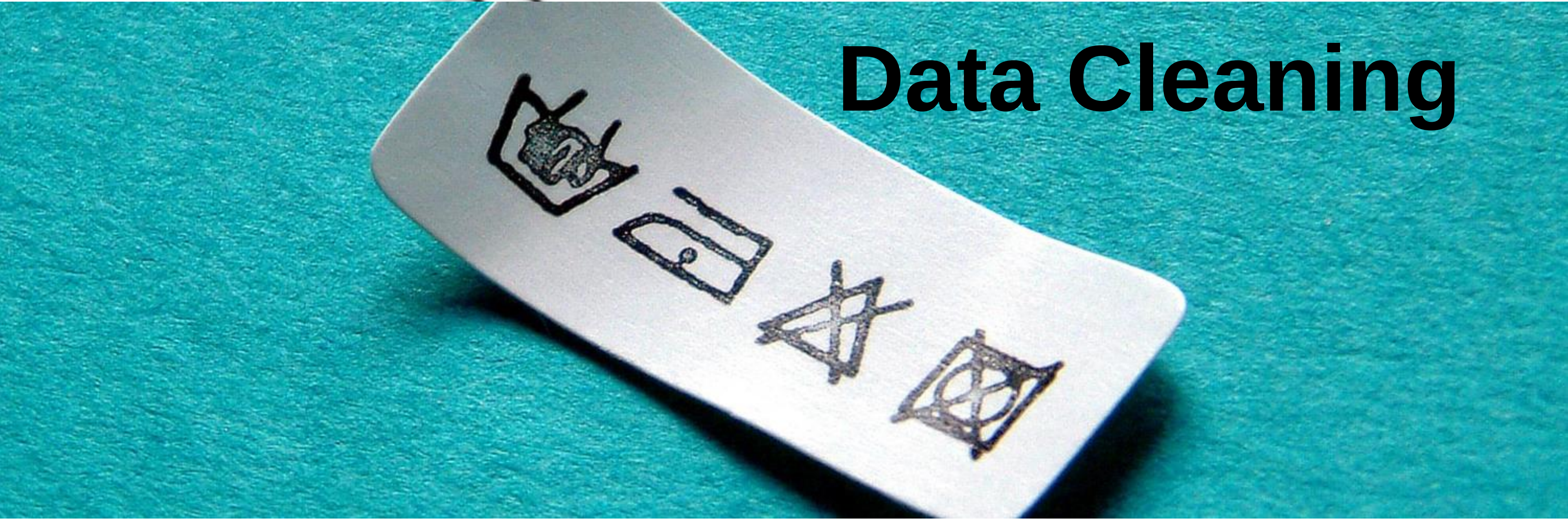
Data Quality and Data Cleaning



Adapted from various GBIF and BID presentations

Key concepts in data quality and data cleaning:

1) What is data cleaning?

2) How does data quality and data cleaning fit into overall data management?

3) Refresher on field and data types

4) Critical data fields for cleaning and standardising of biodiversity data

ENDANGERED WILDLIFE TRUST

1) Data Cleaning

A process used to improve data quality through correction of detected errors and omissions

The purpose of data cleaning is to make the data fit for use, by ensuring correctness & consistency

The process involves:

- Defining and determining error types
- Searching for and identifying error instances
- Correcting the errors
- Documenting error instances and error types
- Modifying data entry procedures to reduce future errors

Correctness

- ***Correctness (Accuracy)***

  Has the correct value been recorded?

  For example:

  Is *Acacia* a bird?

  Is ZM the correct country code for Zimbabwe?

  Can the height of a Fig tree be 1800 m tall?

  Are the geographic coordinates 0 0?

These values are all incorrect and need to be corrected in the dataset

Consistency

- ***Consistency (Precision)***

  How often do you get it right (are you recording the same value for the same thing every time)?

  For example:

  Full Name = Joseph Geoffrey Gawa

  Full Name = Gawa, J.

