A primer for the capture, processing and analysis of biodiversity data

Day 1

ENDANGERED
WILDLIFE TRUST

Dr. Lizanne Roxburgh and Dr. Dominic Henry
Conservation Science Unit, Endangered Wildlife Trust

Day 1

Start	End		Session topic
08h30	10h00	1	Introduction to GBIF, data sharing and metadata; practical exercises
10h00	10h30		Tea break
10h30	12h15	2	Introduction to Data standards and DarwinCore; practical exercise on DwC
12h15	13h30		Lunch
13h30	15h00	3	Practical exercise in data cleaning and standardising using OpenRefine
15h00	15h30		Tea break
15h30	17h00	4	Developing a data collection app

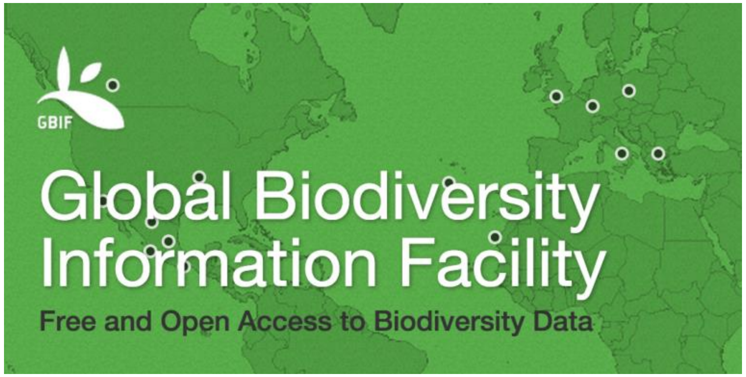
Foundational Biodiversity Information Programme (or FBIP)

science
& technology
Department:
Science and Technology
REPUBLIC OF SOUTH AFRICA

- FBIP is “*a long-term programme to* ***generate, manage and disseminate foundational biodiversity information*** *and knowledge to* ***improve decision-making****, service delivery and create new economic opportunities*”.
- First two sessions today are about good data management practices, and how your data can be shared. How can you make your data useful beyond just your Postdoc/ PhD/ MSc study?

WHAT IS GBIF?

- An **international network** and research infrastructure funded by the world’s governments and aimed at providing anyone, anywhere, **open access to data** about all types of life on Earth.


GBIF
Global Biodiversity Information Facility
Free and Open Access to Biodiversity Data

- It is coordinated through its Secretariat, which is based in Denmark, and it provides participating countries and organizations around the world with **common standards** and **open-source tools** that enable them to share information about where and when species have been recorded.

THE ORIGINS OF GBIF

Organisation for Economic Co-operation and Development

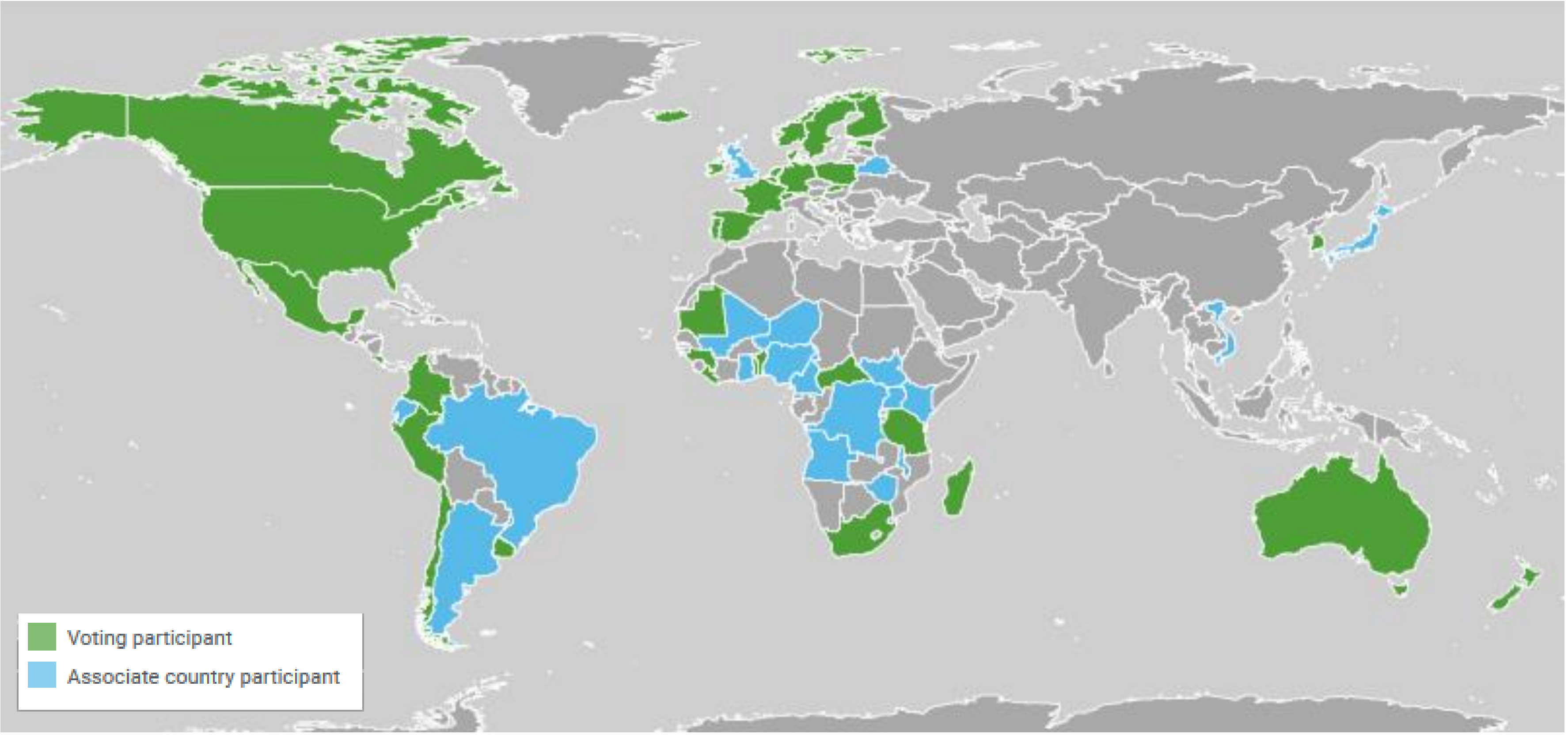
The OECD promotes policies that will improve the economic and social well-being of people around the world. The OECD provides a forum in which governments can work together to share experiences and seek solutions to common problems.

In 1999, a recommendation of the OECD Megascience Forum was that:

"An international mechanism is needed to make biodiversity data and information accessible worldwide"

Following this recommendation, GBIF was established two years later, in 2001.

THE GBIF NETWORK TODAY



- Participation in GBIF is through ‘nodes’ that coordinate data mobilization from national institutions/networks (in South Africa, it is SANBI - GBIF)

https://www.gbif.org/the-gbif-network

WHAT IS DATA MOBILIZATION?

