FBIP Data Challenges: The Need to Address Quality Standards

By: Mahlatse M. Kgatla Email: M.Kgatla@SANBI.org.za

Joint SANBI Biodiversity Information Management & Foundational Biodiversity Information Programme **FORUM** 13–16 August 2018

SANBI Biodiversity for Life
South African National Biodiversity Institute

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

NRF National Research Foundation

Introduction

- FBIP funds projects that generate primary/foundational biodiversity data, including occurrence data.
- Occurrence data has three core descriptors (fields):
  - Taxon ID (what) – identity of organism
  - Locality (where) – exact area where organism was collected/observed
  - Collection date (when) – day/month/year (sometimes even time) when the organism was collected /observed

Introduction

- This data can then be easily packaged using the Darwin Core System.
- Darwin Core in simple terms is a system of defined fields and is used by the GBIF and many other organizations/researchers.
- Darwin Core helps standardize data from different types of projects and allows data sharing.
- FBIP created a simple data and a metadata template using Darwin Core Standards, which grant holders are required to use to capture data.

FBIP Data and Metadata Template

Metadata Sheet Template

Basics

Title of data set	
Taxonomic group covered	
Species / specimen information	
No. of records	
Last updated	
Description (explain what the data set represents)	

Resource Owner

Data set Owner	
Organisation	
Contact person	
Position	
Phone	
Email	
Homepage	
Address	

Institution ID	Collection ID	Dataset ID	Basis of Record	Catalogue Number	Occurrence Remarks	Record Number	Recorded By	Individual ID	Individual Count	Life Stage	Reproductive Condition	Establishment Means	Preparations	Disposition	Associated Reference	Associated Sequences
An identifier for the institution having custody of the object(s) or information referred to in the record.	An identifier for the collection or dataset from which the record was derived.	An identifier for the set of data.	eg. "PreservedSpecimen", "FossilSpecimen", "LivingSpecimen", "HumanObservation",		Example: "found dead on road".	Often serves as a link between field notes and an Occurrence record, such as a specimen collector's number.	Example: "Oliver P. Pearson; Anita K. Pearson"	An identifier for an individual or named group of individual organisms represented in the Occurrence. Meant to accommodate resampling of the same individual or group for monitoring purposes.	The number of individuals represented present at the time of the Occurrence.			Examples: "native", "introduced", "naturalised", "invasive", "managed".	Examples: "skin; skull; skeleton", "whole animal (ETOH); tissue (EDTA)", "fossil", "cast", "photograph", "DNA extract".	Examples: "in collection", "missing", "voucher elsewhere", "duplicates elsewhere".	literature associated with the Occurrence.	identifiers (publication, global unique identifier, URI) of genetic sequence information associated with the Occurrence.

