEN	2013
HUMAN_OBSERVATION	2012
LITERATURE	2011
LIVING_SPECIMEN	2010

Kingdom	Rank
Animalia	CLASS
Bacteria	FAMILY

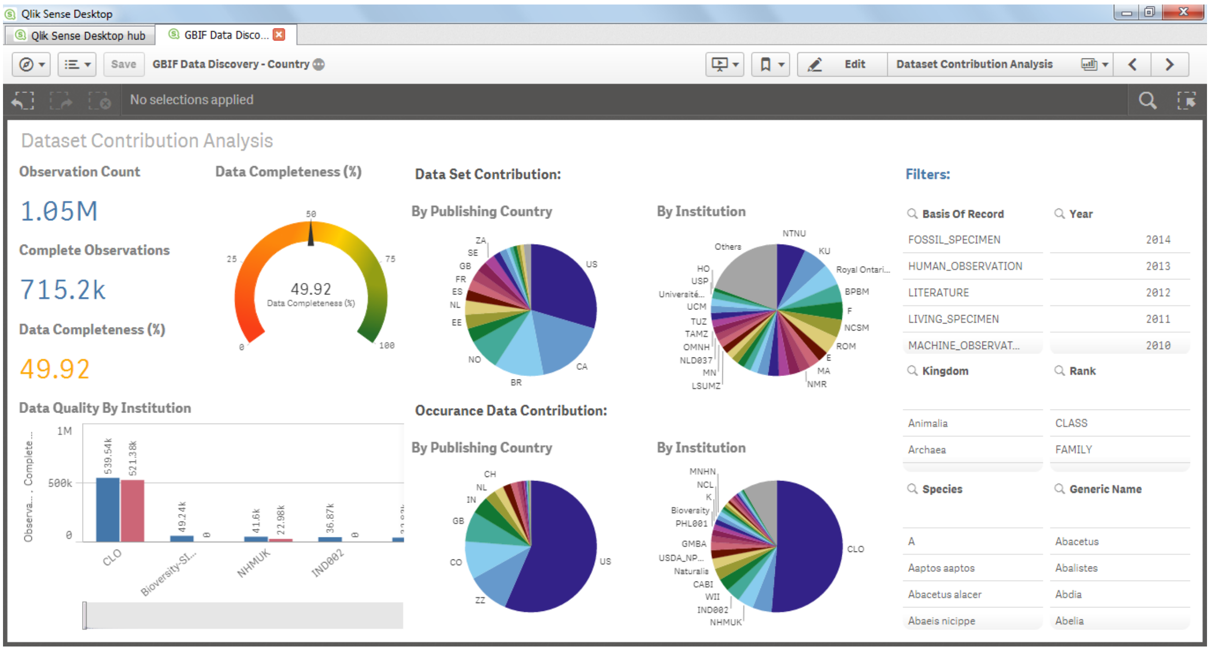
Species	Generic Name
Abalistes stellaris	Abelia
Abdopus horridus	Abelmoschus
Abelmoschus esculentus	Abies
Abelmoschus ficulneus	Abropelta

Data Analysed: Years-2008 to 2014; Basis of Record-Unknown

Insights: Unknown Basis of Record is more in Plantae. The Data Completeness has increased over the period of time with respect to Basis of Records as Unknown.

Recommended Actions: Ensure Plantae surveys are better conducted and recorded. Also determine best practices which have reduced the errors and encourage such best practices.

9. Dashboard: Data Set Contribution Analysis



Qlik Sense Desktop

Qlik Sense Desktop hub | GBIF Data Disco...

Save | GBIF Data Discovery - Country | Edit | Dataset Contribution Analysis

No selections applied

Dataset Contribution Analysis

Observation Count

1.05M

Complete Observations

715.2k

Data Completeness (%)

49.92

Data Quality By Institution

Data Completeness (%)

Data Set Contribution:

By Publishing Country

By Institution

Occurance Data Contribution:

By Publishing Country

By Institution

Filters:

Basis Of Record: FOSSIL_SPECIMEN, HUMAN_OBSERVATION, LITERATURE, LIVING_SPECIMEN, MACHINE_OBSERVAT...