Mobilization – making data available for use/ re-use by others.

*In the past, this meant publishing a print document that could be read by many people. Today, mobilization usually means making something available on the web – truly mobile data is published to a **freely accessible web site**.*

Once mobilized, data can be downloaded and imported for use in other analyses, possibly in combination with similar data from other data sources, possibly combined with other kinds of data so they can be used to create new insights.

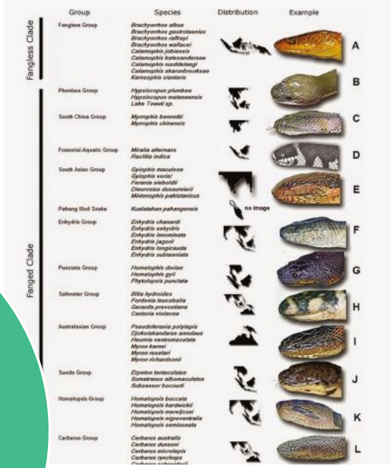
By encouraging and helping institutions to publish data according to common standards, **GBIF** enables research not possible before, and informs better decisions to conserve and sustainably use the biological resources of the planet.

What is Biodiversity Data?

Biodiversity data – bits of information about different kinds of organisms that have been observed somewhere in space and time.

Occurrence data

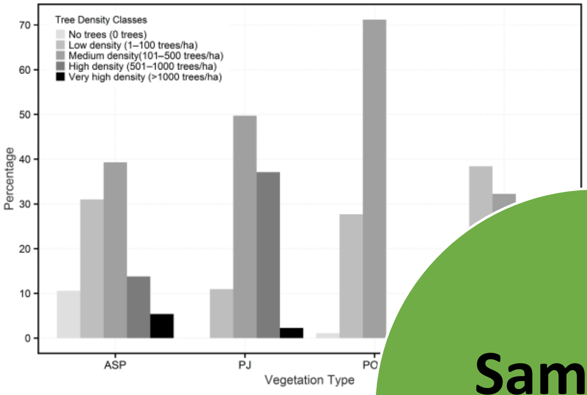
Individuals of 1 species occurring in a place and time



Group
Species
Distribution
Example
Fangless Clade
Fanged Clade
Fangless Group
Plumbea Group
South China Group
Fossorial-Aquatic Group
South Asian Group
Pahang Mud Snake
Enhydris Group
Punctata Group
Saltwater Group
Australasian Group
Sunda Group
Homalopsis Group
Cerberus Group
no image
A
B
C
D
E
F
G
H
I
J
K
L

Checklists

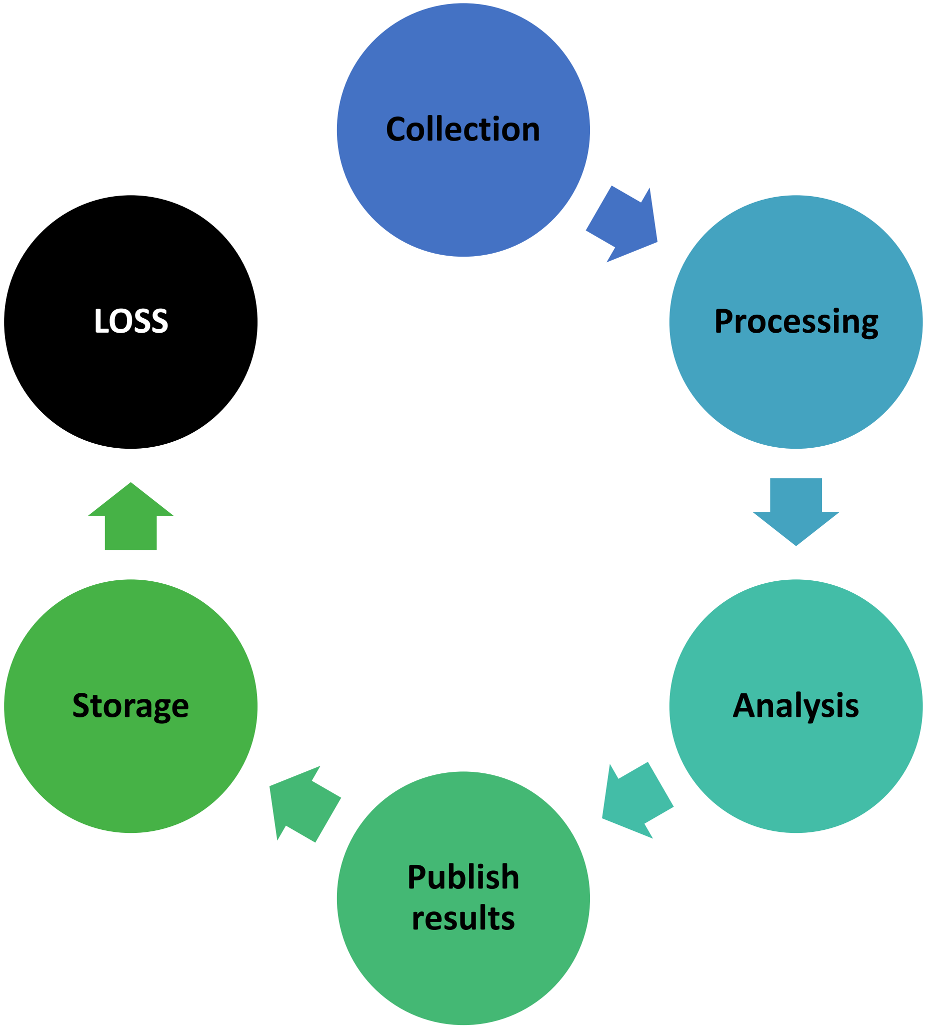
A list of species occurring in a particular place



Sampling events

Individuals of 1 species occurring in a place and time, recorded with a particular effort