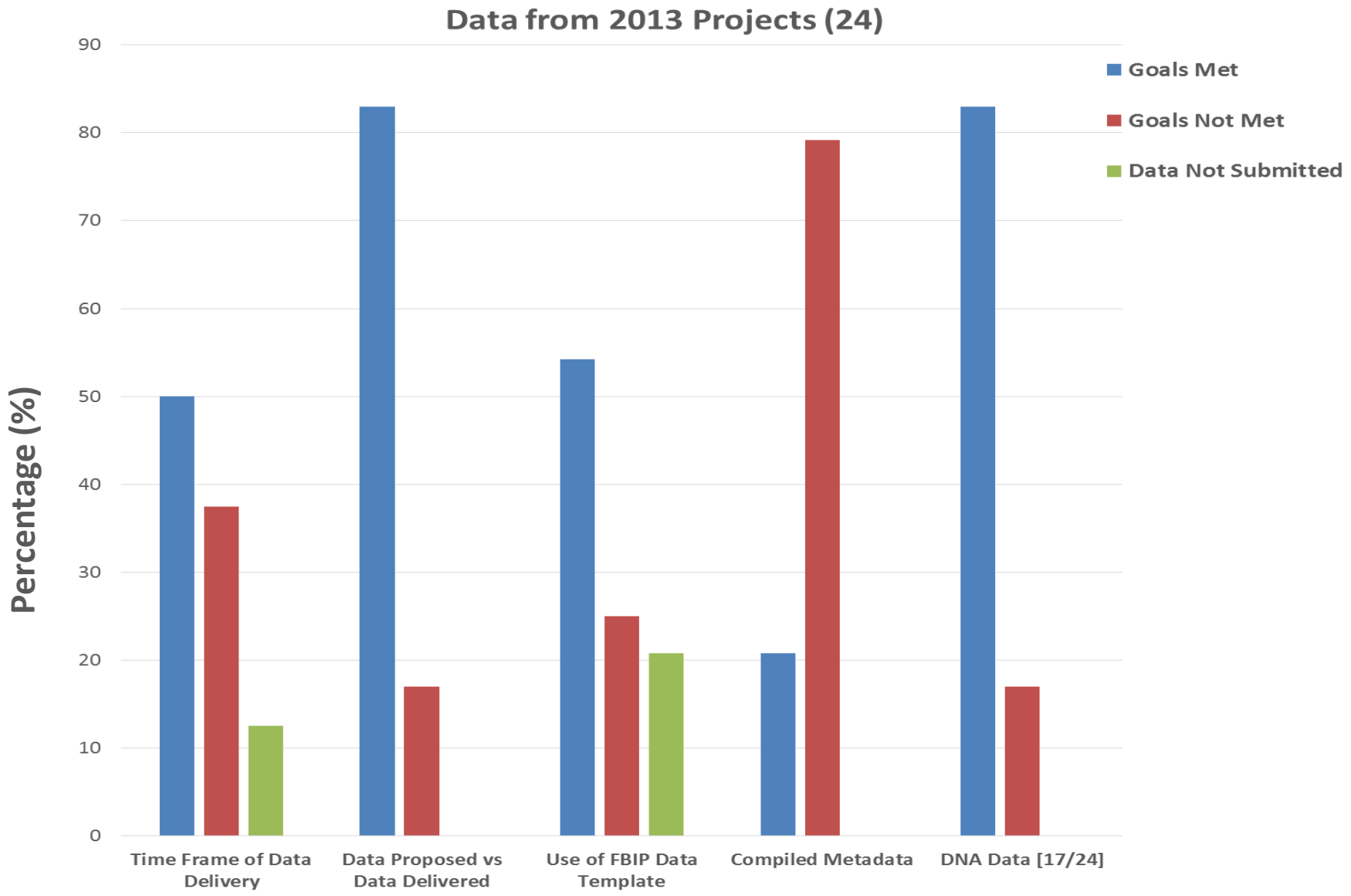
Challenges

- Two main types of data challenges
  - FBIP related challenges
  - General data challenges

FBIP Related Challenges

- Time frame of data delivery
- Data proposed vs data delivered
- Use of FBIP data template
- Compiled Metadata
- DNA data – not submitted to BOLD, and / or no accession numbers for record in BOLD / Genbank