Year: 2014, 2013, 2012, 2011, 2010

Kingdom: Animalia, Archaea

Rank: CLASS, FAMILY

Species: A, Aaptos aaptos, Abacetus alacer, Abaeis nicippe

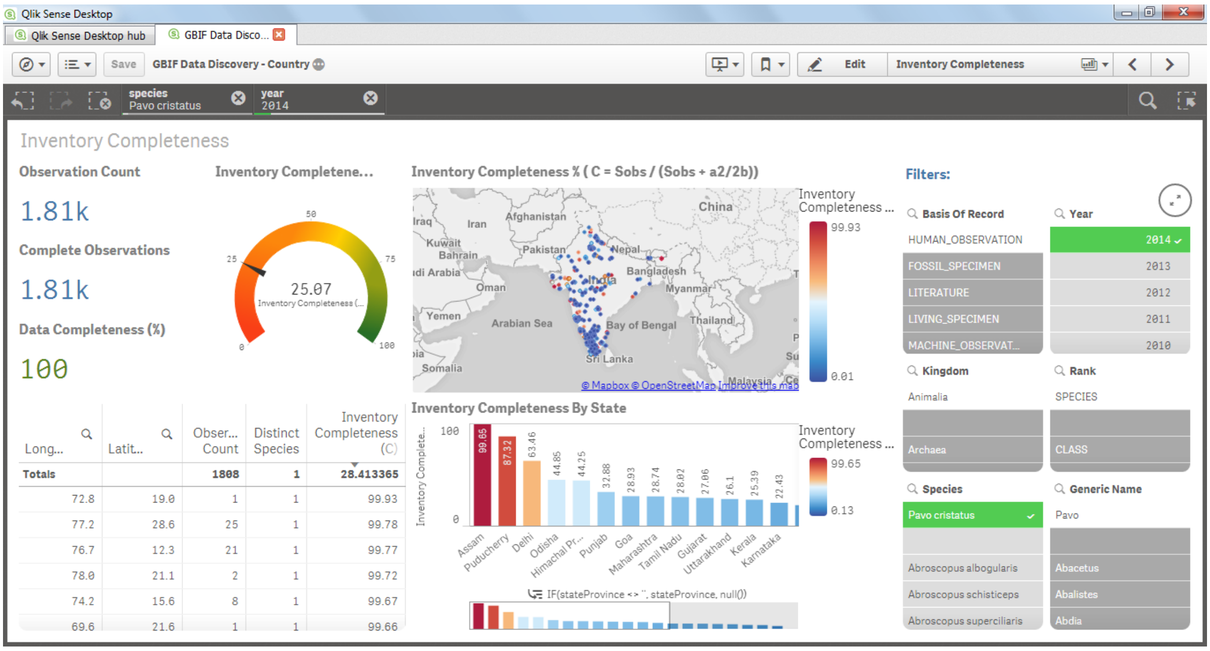
Generic Name: Abacetus, Abalistes, Abdia, Abelia

Data Analysed: Years-2008 to 2014

Insights: Data sets are mainly contributed for India by US and Canada, by Institutions- CLO, NTNU, etc.

Recommended Actions: We can identify which Institutions data set have less data completeness % and they can be asked to take corrective actions

10. Dashboard: Inventory Completeness - Sousa-Baena et al method (2013)



Qlik Sense Desktop

Qlik Sense Desktop hub | GBIF Data Disco...

Save | GBIF Data Discovery - Country | Edit | Inventory Completeness

species Pavo cristatus | year 2014

Inventory Completeness

Observation Count

1.81k

Complete Observations

1.81k

Data Completeness (%)

100

Inventory Completene...

Inventory Completeness % (C = Sobs / (Sobs + a2/2b))

Inventory Completeness By State

Long...	Latit...	Obser... Count	Distinct Species	Inventory Completeness (C)
Totals		1808	1	28.413365
72.8	19.0	1	1	99.93
77.2	28.6	25	1	99.78
76.7	12.3	21	1	99.77
78.0	21.1	2	1	99.72
74.2	15.6	8	1	99.67
69.6	21.6	1	1	99.66

IF(stateProvince <> '', stateProvince, null())

Filters:

Basis Of Record: HUMAN_OBSERVATION, FOSSIL_SPECIMEN, LITERATURE, LIVING_SPECIMEN, MACHINE_OBSERVAT...