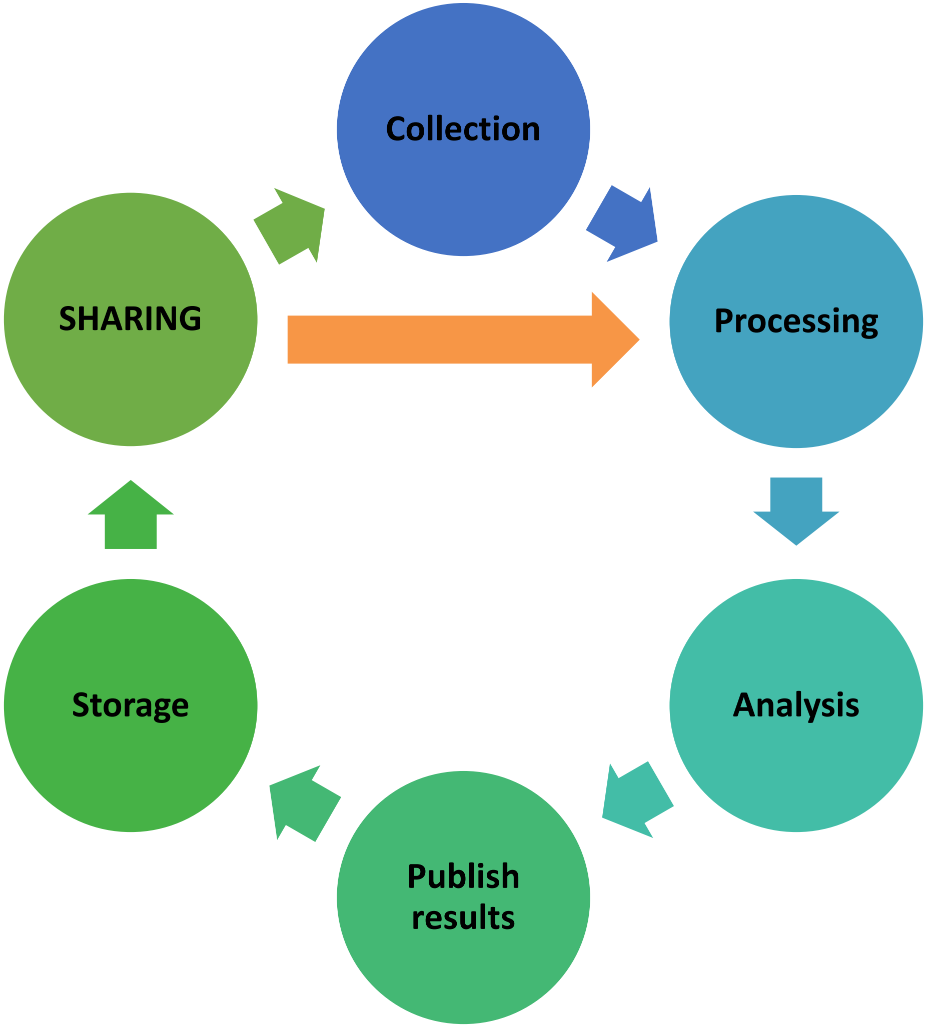
Biodiversity Data Life Cycle



Typically, biodiversity data in a project would go through the cycle of collection, processing, analysis, publication of analysed results, storage on a computer or hard drive, and ultimately loss after some years or decades

ENDANGERED WILDLIFE TRUST
Protecting forever, together.

Biodiversity Data Life Cycle

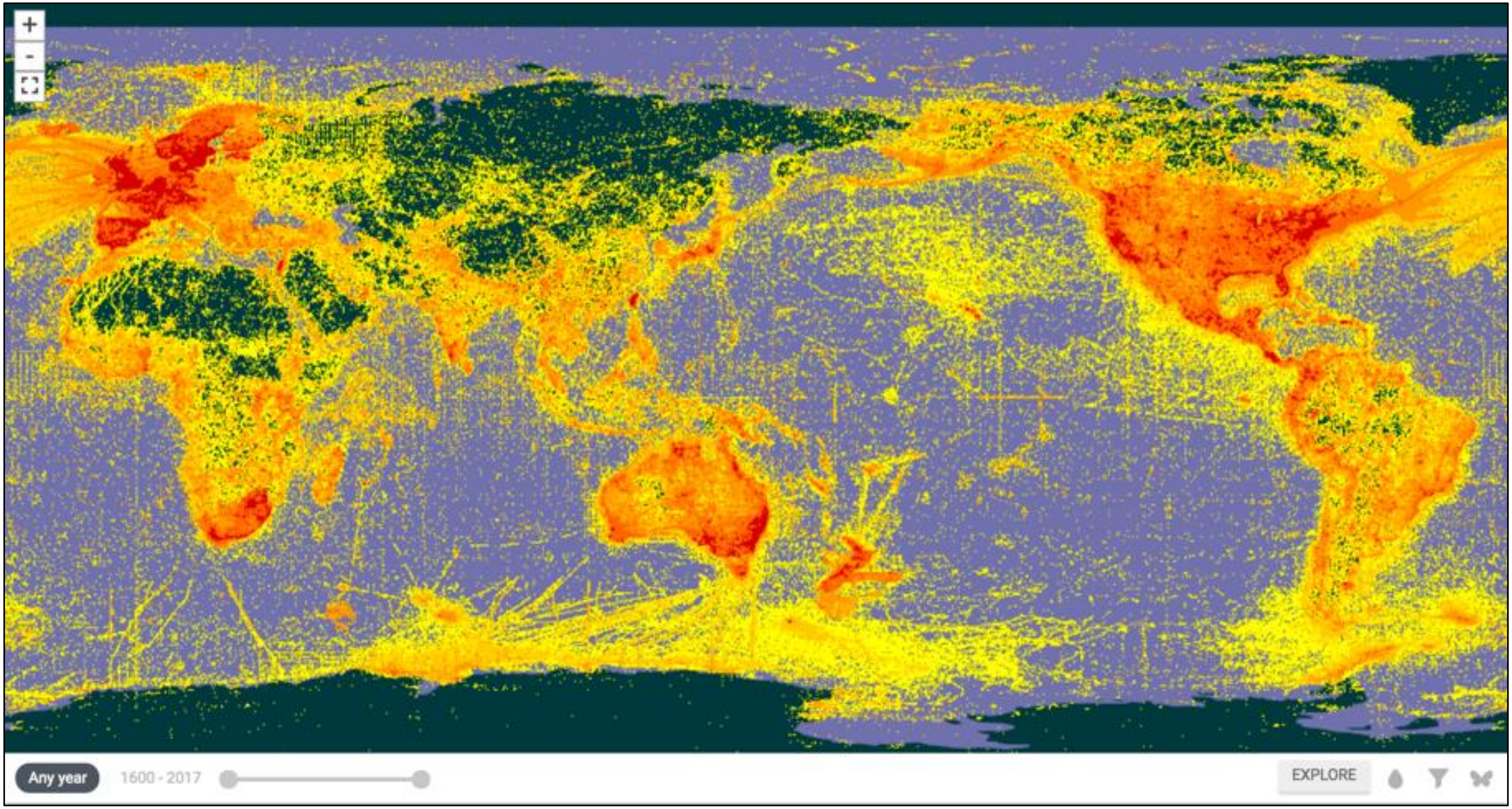


However, a much preferable lifecycle would not end with data loss, but rather with data sharing. These data could then be combined with other data sources, and re-enter the cycle of processing, analysis, publication of analysed results, and further sharing.

GBIF is arguably one of the most important avenues for sharing and reuse of biodiversity data.

ENDANGERED WILDLIFE TRUST
Protecting forever, together.

