CESP Project: Strengthening Zimbabwe's GBIF node through collaboration with GBIF Spain

GBIF Zimbabwe

MINISTERIO DE CIENCIA, INNOVACIÓN Y UNIVERSIDADES

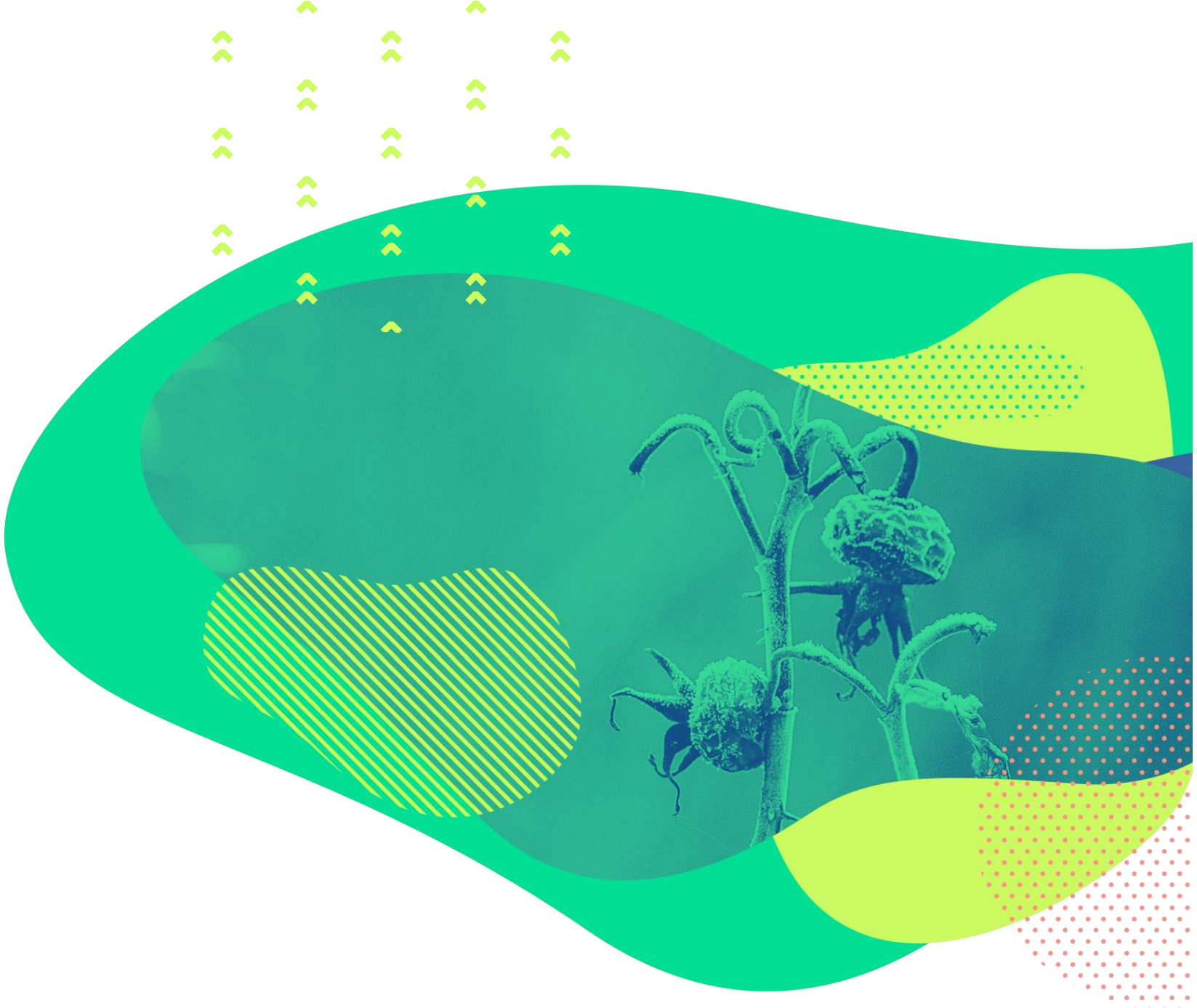
CSIC Consejo Superior de Investigaciones Científicas

REAL JARDÍN BOTÁNICO

Gbif.es

GBIF

Basic guidelines about data quality and data cleaning



CESP Project: Strengthening Zimbabwe's GBIF node through collaboration with GBIF Spain

