



What is it?

You might want to integrate a dataset that you discovered and that fits to your own data in order to **verify your results** or as a starting point for an **integrative study** or just to **test a new hypothesis** for a follow-up study.

Integration is the merging of multiple datasets from different sources, like your recently collected data with former data from other owners, resulting in a new, bigger dataset. You can manually integrate data or save time and use the automatic integration of GFBio (implementation is in preparation). This requires to take care about a common syntax and terminology right from the start (see [Fact-Sheets 'Collect'](#) and ['Describe'](#)). Again, you can use GFBio services, like the Terminology- and Description-Tool, that will soon be available. When other authors' data are reused, it is fundamental to provide credit to the data creators through a robust **data citation mechanism**.

How to do it?

1. For an efficient integration, good data management practices should have been performed previously. Make data management easy by making use of [GFBio tools and workbenches](#) for data collection, assurance, description, submission and discovery as soon as they are available! (Stay informed about new Tools and Functions via [Twitter](#) and [Facebook](#)!)
2. **Understand the data** and assess suitability for the required purpose. You can use visualisation and aggregation tools (GFBio works on their implementation), like geographic maps or descriptive statistics for a better understanding.
3. **Agree on semantics** during the collect and describe stages! This will facilitate the integration process later on.
4. Ensure that formats and parameters are compatible (datum, resolution, metric units).
5. Document the relationships among data sets from different sources.
6. Bear in mind to use unique identifiers to prevent duplication of the used data set.
7. Document the data integration process (script, workflow) and describe it in the metadata.
8. Document your data analysis.
9. Establish a proper citation of the data sets you used (**data provenance**)!
10. Provide feedback to data creators about data quality and metadata quality. This might facilitate data integration for further users.

Who does it?

Data reusers in general, that can be modellers or researchers in integrative or comparative studies.

Key elements

- Proper data management practices in previous data life cycle stages
- Documentation of the integration and analysis workflows
- Data citation
- Data provenance
- Feedback to data owner

Useful links

<http://www.dataone.org/education-modules> (DataONE - Education Modules)

<https://www.dataone.org/best-practices/document-integration-multiple-datasets> (DataONE - Integration)

<http://www.dcc.ac.uk/sites/default/files/documents/DC%20101%20Transform.pdf> (DCC - Digital Curation)

http://ukdataservice.ac.uk/media/104397/data_citation_online.pdf (Economic & Social Research Council)