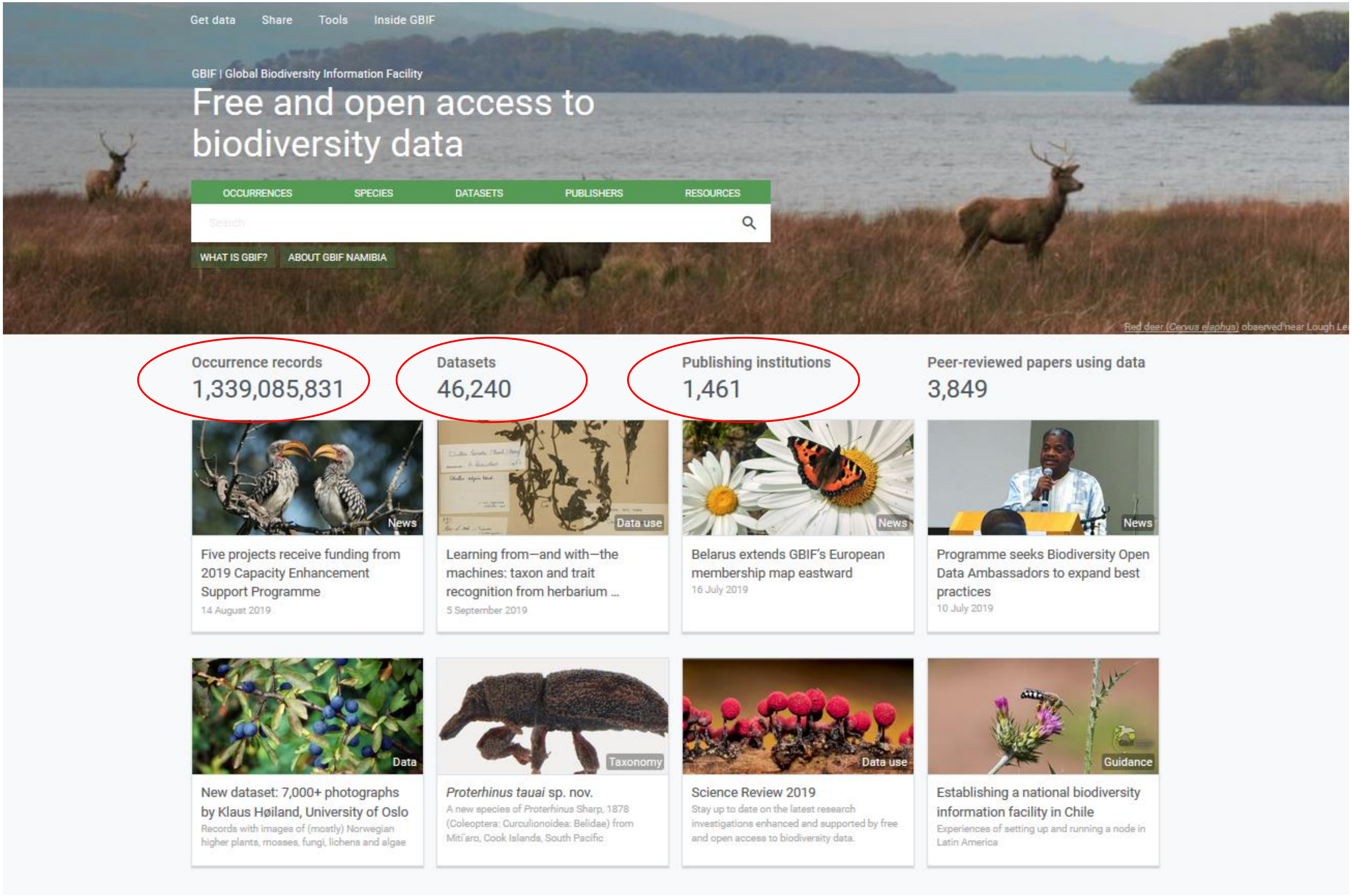
GBIF.ORG - DATA DISTRIBUTION



This map represents the biodiversity data available on GBIF. Each dot represents evidence of species occurrence with standardized information on what was observed, where, by whom, when, and based on what evidence?

www.gbif.org/occurrence

The GBIF website – www.gbif.org



Get data Share Tools Inside GBIF

GBIF | Global Biodiversity Information Facility

Free and open access to biodiversity data

OCCURRENCES SPECIES DATASETS PUBLISHERS RESOURCES

Search

WHAT IS GBIF? ABOUT GBIF NAMIBIA

Red deer (*Cervus elaphus*) observed near Lough Le

Occurrence records	Datasets	Publishing institutions	Peer-reviewed papers using data
1,339,085,831	46,240	1,461	3,849

News
Five projects receive funding from 2019 Capacity Enhancement Support Programme
14 August 2019

Data use
Learning from—and with—the machines: taxon and trait recognition from herbarium ...
5 September 2019

News
Belarus extends GBIF's European membership map eastward
16 July 2019

News
Programme seeks Biodiversity Open Data Ambassadors to expand best practices
10 July 2019

Data
New dataset: 7,000+ photographs by Klaus Høiland, University of Oslo
Records with images of (mostly) Norwegian higher plants, mosses, fungi, lichens and algae

Taxonomy
Proterhinus tauai sp. nov.
A new species of *Proterhinus* Sharp, 1878 (Coleoptera: Curculionoidea: Belidae) from Miti'aro, Cook Islands, South Pacific

Data use
Science Review 2019
Stay up to date on the latest research investigations enhanced and supported by free and open access to biodiversity data.

Guidance
Establishing a national biodiversity information facility in Chile
Experiences of setting up and running a node in Latin America