Katia Cezón
katia@gbif.es

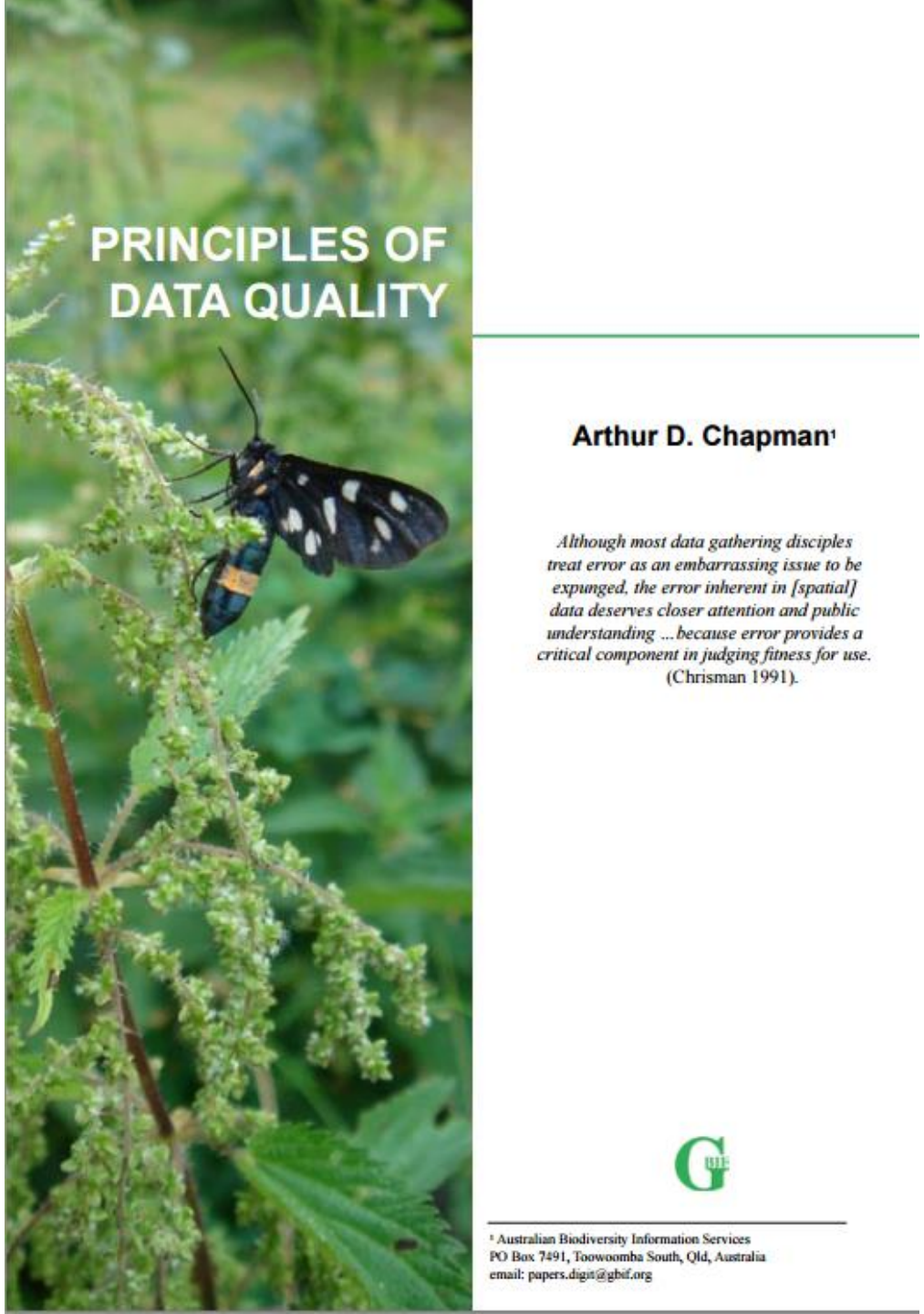
Gbif.es

Data quality and data cleaning

Reference books

- Chapman, A. D. 2005. **Principles of Data Quality**, version 1.0. Report for the Global Biodiversity Information Facility, Copenhagen. ISBN 87-92020-03-8. Available at http://www.gbif.org/orc/?doc_id=1229.