FBIP Related Challenges



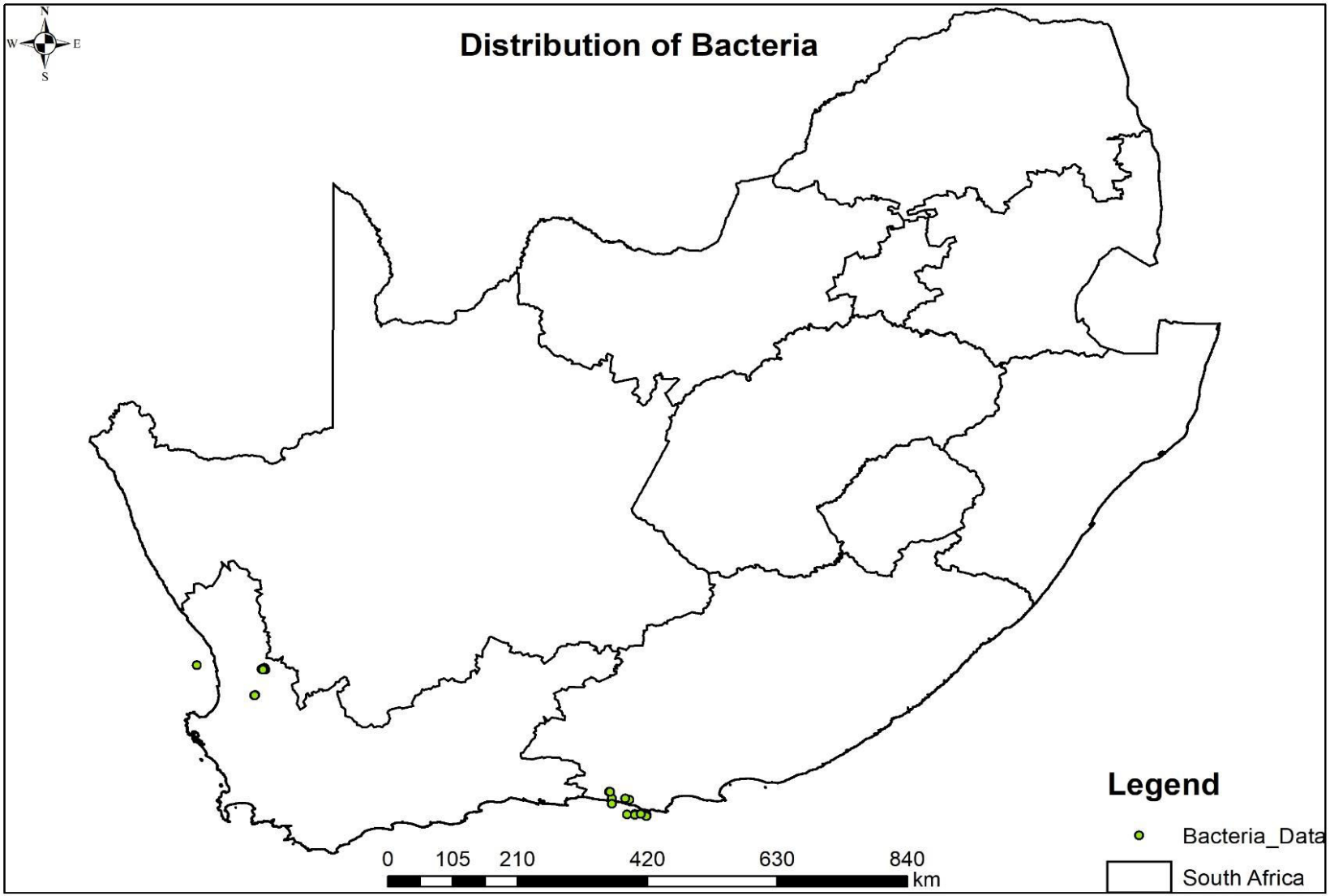
General Data Challenges

- Locality
- Collection date
- Taxon ID

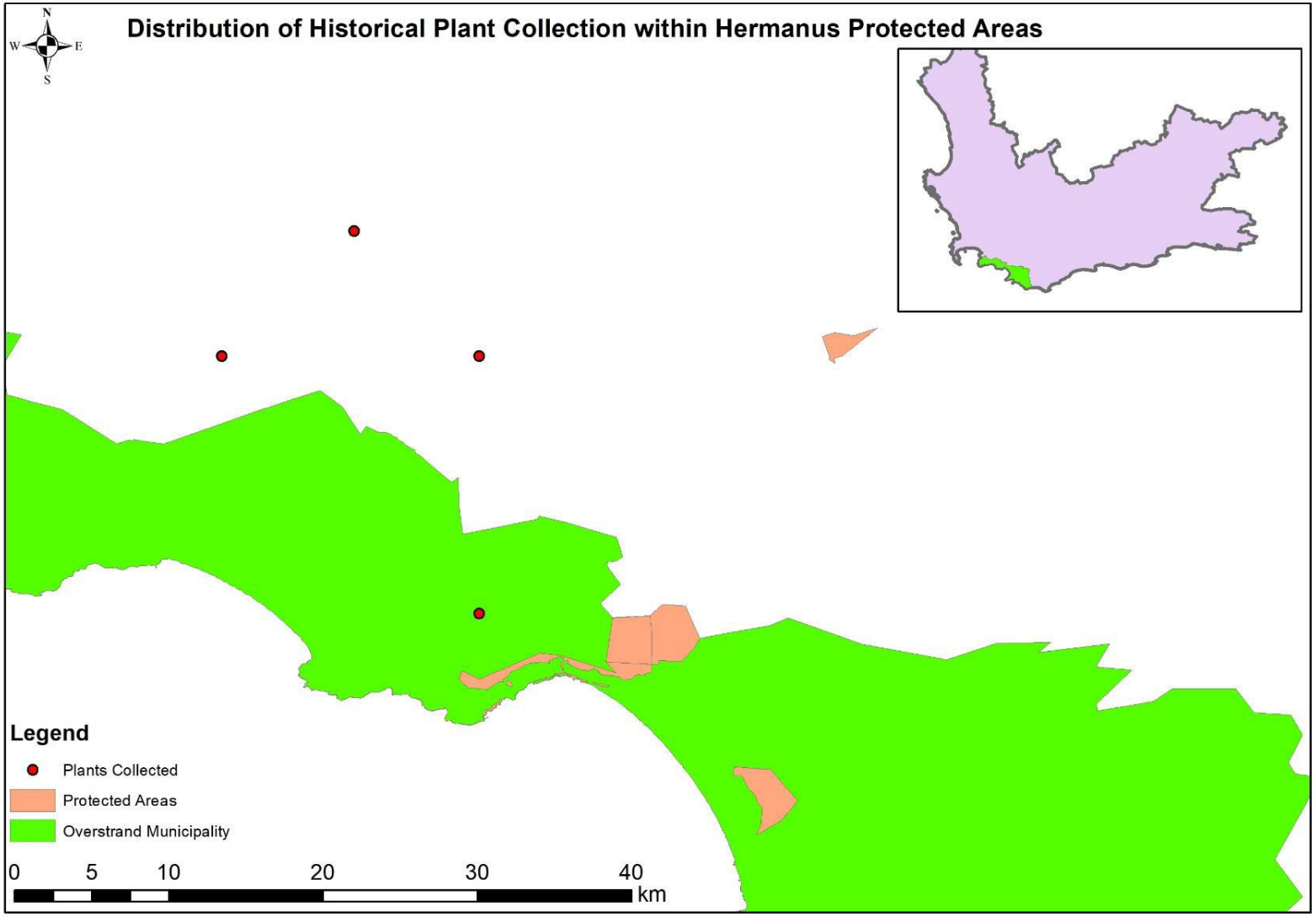
Locality

- Coordinates field left empty.
- Latitude and Longitude swapped.
- Conversion from Degrees-Minutes-Second and Degrees-Decimal Minutes to Decimal Degrees.
- Completely inaccurate coordinates.
- Conflict between coordinates and locality description.