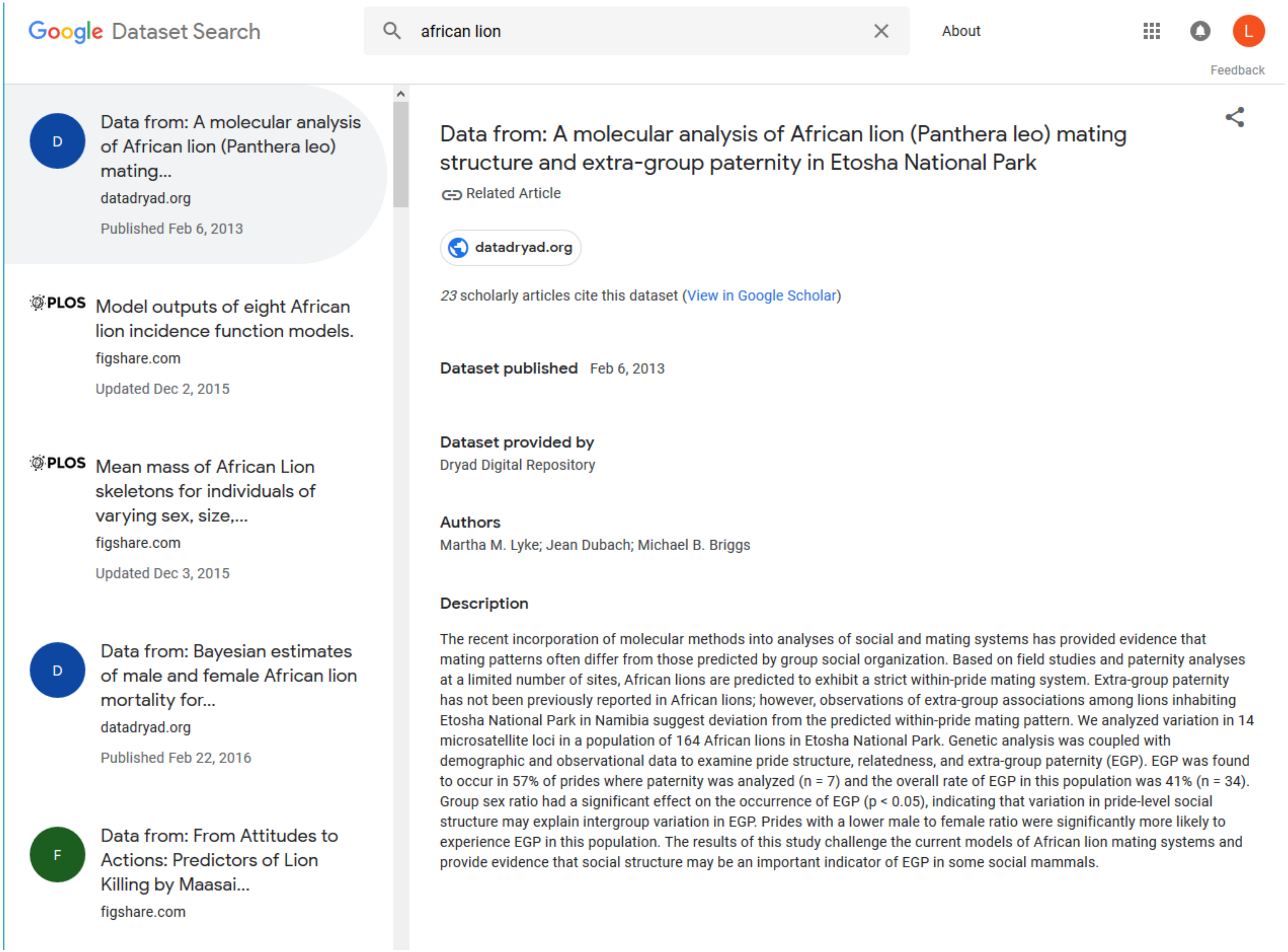
Where else can data be shared?

A sample of online data-sharing platforms:

- GBIF – for organismal occurrence data
- Dryad – for article-related data
- Encyclopedia of Life – has a page for each species
- DataONE - Data Observation Network for Earth – environmental science data
- figshare – for sharing of institutional scientific data and outputs

Google Dataset search
– provides a single place to search all data platforms

https://toolbox.google.com/datasetsearch



Data from: A molecular analysis of African lion (Panthera leo) mating...
datadryad.org
Published Feb 6, 2013

Model outputs of eight African lion incidence function models.
figshare.com
Updated Dec 2, 2015

Mean mass of African Lion skeletons for individuals of varying sex, size,...
figshare.com
Updated Dec 3, 2015

Data from: Bayesian estimates of male and female African lion mortality for...
datadryad.org
Published Feb 22, 2016

Data from: From Attitudes to Actions: Predictors of Lion Killing by Maasai...
figshare.com

Data from: A molecular analysis of African lion (Panthera leo) mating structure and extra-group paternity in Etosha National Park

Related Article

datadryad.org

23 scholarly articles cite this dataset (View in Google Scholar)

Dataset published Feb 6, 2013

Dataset provided by
Dryad Digital Repository

Authors
Martha M. Lyke; Jean Dubach; Michael B. Briggs

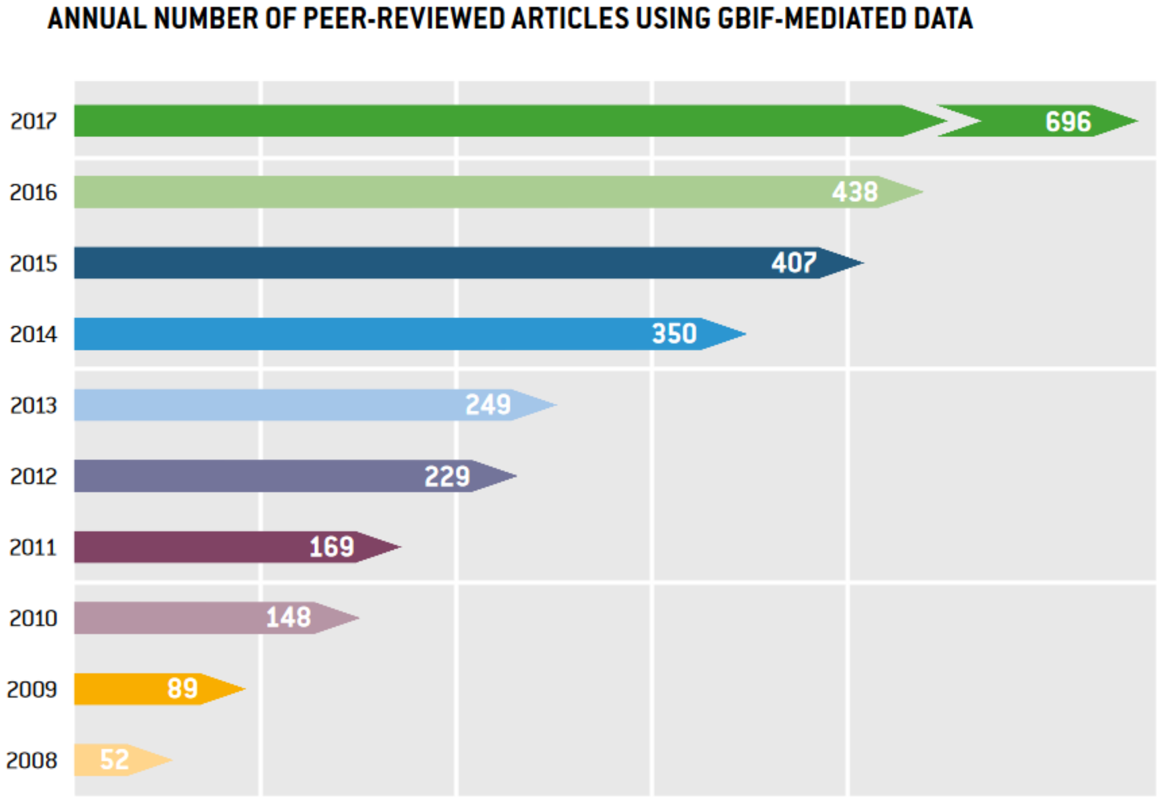
Description

The recent incorporation of molecular methods into analyses of social and mating systems has provided evidence that mating patterns often differ from those predicted by group social organization. Based on field studies and paternity analyses at a limited number of sites, African lions are predicted to exhibit a strict within-pride mating system. Extra-group paternity has not been previously reported in African lions; however, observations of extra-group associations among lions inhabiting Etosha National Park in Namibia suggest deviation from the predicted within-pride mating pattern. We analyzed variation in 14 microsatellite loci in a population of 164 African lions in Etosha National Park. Genetic analysis was coupled with demographic and observational data to examine pride structure, relatedness, and extra-group paternity (EGP). EGP was found to occur in 57% of prides where paternity was analyzed (n = 7) and the overall rate of EGP in this population was 41% (n = 34). Group sex ratio had a significant effect on the occurrence of EGP ($p < 0.05$), indicating that variation in pride-level social structure may explain intergroup variation in EGP. Prides with a lower male to female ratio were significantly more likely to experience EGP in this population. The results of this study challenge the current models of African lion mating systems and provide evidence that social structure may be an important indicator of EGP in some social mammals.