- Chapman, A. D. 2005. **Principles and Methods of Data Cleaning: Primary Species and Species-Occurrence Data**, version 1.0. Report for the Global Biodiversity Information Facility, Copenhagen. Available online at http://www.gbif.org/document/80528.

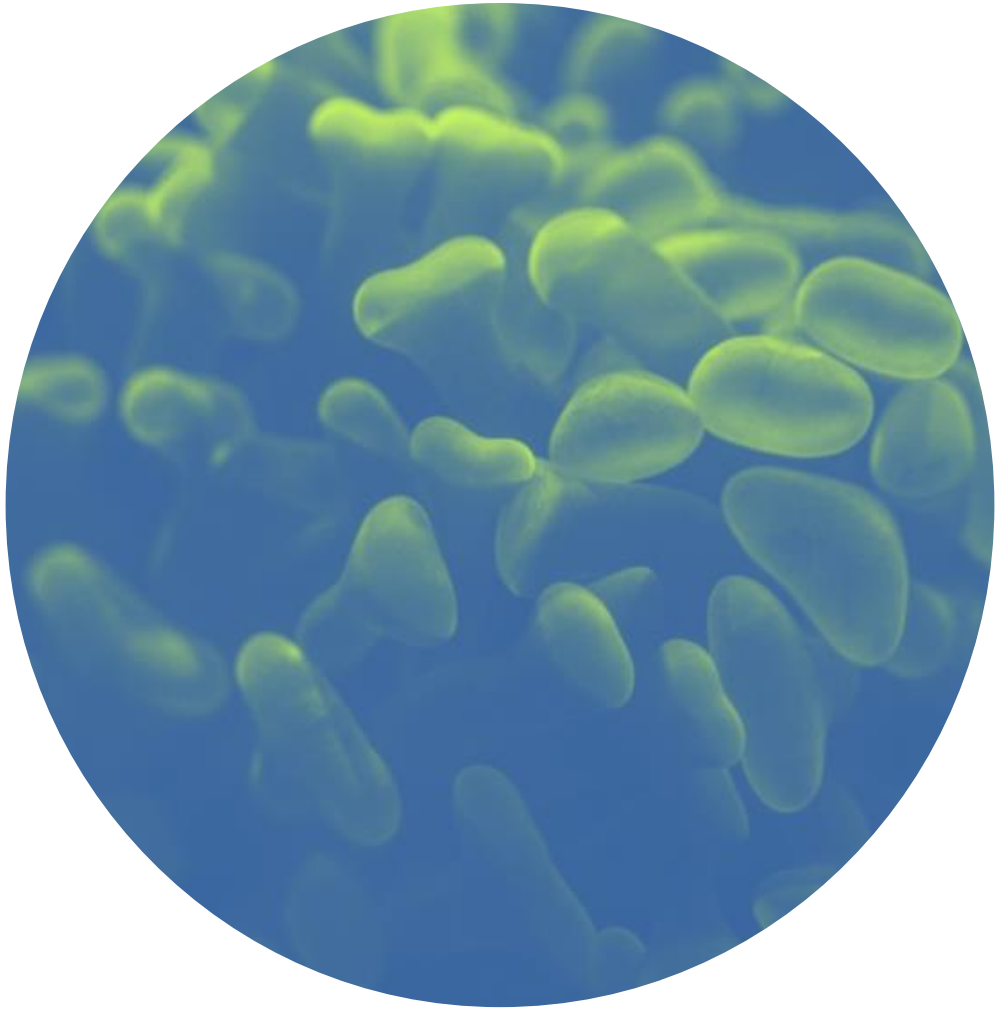


Data quality concepts

Data quality

What is it?

- *…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 only when someone uses the data to do something useful…*