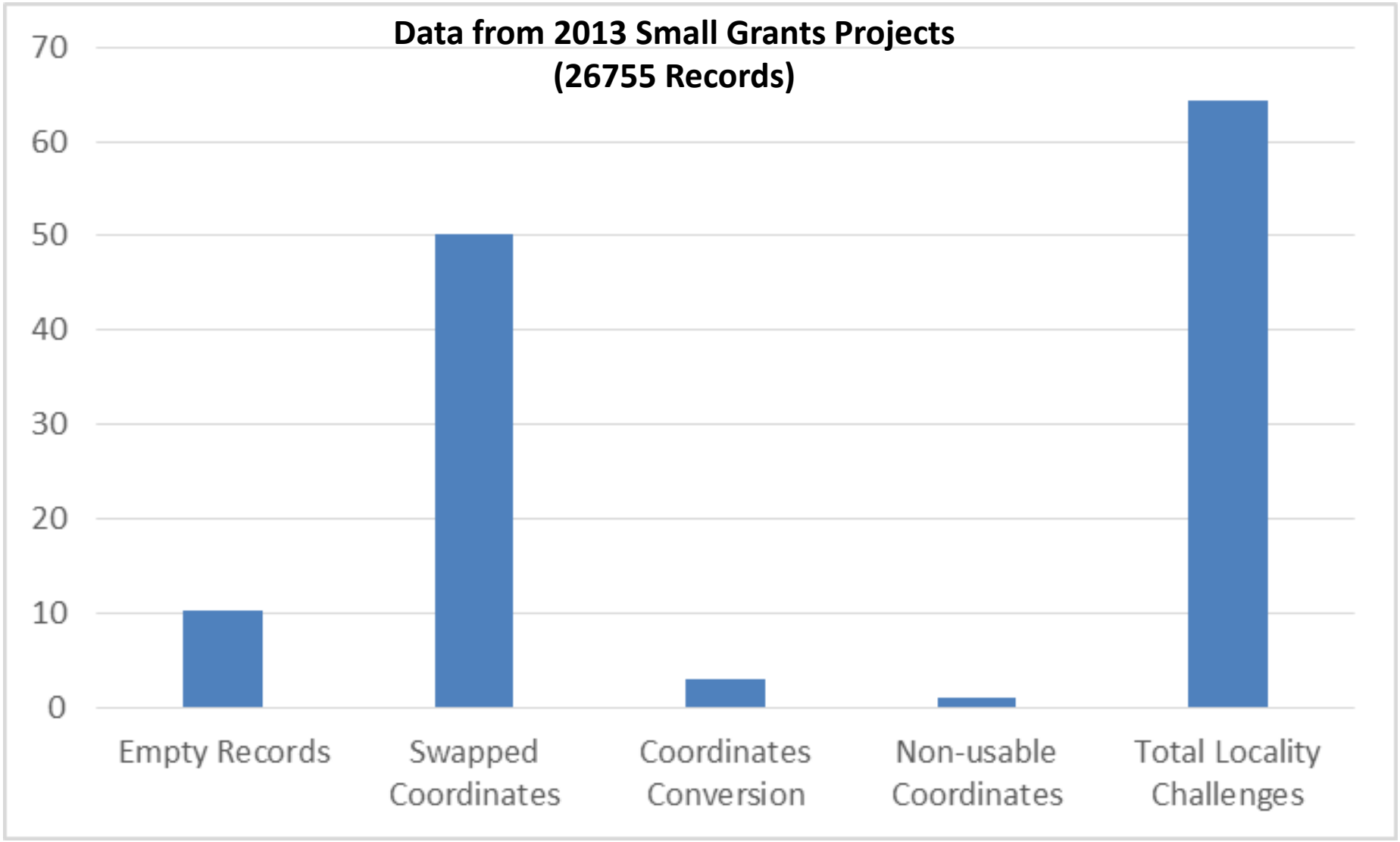
Locality


Distribution of Bacteria
N
W
E
S
0
105
210
420
630
840
km
Legend
Bacteria_Data
South Africa

Locality


Distribution of Historical Plant Collection within Hermanus Protected Areas
N
W
E
S
Legend
Plants Collected
Protected Areas
Overstrand Municipality
0
5
10
20
30
40
km