Why share data?

"Data reuse is the fundamental goal of data sharing"

Examples of reuses: Red Lists, climate change modelling, national biodiversity assessments, identification of critical biodiversity areas, environmental impact assessments, alien invasive species research

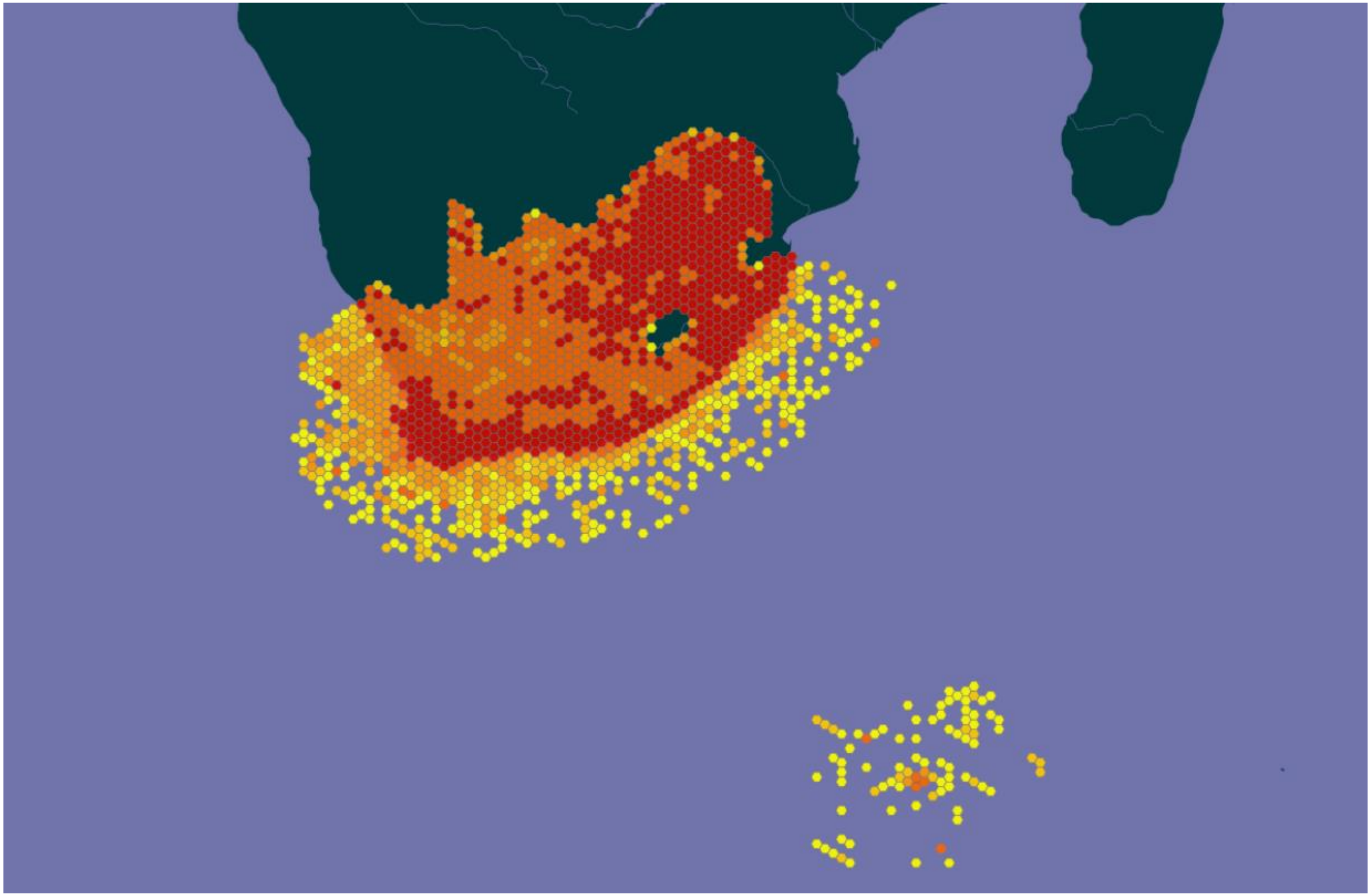


GBIF Secretariat. (2018). GBIF Science Review 2018. https://doi.org/10.15468/VA9B-3048

ENDANGERED WILDLIFE TRUST
Protecting forever, together.