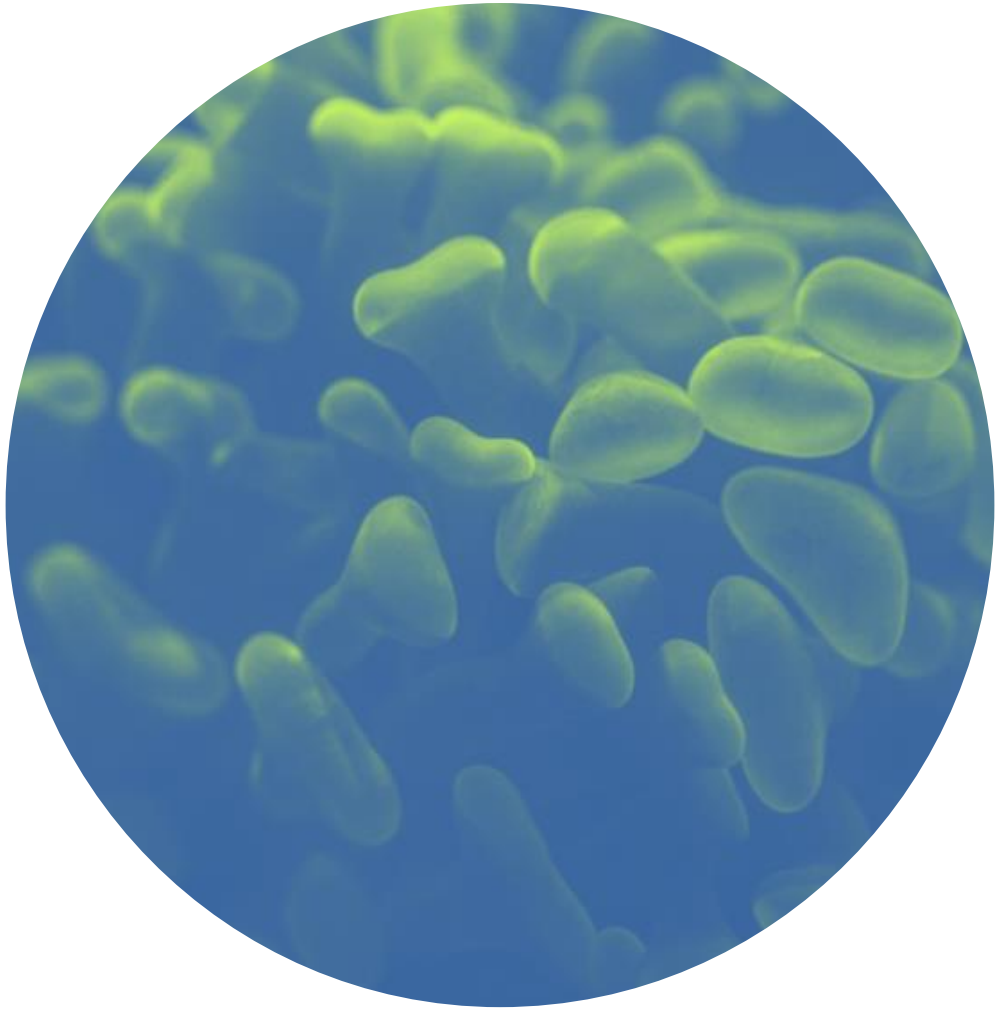
Chapman, A. D. 2005

Data quality

What is it?

Fitness for use

- Potential use. Data quality is relative to its intended use. Data may not be useful for one user**, but it can be useful for others**.



Data quality

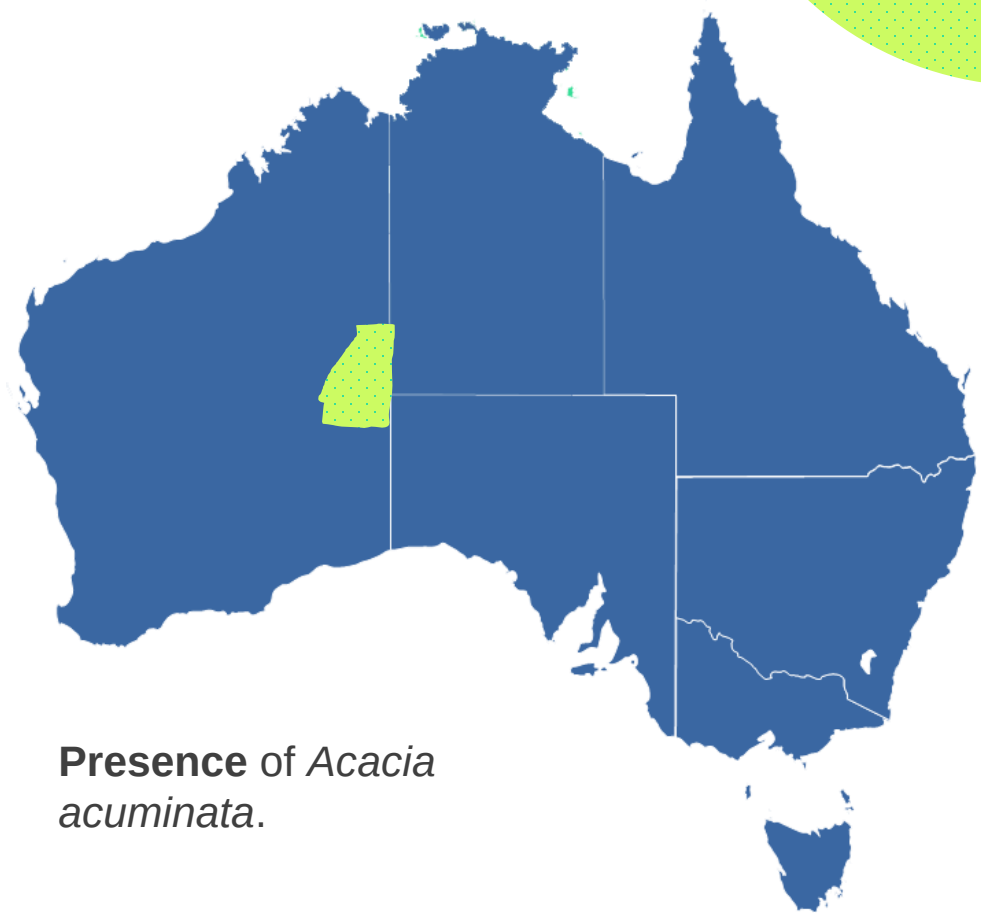
What is it?