Locality

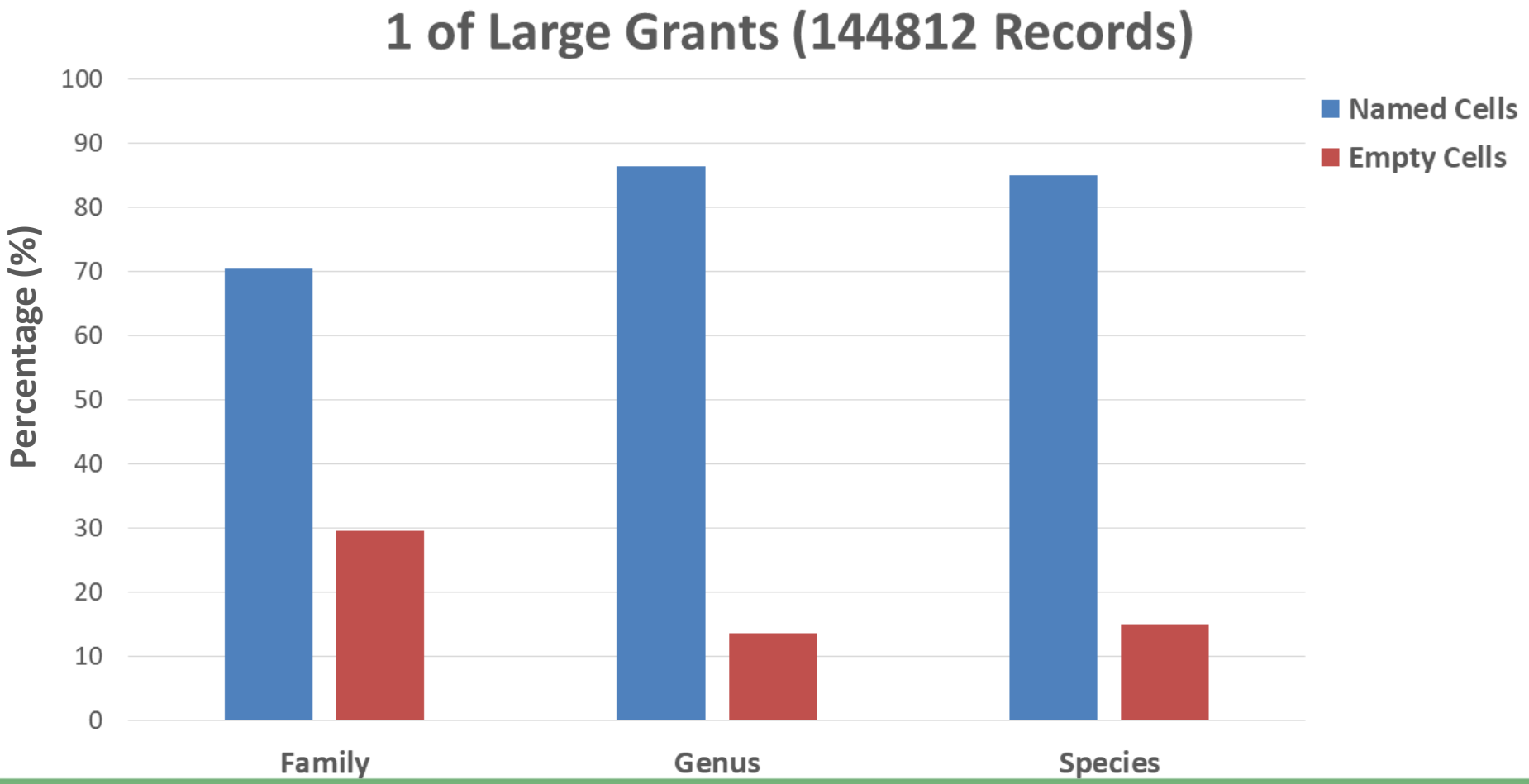

Data from 2013 Small Grants Projects
(26755 Records)
Percentages
70
60
50
40
30
20
10
0
Empty Records
Swapped Coordinates
Coordinates Conversion
Non-usable Coordinates
Total Locality Challenges