  Full Name = J. G. Gawa

  Full Name = Joe

  These are all the same person but their name has not been consistently captured in the dataset, and should be standardised

Data cleaning tools

*Ideally data cleaning should be as **automated** and **replicable** as possible, should be done using specialized tools and should be documented.*

Data cleaning tools:

- are used for working with messy data.
- are software programmes that assist the process of identifying and repairing inaccurate, incomplete, redundant, or non-conforming data. They help to find the patterns, missing values, character sets and other characteristics of your data values.
- find and clean duplicate data, bad entries and incorrect information.

In summary, a data cleaning tool assists, identifies, repairs, discovers, analyzes, cleans, and transforms messy data

Fitness for Use

“...data quality is related to use and cannot be assessed independently of the user. In a database, the data have no actual quality or value (Dalcin 2004); they only have potential value that is realized when someone uses the data to do something useful.“

Data that is fit for use is:

- accessible,
- accurate,
- complete / comprehensive,
- consistent with other sources,
- standardised,
- well documented in metadata,
- easy to read and easy to interpret

Reference: Chapman, A. D. 2005. Principles of Data Quality, version 1.0. Report for GBIF, Copenhagen. ISBN 87-92020-03-8.

In summary

Your data should:

- be **clean** - Correct and Consistent
- follow a **biodiversity data standard**
- be **well structured**
- be documented in **metadata -** Fit for use, complete, easy to interpret

This is critical not only for sharing your data with others, but also for importing your data into other software packages for statistical analysis, GIS mapping etc

When would you need to clean data?

- Before analysing or mapping any of your own datasets
- Before sharing your datasets with anyone else, or uploading to a data portal
- Before using data that you have downloaded from a website such as GBIF
- Before using data that anyone else has shared with you

2) How does data cleaning fit into data management?

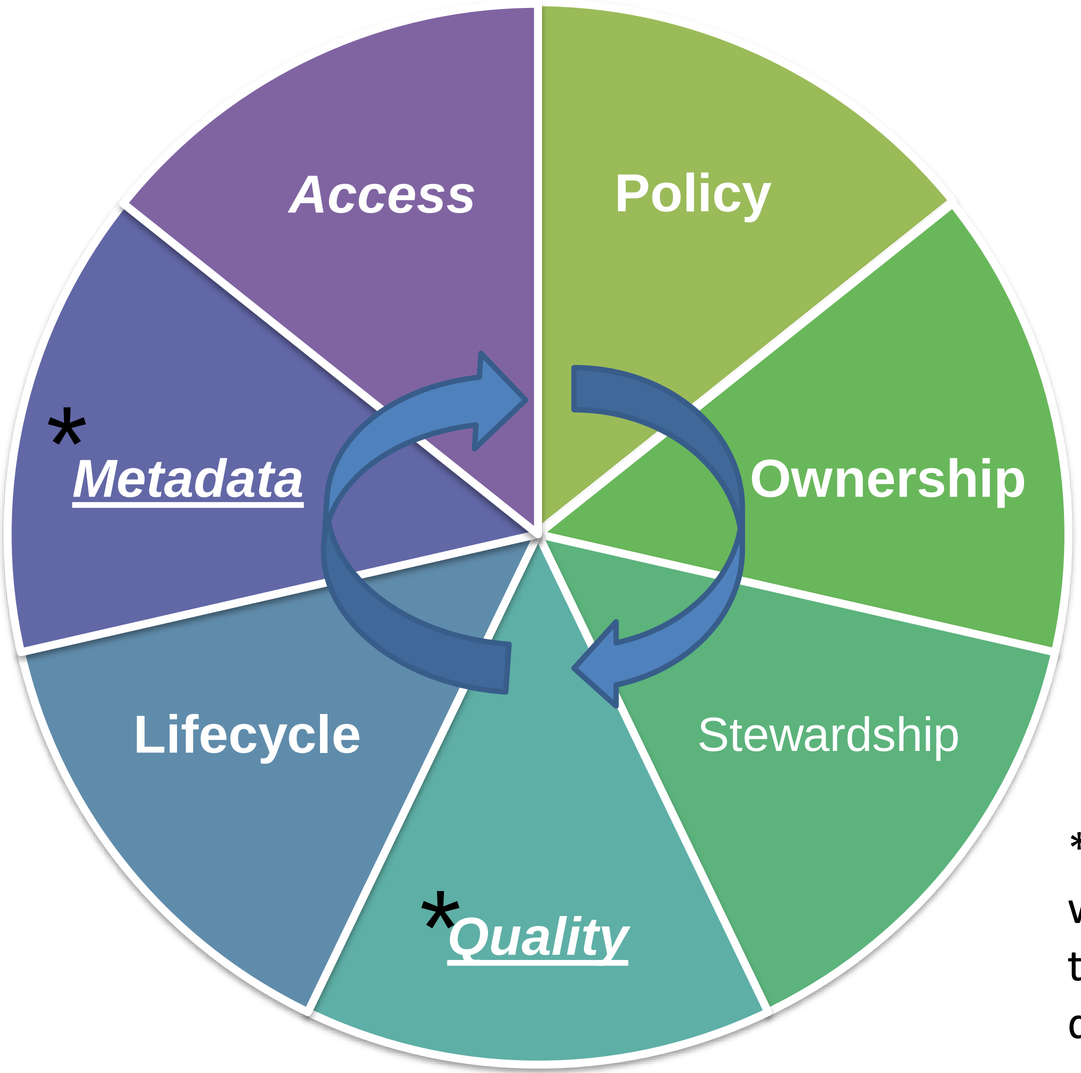
- Data Management is the whole range of activities involved in the handling of data.