Is it a good quality data?

Fitness for use

- Does *Acacia acuminata* occur in Australia?
- Does *Acacia acuminata* occur

  Ngaanyatjara protected area?



Presence of *Acacia acuminata*.

Data quality

What is it?

Are oysters tasty?

Fitness for use parameters

- Accessibility
- Accuracy
- Timely
- Completenes and Comprehensiveness
- Consistency
- Relevance
- Readability and interpretation

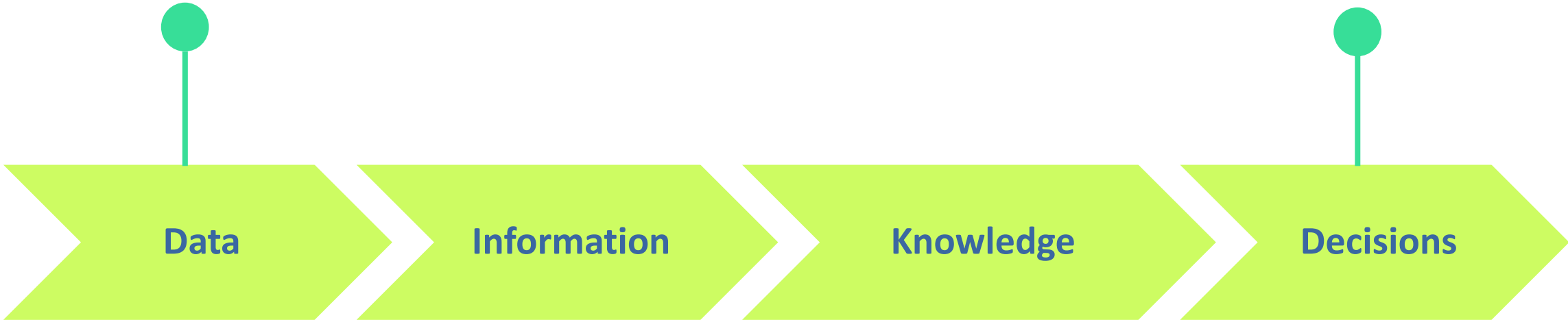
Fitness for use parameters

- Accessibility
- Accuracy
- Timely
- Completenes and Comprehensiveness
- Consistency
- Relevance
- Readability and interpretation

Redman (2001), suggested that *for data to be fit for use they must be accessible, accurate, timely, complete, consistent with other sources, relevant, comprehensive, provide a proper level of detail, be easy to read and easy to interpret.*

Why data quality is important?