GBIF data for South Africa

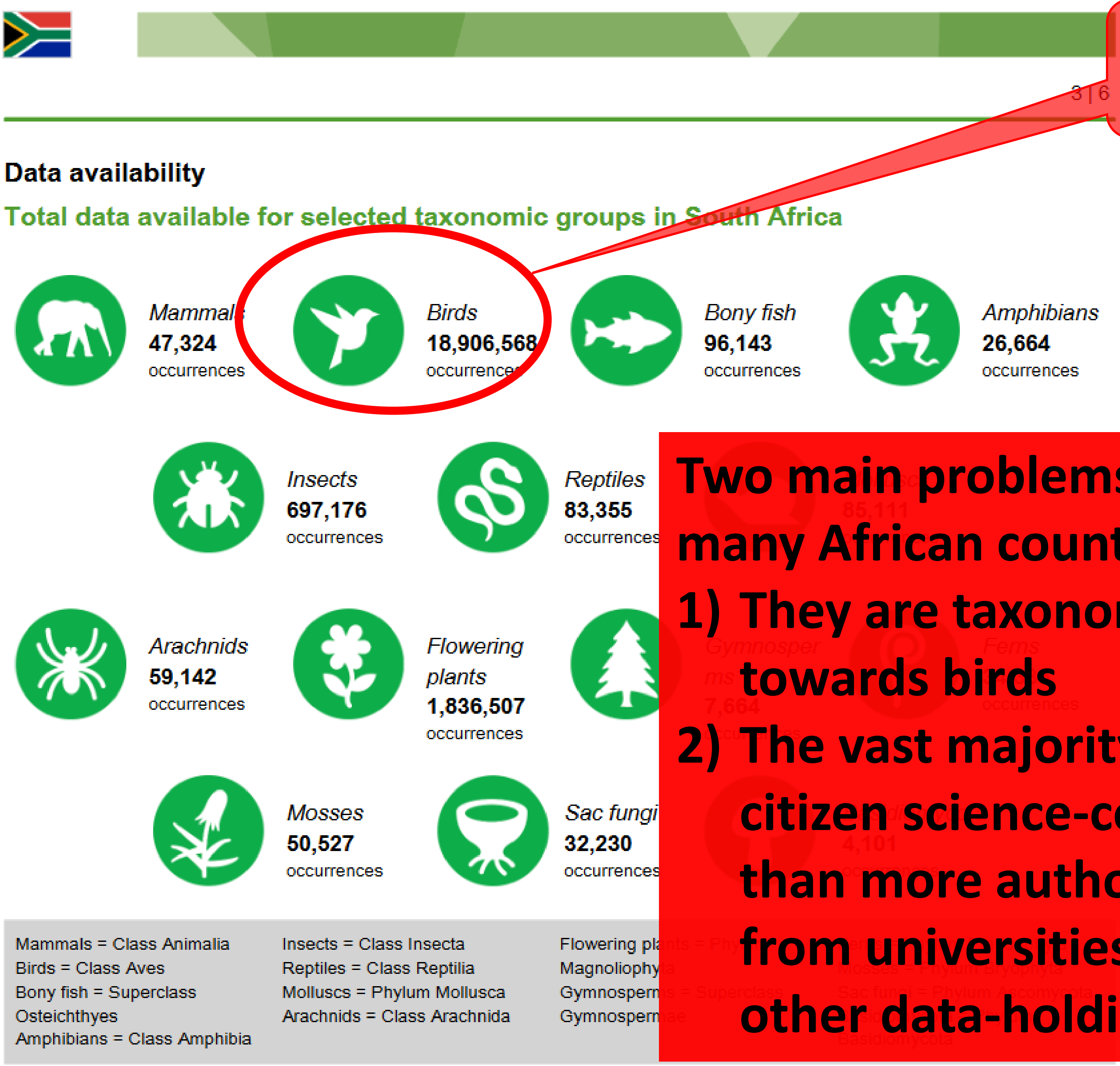
Total number of occurrences published for South Africa: 22,320,629



GBIF
Global Biodiversity
Information Facility

GBIF data for South Africa

Most occurrence data is of birds



3 | 6

Data availability

Total data available for selected taxonomic groups in South Africa

Mammals **47,324** occurrences

Birds **18,906,568** occurrences

Bony fish **96,143** occurrences

Amphibians **26,664** occurrences

Insects **697,176** occurrences

Reptiles **83,355** occurrences

Arachnids **59,142** occurrences

Flowering plants **1,836,507** occurrences

Mosses **50,527** occurrences

Sac fungi **32,230** occurrences

Mammals = Class Animalia
Birds = Class Aves
Bony fish = Superclass Osteichthyes
Amphibians = Class Amphibia

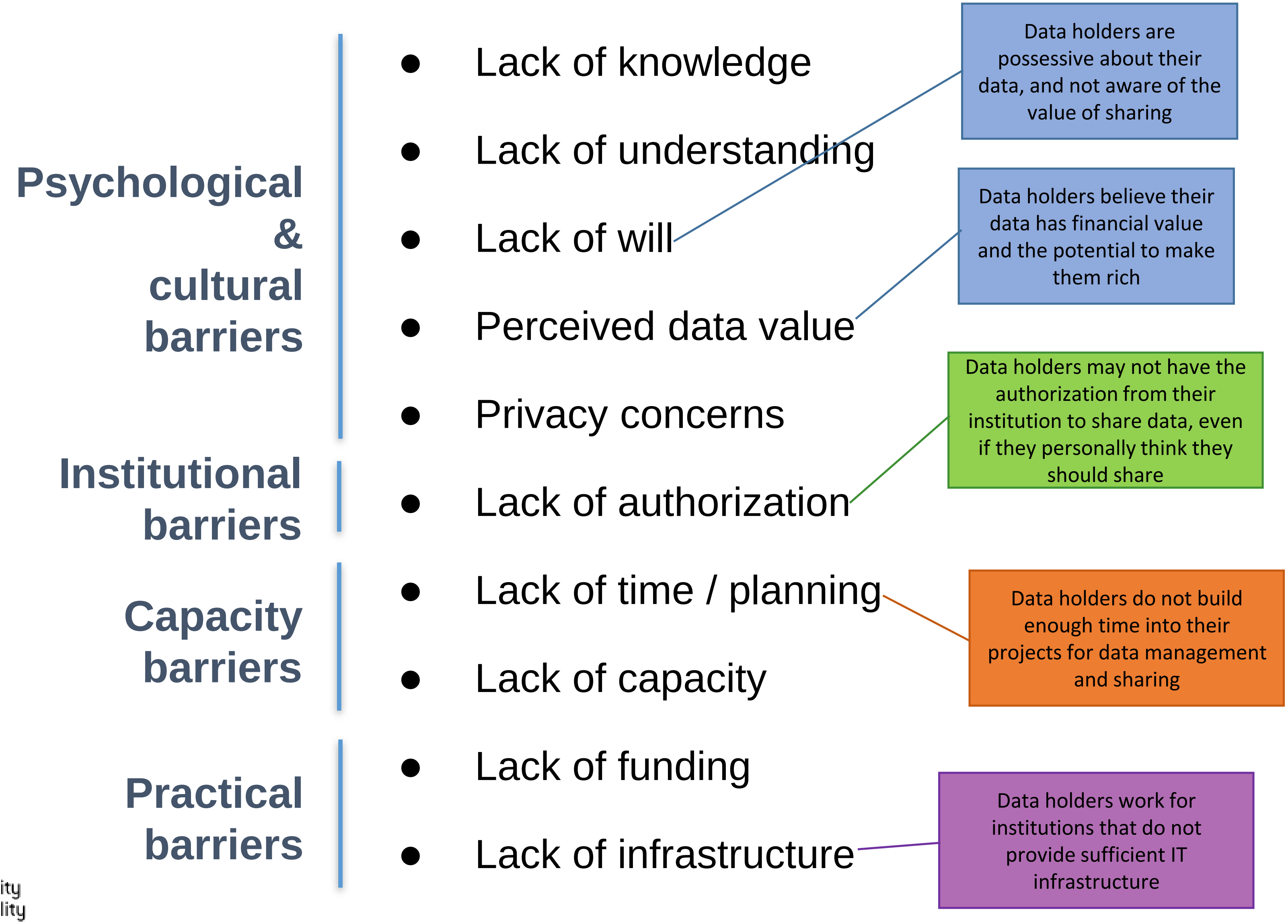
Insects = Class Insecta
Reptiles = Class Reptilia
Molluscs = Phylum Mollusca
Arachnids = Class Arachnida