Collection Date

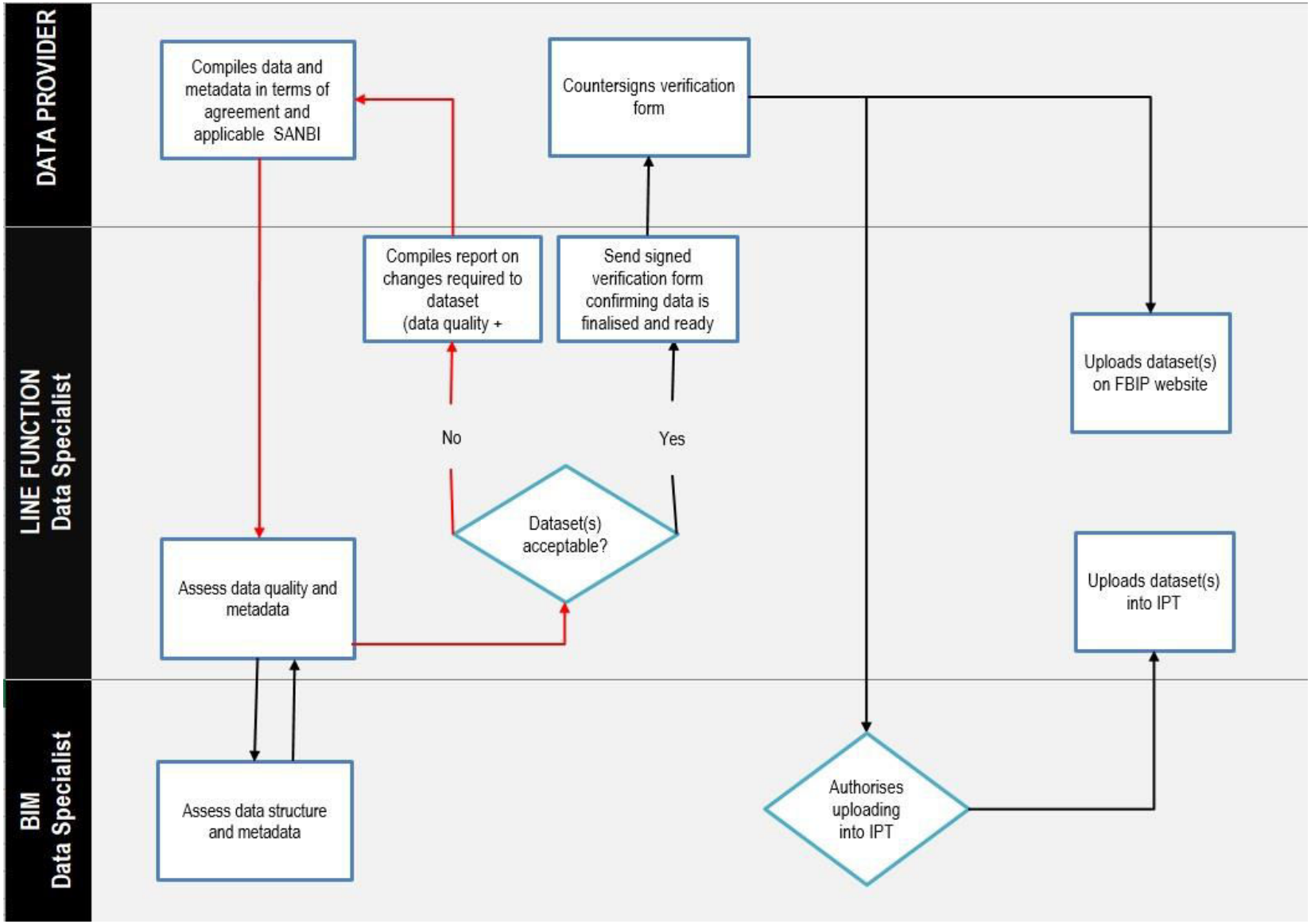
- The date when the organism was collected or observed.
- Correct - date written in Year/Month/Day (in their own fields).
- Fields often left empty.
- Example: 1 of large grants (total of 144812 records)
  - 24638 (17%) records with no date.
  - 16030 (11.1%) with year only.
  - 5370 (3.7%) with year and month only.
  - 98772 (68.2%) with Year/Month/Day.

Taxon ID

- 3 classification fields– family, genus, species.



Ensuring Data Standards



Conclusion

- Researchers need to pay more attention to the data they generate.
- Addressing the challenges highlighted will increase the value of data and therefore the value of the programme, ensuring continued funding.
- Standardized data sets have more value because it is easy to share and integrate into other data sets.


(Jack Ma – Executive Chairman of Alibaba Group)

- Entering an age of data which is predicted to become more valuable than oil.
- We in the biodiversity conservation need to step up our game in this regard.

Thank You