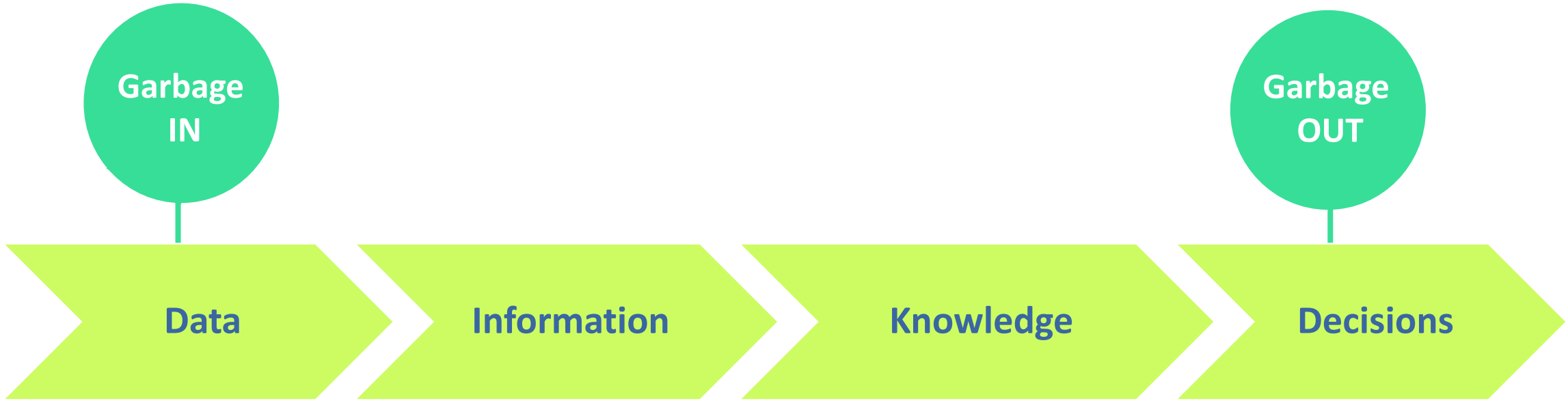
From data to decisions



Better data, better decisions

Why data quality is important?

From data to decisions

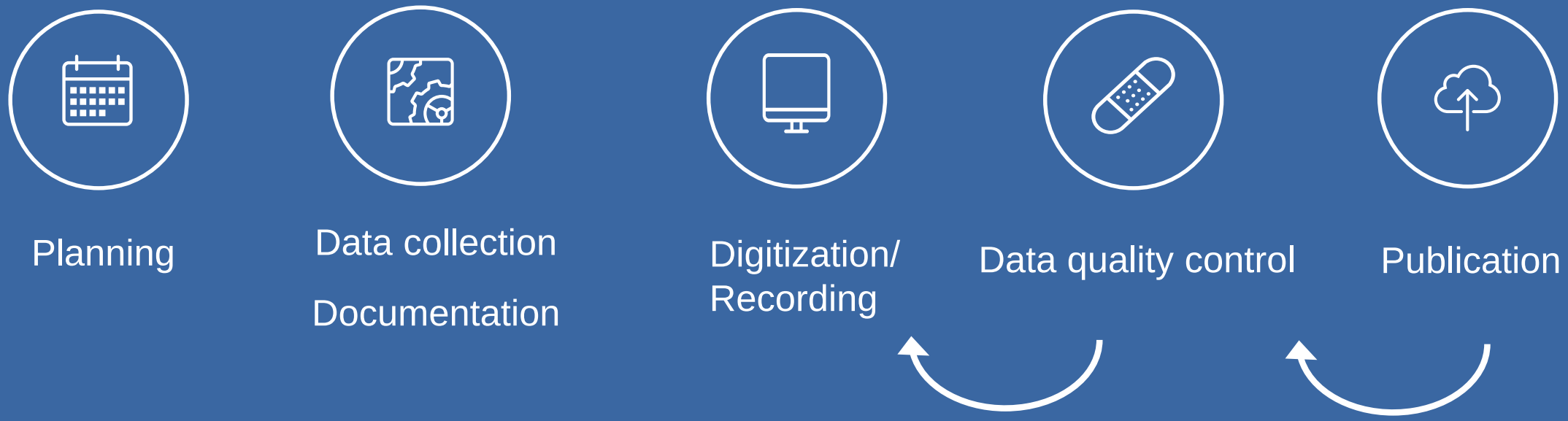


Quality problems lead to poor quality results, analysis and decisions.

Loss of data quality can occur at every step

Biodiversity data flow and processes in the data value chain

When problems occur?