- These activities include:

Institutional Data Policies; Data Ownership; Data Documentation and Metadata; Data Quality; Data Access etc

- Some of these activities are your responsibility, while others are mainly an institutional responsibility (i.e., your university, institution, department etc)

Data management activities in summary:



* these two activities are what we are focussing on most in this training, as the responsibility to do them lies mostly with you as an individual

Two golden rules for Biodiversity Data Management

- follow **international standards for cleaning and formatting data**. (*Ideally these should be standardised within institutions, and definitely within datasets*)

- **write the metadata** about that dataset that you collected or compiled. (*Ideally your institution should keep a dataset inventory and a catalogue of all metadata*)

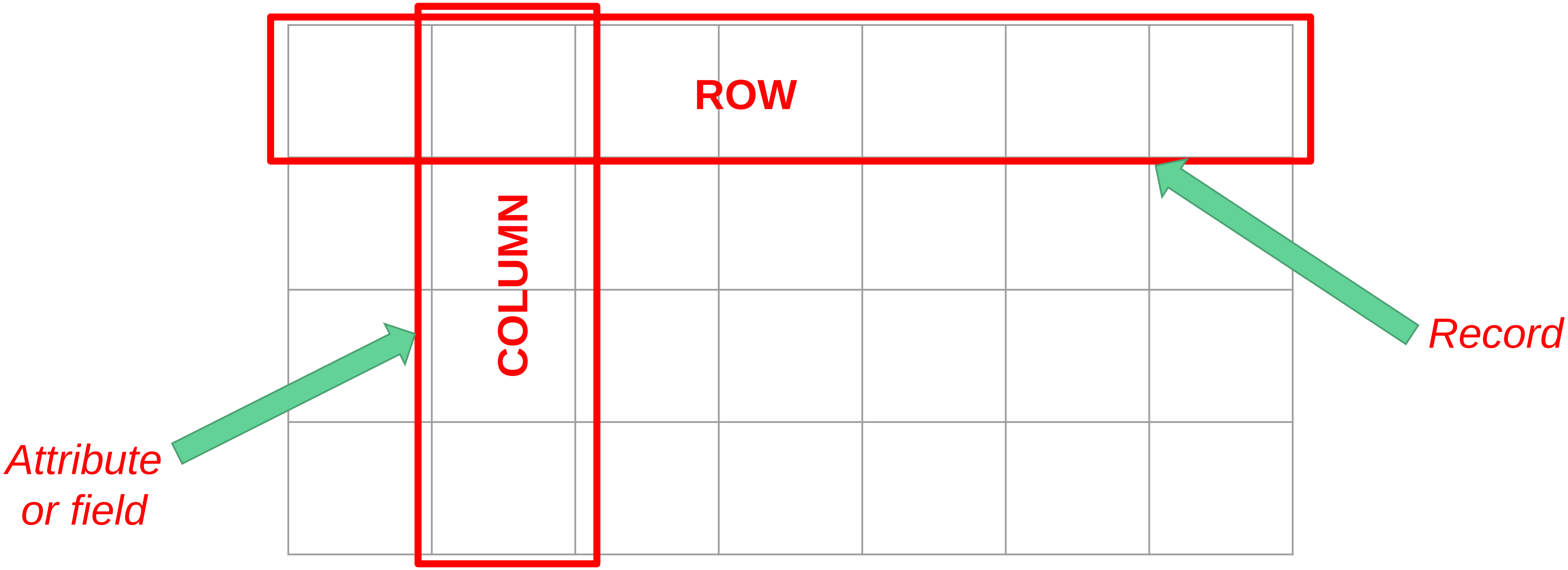
Mark J. Costello, John Wieczorek (2014) Best practice for biodiversity data management and publication. Biological Conservation.

3) Data Tables, Fields and Data types

Data Tables
Field Types / Data Types

Data Tables

e.g., a database table or a spreadsheet in Excel, where data is stored in columns and rows. Each row represents a record, and each column represents an attribute (such as date or species or location) – also known as a field



Field Types

Field should have a particular data type

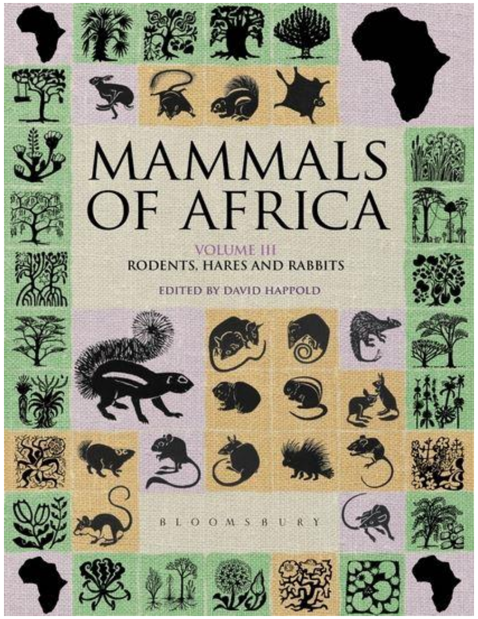
The information that you store in a column or field should be consistent and of a single data type

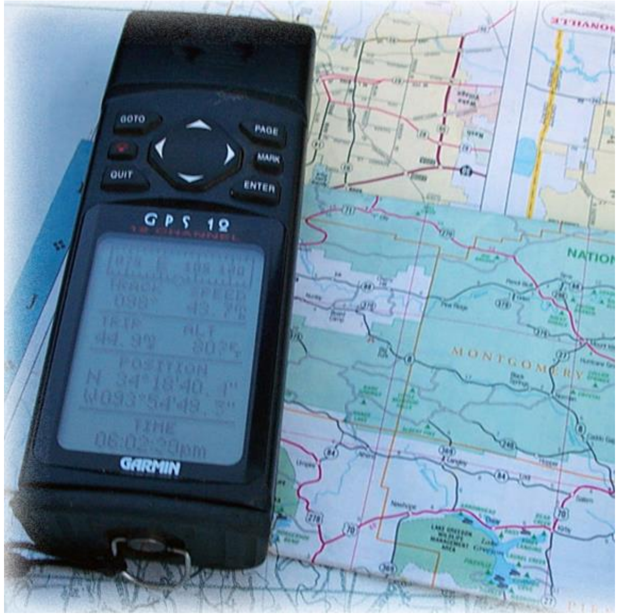
Data Types:

a. numeric - 2 types: integer or decimals - **Numbers**
b. alphanumeric - character, string - **Text**
c. date/time - date, time or date and time – **Standardised format**
d. memo - long text, longchar, blob - **Unstructured Text**
e. boolean - 1/0, 0/1: Yes/No, Y/N, True/False - **Binary**

4) Critical data fields for cleaning and standardising are TAXONOMIC and SPATIAL

i.e., what something is and where it was observed

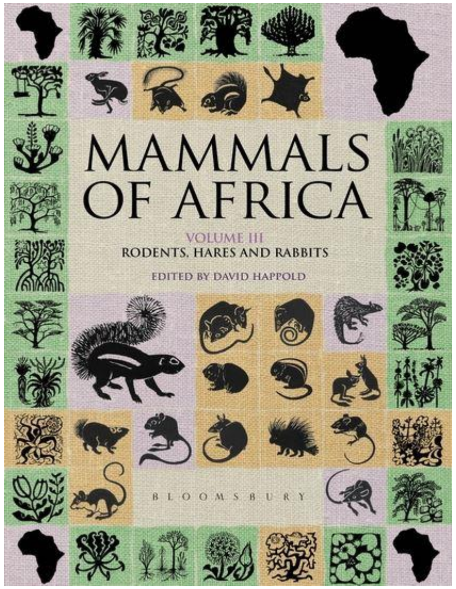




This information is captured in the DwC classes Location and Taxon

Taxonomic data: definition

- **Names** (scientific, vernacular, rank, hierarchy, …)
- **Status** (synonyms, valid names, …)
- **References** (author, date and location)
- **Identification** (by whom and when?)
- **Quality terms** (ID certainty, …)



MAMMALS OF AFRICA

VOLUME III

RODENTS, HARES AND RABBITS

EDITED BY DAVID HAPPOLD

BLOOMSBURY

Taxonomic data

Scientific/common name **= entry point**

⇩

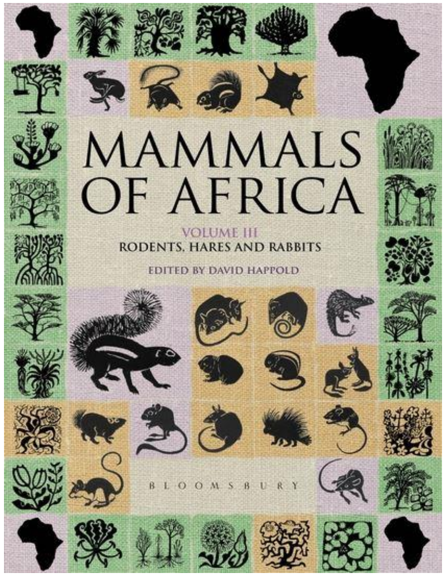
Possible errors and solutions:

- Incorrect identifications (calls for help from a taxonomist)
- Typos (data cleaning)
- Wrong format (data cleaning)

