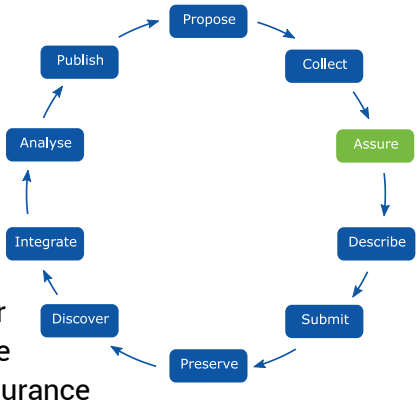
# Assure




## What is it?

To assure and control the quality of data, errors should be prevented from entering a dataset or eliminated from a dataset. GFBio will offer a visualization-tool that can, for instance, create a graphical overlay with other data to detect errors. You will be able to transform from different coordinate systems or just to create a distribution of your data on a graphical display. Assurance is important in order to **monitor and maintain data quality** during the whole data life cycle. Data quality is one of the main challenges when promoting data reuse, therefore it should be ensured when data are collected, entered and analyzed. As soon as **data are 'fit for use'**, they should be accessible, accurate, complete, consistent with other sources, relevant, comprehensive, provide a proper level of detail, be easy to read and easy to interpret.

## How to do it?

As soon as the data are entered in spreadsheets/databases, you can start with basic quality assurance.

1. Data entered by hand (by more than one person) should be double-checked.
2. Check the consistency of the format throughout the data set.
3. Document the cleaning process in a script (transformation workflow/mapping rules).
4. **Data cleaning process**: Perform data cleaning (data entry errors, measurement errors, distillation errors, data integration errors). Detect and delete errors and discrepancies to enhance data quality.
    a. Identify missing, impossible, or anomalous values and discrepancies using GFBio-visualization-tools (e.g. geographic distribution) and analysis-tools to visualize outliers (statistical and graphical summaries). (GFBio is currently working on that.)
    b. Verification: ensure that your transformation will cure all errors found.
    c. Transform the data using GFBio-transformation-tools (soon available).
    d. Backflow of cleaned data into source and replacement of errors.
    e. Save this cured dataset as a new version to avoid irreparable distillation errors (Versioning).
5. Also at later stages of the data life cycle, like during analysis or interpretation of results for a subsequent publication, further discrepancies can be detected and eliminated.
6. Communicate data quality using either coding within the data set that indicates quality, or in the metadata.

## Who does it?

Currently every **data producer and reuser**, who does his own research or is part of a research programme (like ecologists, geo-scientists, geneticists, modellers etc.).

## Key elements

- Check for accuracy and consistency of data structure and format through semi-automated tools soon provided by GFBio.
- Detect errors through GFBio-visualization and analysis-tools (statistical/graphical analysis)
- Go through the data cleaning process and make a script of it.
- Version your data sets.
- Document the quality of data by flags, metadata, coding.

## Useful links

https://www.dataone.org/best-practices (Best-Practices-Primer)
http://www.youtube.com/watch?v=i2jcOJOFUZg (MANTRA Video with Jeff Haywood - Importance of good file management in research)
http://openrefine.org/ (useful tool)