Mechanisms for improvement

Mechanisms for improvement

Prevention

It is better to **prevent errors** than to cure them later and it is, by far, the cheaper option (Redman 2001).

Detection

Find and identify errors that have already been incorporated into the database

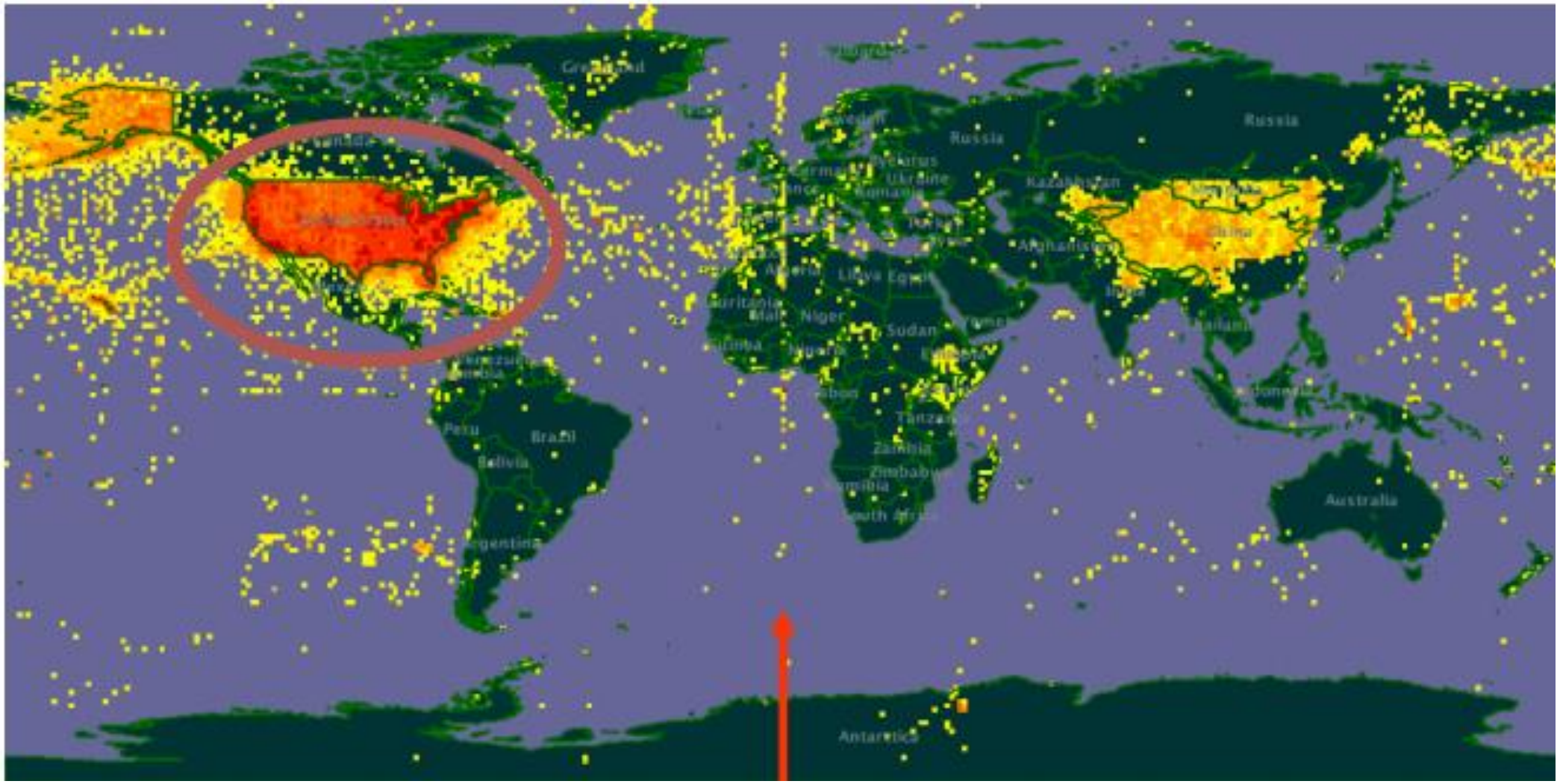
Cleaning and corrections

Cleanup process of the database

Documentation and recommendations

Documentation helps to feedback mechanisms ensure that the error doesn’t occur again