Taxonomic data

Taxonomic data: common MISTAKES to avoid

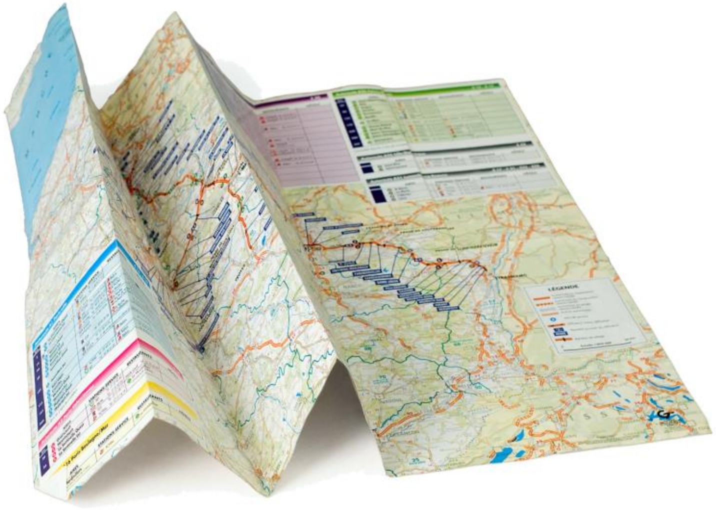
- Missing data (e.g., subspecies included but not the species)
- Incorrect values (typos, information in the wrong field, use of symbols « ?? », …)
- Uncertainty in at least one name of the binominal nomenclature
- Duplicates (synonyms, several valid names…)
- Inconsistent data using different checklists



Spatial data

Spatial data is key information to determine the fitness-for-use of primary biodiversity data, as this information can be used for:

- Species distribution modelling
- Selection of areas to protect
- Resource and environmental management
- Climate change modelling
- etc ....

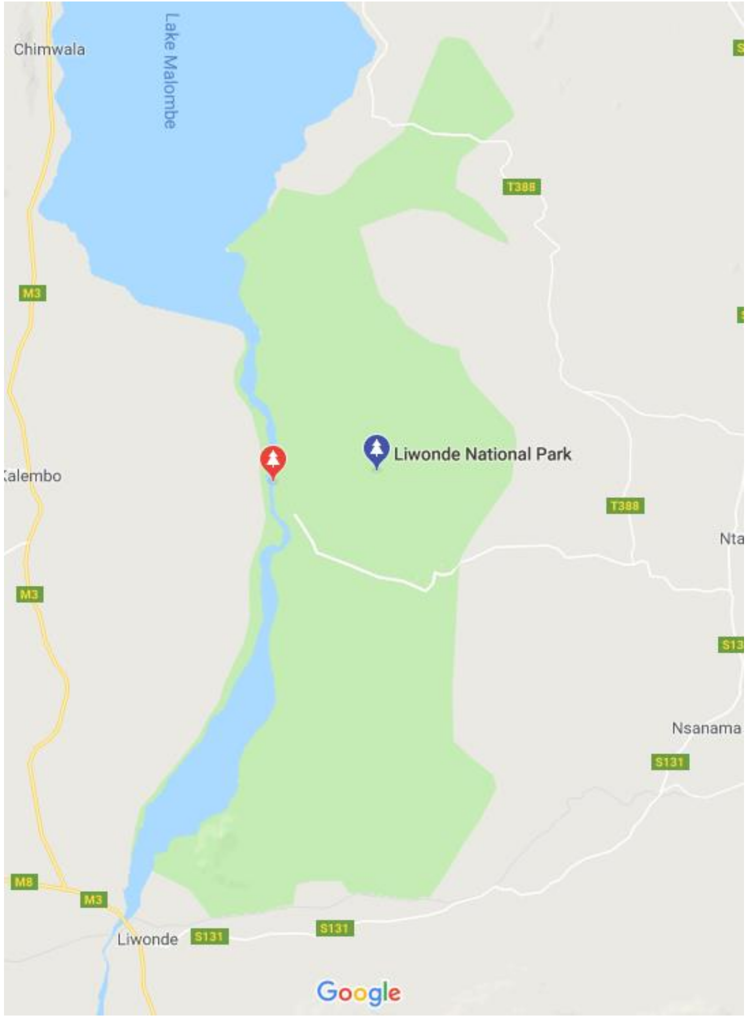


Spatial data

Spatial data: definition

What are we talking about?

