FBIP Data Challenges: The Need to Address Quality Standards

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

<table>
<thead>
<tr>
<th>Title of data set</th>
</tr>
</thead>
<tbody>
<tr>
<td>Taxonomic group covered</td>
</tr>
<tr>
<td>Species / specimen information</td>
</tr>
<tr>
<td>No. of records</td>
</tr>
<tr>
<td>Last updated</td>
</tr>
<tr>
<td>Description (explain what the data set represents)</td>
</tr>
</tbody>
</table>

## Resource Owner

<table>
<thead>
<tr>
<th>Data set Owner</th>
</tr>
</thead>
<tbody>
<tr>
<td>Organisation</td>
</tr>
<tr>
<td>Contact person</td>
</tr>
<tr>
<td>Phone</td>
</tr>
<tr>
<td>Email</td>
</tr>
<tr>
<td>Homepage</td>
</tr>
</tbody>
</table>

## Institutional Information

<table>
<thead>
<tr>
<th>Institution ID</th>
</tr>
</thead>
<tbody>
<tr>
<td>Collection ID</td>
</tr>
<tr>
<td>Dataset ID</td>
</tr>
<tr>
<td>Basis of Record</td>
</tr>
<tr>
<td>Catalogue Number</td>
</tr>
<tr>
<td>Occurrence Remarks</td>
</tr>
<tr>
<td>Record Number</td>
</tr>
<tr>
<td>Recorded By</td>
</tr>
<tr>
<td>Individual ID</td>
</tr>
<tr>
<td>Individual Count</td>
</tr>
<tr>
<td>Reproductive Condition</td>
</tr>
<tr>
<td>Establishment Means</td>
</tr>
<tr>
<td>Preparations</td>
</tr>
<tr>
<td>Disposition</td>
</tr>
<tr>
<td>Associated Reference</td>
</tr>
<tr>
<td>Associated Sequences</td>
</tr>
</tbody>
</table>

- **Institution ID**: An identifier for the institution having custody of the data set or information referred to in the record.
- **Collection ID**: An identifier for the collection or the dataset from which the record was derived.
- **Dataset ID**: An identifier for the data set.
- **Basis of Record**: Often serves as a link between field notes and an Occurrence record, such as a specimen collector’s number.
- **Catalogue Number**: Example: “found dead on road”.
- **Occurrence Remarks**: Example: “Oliver P. Pearson; Anita K. Pearson”.
- **Recorded By**: An identifier for an individual or named group of individual organisms represented in the Occurrence. Used to accommodate resampling of the same individual or group for monitoring purposes.
- **Individual ID**: The number of individuals represented present at the time of the Occurrence.
- **Individual Count**: Examples: “skin, skull”, “skeletal”, “whole animal (EUTA)”, “tissue (EUTA)”, “fossil”, “cast”, “photograph”, “DNA extract”.
- **Establishment Means**: Literature associated with the Occurrence.
- **Preparations**: Identifiers (publication, global unique identifier, UNI) 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

Data from 2013 Projects (24)

- **Time Frame of Data Delivery**
- **Data Proposed vs Data Delivered**
- **Use of FBIP Data Template**
- **Compiled Metadata**
- **DNA Data [17/24]**

**Categories and Data:***
- Goals Met
- Goals Not Met
- Data Not Submitted
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

Legend
- Bacteria_Data
- South Africa

0 105 210 420 630 840 km
Locality

distribution of historical plant collection within Hermanus protected areas

Legend
- Plants Collected
- Protected Areas
- Overstrand Municipality
Locality

Data from 2013 Small Grants Projects
(26755 Records)
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.

1 of Large Grants (144812 Records)

- **Family**: Named Cells (70%) and Empty Cells (30%)
- **Genus**: Named Cells (90%) and Empty Cells (10%)
- **Species**: Named Cells (90%) and Empty Cells (10%)
Ensuring Data Standards

DATA PROVIDER

- Compiles data and metadata in terms of agreement and applicable SANBI
- Countersigns verification form

LINE FUNCTION
Data Specialist

- Compiles report on changes required to dataset (data quality ±)
- Send signed verification form confirming data is finalised and ready

- Assesses data quality and metadata
- Dataset(s) acceptable?

BIM
Data Specialist

- Assesses data structure and metadata

- Authorises uploading into IPT
- Uploads dataset(s) on FBIP website
- Uploads dataset(s) into IPT
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