Flowering plants = Phy
Magnoliophy
Gymnosperms = Superclass
Gymnospermae

Two main problems with data from many African countries:

1) **They are taxonomically skewed towards birds**
2) **The vast majority of the data is citizen science-collected, rather than more authoritative data from universities, museums and other data-holding institutions**

Why doesn't everyone share their data?
Examples of barriers to data publishing



Psychological & cultural barriers

- Lack of knowledge
- Lack of understanding
- Lack of will
- Perceived data value
- Privacy concerns

Institutional barriers

- Lack of authorization

Capacity barriers

- Lack of time / planning
- Lack of capacity

Practical barriers

- Lack of funding
- Lack of infrastructure

Data holders are possessive about their data, and not aware of the value of sharing

Data holders believe their data has financial value and the potential to make them rich

Data holders may not have the authorization from their institution to share data, even if they personally think they should share

Data holders do not build enough time into their projects for data management and sharing

Data holders work for institutions that do not provide sufficient IT infrastructure

GBIF
Global Biodiversity
Information Facility

Why publish your data online?

- Leaving a legacy, allowing future researchers to reuse your data and acknowledge your contribution to science
- This is the future: research funders and journal publishers will no longer allow you to hoard your data; there are opportunities for you to make use of others’ data and collaborate on research projects

Incentives for publishing data

GBIF promotes a culture in which people recognize the benefits of publishing open-access biodiversity data

By publishing data, you will

- **contribute to global knowledge about biodiversity**, and thus to the solutions that will promote its conservation and sustainable use.
- **reveal new opportunities for collaboration** with other data owners and researchers.
- be **properly credited** for your work to create and curate biodiversity data.
- **gain/ maintain access to funding**, as some funding agencies now require researchers receiving public funds to make data freely accessible at the end of a project.
- be able to **trace the usage and citations** of your published data (through GBIF).

GBIF
Global Biodiversity
Information Facility

How can you publish your data?

- Prepare **metadata** to accompany the dataset *(Metadata is data about data)*
- Prepare your data according to **DarwinCore standards** in an Excel spreadsheet (or other data management software)
- *Contact SANBI GBIF (the South African GBIF node) (or the EWT) about publishing the data to GBIF*

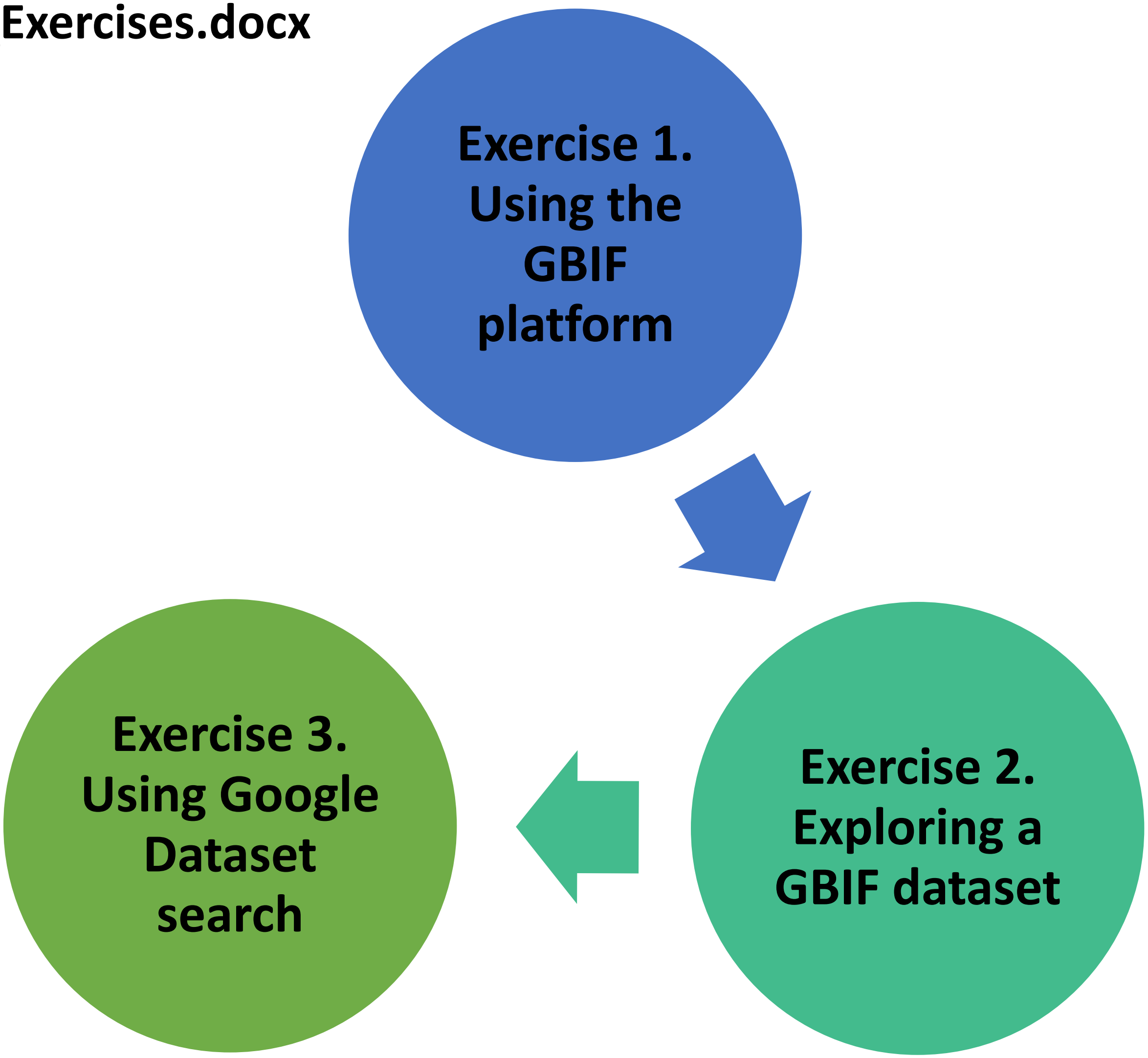
GBIF
Global Biodiversity
Information Facility

Day 1

Start	End		Session topic
08h30	10h00	1	Introduction to GBIF, data sharing and metadata; practical exercises
10h00	10h30		Tea break
10h30	12h15	2	Introduction to Data standards and DarwinCore; practical exercise on DwC
12h15	13h30		Lunch
13h30	15h00	3	Practical exercise in data cleaning and standardising using OpenRefine
15h00	15h30		Tea break
15h30	17h00	4	Developing a data collection app

Practical exercises

File: Day1_GBIF_Exercises.docx



Exercise 1. Using the GBIF platform

Exercise 2. Exploring a GBIF dataset

Exercise 3. Using Google Dataset search

ENDANGERED WILDLIFE TRUST
Protecting forever, together.