Year: 2014 ✓, 2013, 2012, 2011, 2010

Kingdom: Animalia, Archaea

Rank: SPECIES, CLASS

Species: Pavo cristatus ✓, Abroscopus albogularis, Abroscopus schisticeps, Abroscopus superciliaris

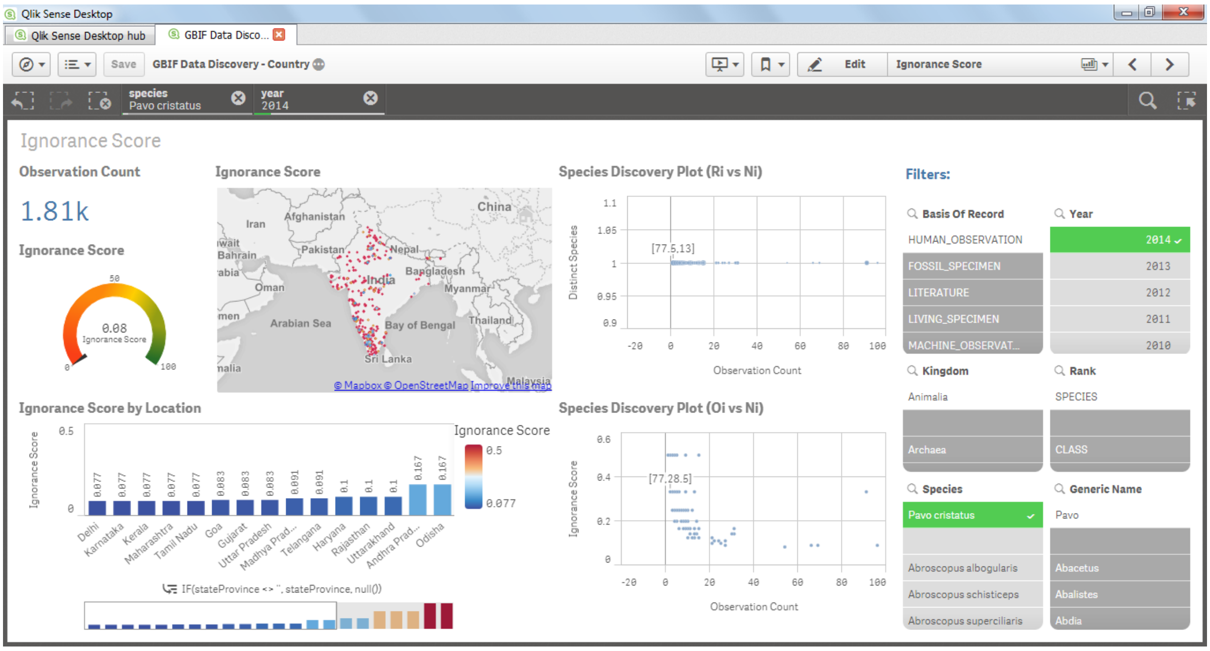
Generic Name: Pavo, Abacetus, Abalistes, Abdia

Data Analysed: Species-Pavo Cristatus (Indian Peacock); Year-2014

Insights: Inventory completeness is high in states like Assam, Pudduchery & Delhi, but lower in Central Part of India

Recommended Actions: Locations where Inventory Completeness is less can be focused upon to take up targeted surveys and improve data collection

11. Dashboard: Ignorance Score - Mair L & Ruete A (2016) method (2013)



species Pavo cristatus

year 2014

Data Analysed: Species-Pavo Cristatus (Indian Peacock); Year-2014

Insights: Ignorance score is lower in south-western parts of India. It is higher in Assam and Himachal Pradresh

Recommended Actions: Locations where Ignorance Score is High are under-sampled and so such areas can be focused upon to take up targeted surveys and improve sampling

Reference

- 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. doi:10.1111/ddi.12136
- Mair L & Ruete A (2016) Explaining Spatial Variation in the Recording Effort of Citizen Science Data across Multiple Taxa. PLoS ONE 11(1): e0147796. doi:10.1371/journal.pone.0147796
- 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. doi:10.1111/geb.12468