- A place description
- Polygon *e.g*. a protected area, or a province
- Grid system e.g. Quarter degree grid cell (or QDGC), pentads
- Latitude and longitude

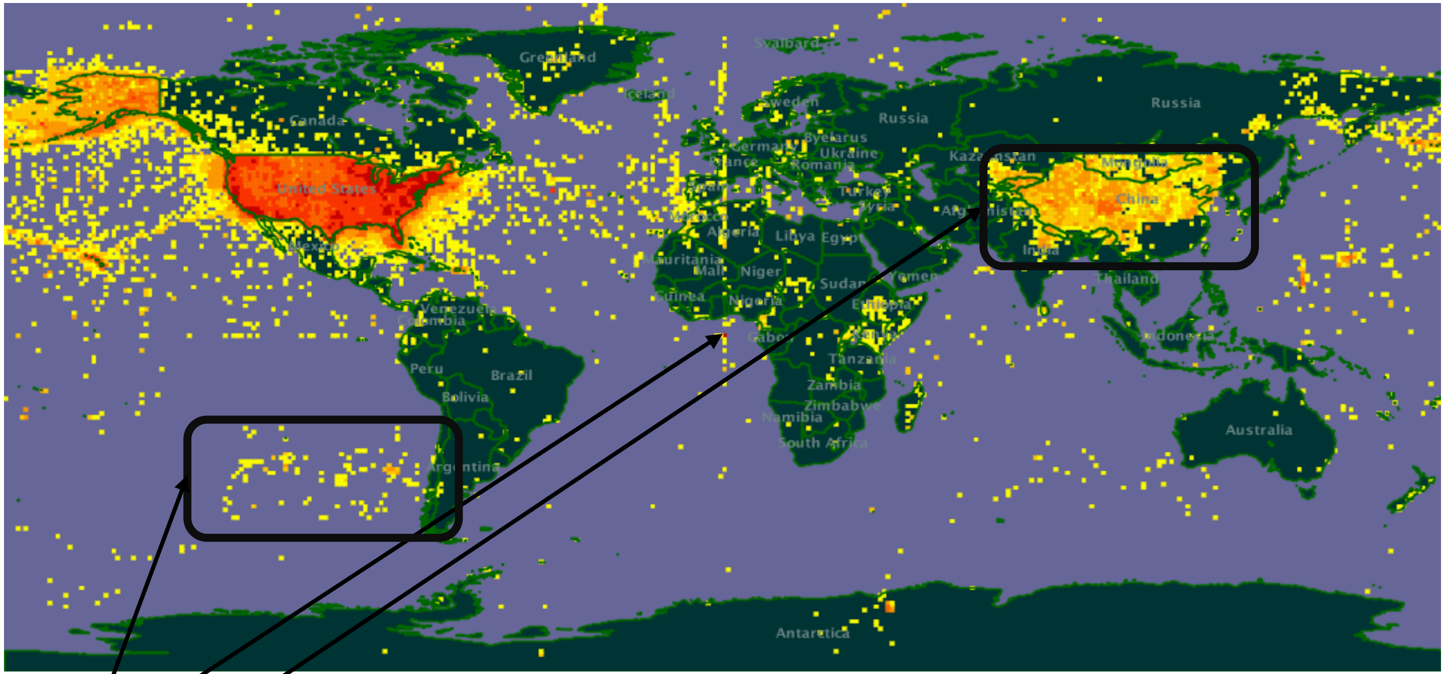


Spatial data: common MISTAKES to avoid

- Inverting coordinates (swopping latitude and longitude)
- Using 0 values for coordinates, when the value is unknown
- Not providing a coordinate reference system
- Incorrectly converting coordinates from one format to another

Spatial data MISTAKES – an example

The below image is of an early GBIF map showing USA data, which provides an example of some common mistakes:



- 0,0 coordinates, plotting in the ocean off West Africa
- Coordinates were inverted, showing a mirror effect on China and slight mirror effect west of Chile

Next steps

- Brief introduction to OpenRefine, which is a data cleaning tool
- Use OpenRefine to tackle some common problems in biodiversity datasets