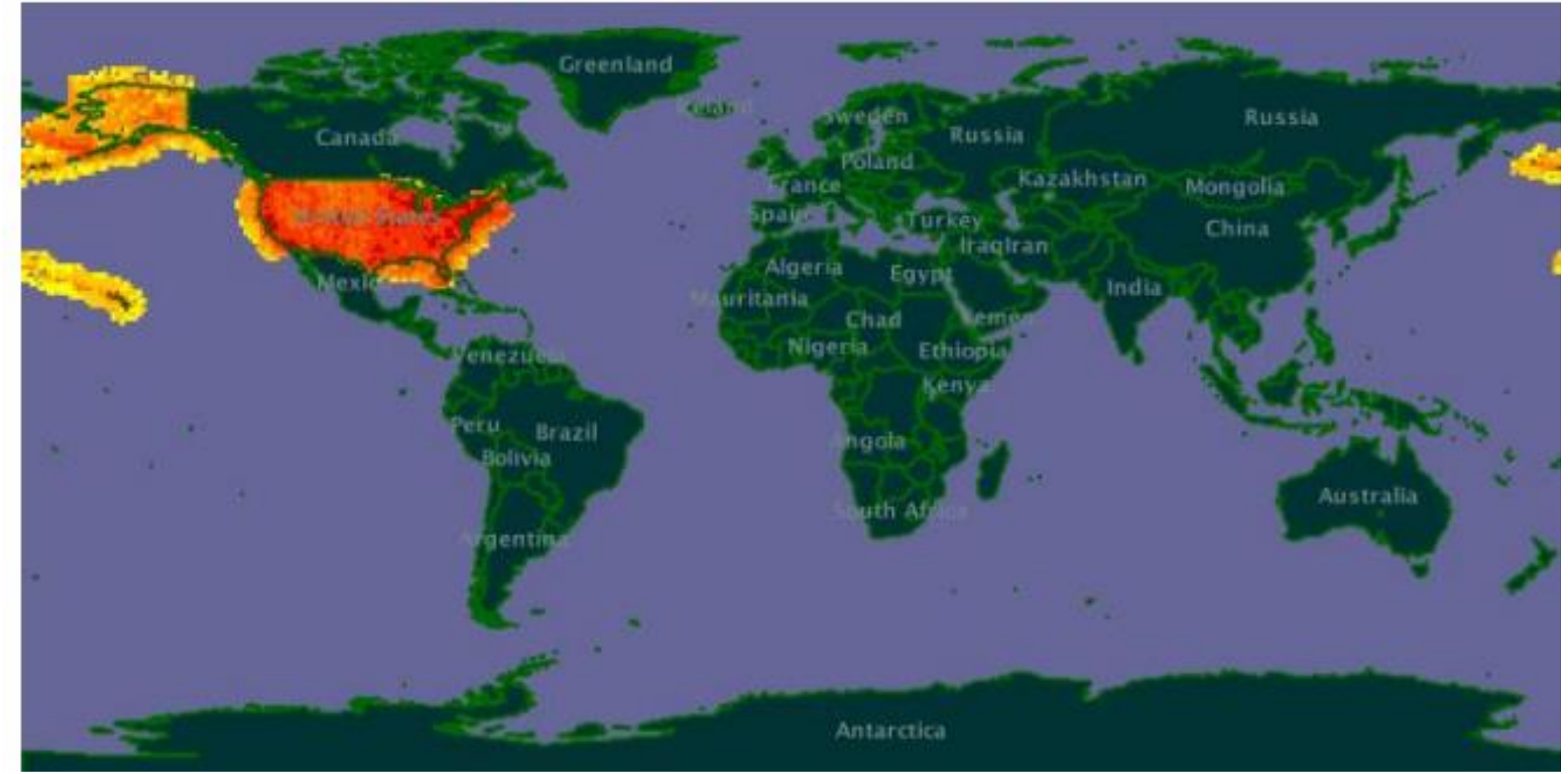
Example



Tim Hirsch


After some cleaning…

Example



Tim Hirsch

“To decide to clean the data first and worry about prevention later, usually means that error prevention never gets satisfactorily carried out and in the meantime more and more errors are added to the database (Chapman 2005)”

GBIF Zimbabwe

MINISTERIO DE CIENCIA, INNOVACIÓN Y UNIVERSIDADES

CSIC
Consejo Superior de Investigaciones Científicas

REAL JARDÍN BOTÁNICO

Gbif.es

GBIF