Edaphobase Data Quality Checklist
for incoming data packets to be imported into Edaphobase

For the users of this checklist:

Consider the data quality-control check as similar to the review of a manuscript:

A) for *technical quality*: especially obvious errors, completeness (no gaps in data and metadata), correctness (information in correct column), etc.
B) for *scientific quality*: consistency (= plausibility), comprehensibility (cryptic abbreviations ...), probability of the data entries, appropriate methods, etc. [Taxonomic checks are carried out by the taxonomist(s) responsible for the taxonomic group].

1. The following questions are worded such that they can be answered as 'yes', 'no', or 'not applicable (N/A)'. Marked black boxes are acceptable for Edaphobase. Ticked red boxes indicate discrepancies which have to be clarified before importing the data (especially in the case of mandatory fields) or - if necessary after consultation with the data provider - are nevertheless acceptable for importing the data into Edaphobase.
2. **Mandatory fields** are marked with *****; **"highly recommended" fields** are marked with **(*)**

Standard for future new data is that all incoming data for import to Edaphobase will run through the Edaphobase Import Wizard. The present checklist is thus created as a quality control *after* the Import Wizard has outputted the data.

All changes to the data must be recorded and documented in the accompanying data (who supplemented/corrected what and when).

Data Controller Information (Controller-1):

Name of the Data Controller:_____

File name of the data packet: _____

Data provider (Name): _____

Data receipt (Date): _____

Storage location (server directory) of the original version: _____

Data type (raw data, literature, collection ...): _____

Original-Data format (xls, txt, csv, Access, etc.): _____

File name of the Import Wizard text files: _____

1. Accompanying data required for all data packets (metadata check)

	Old Data	New Data
Does the data package represent new data or existing ("old") data to be altered/replaced?	☐	☐

	Yes	No	N/A
Is it specified if "data owner" and "data provider" are identical?	☐	☐	☐
Has the data provider consented to the **data policy*** with his/her signature?	☐	☐	☐
Is it specified whether the data should be **freely available online*** after import into Edaphobase (in the Edaphobase portal and other biodiversity databases)? If there are restrictions, is the type of restriction specified? ☐ free availability agreed to ☐ the following restrictions have been requested: _____ _____	☐	☐	☐
Is decided whether the data should be temporarily anonymized (embargo)?? ☐ Embargo desired, duration: _____	☐	☐	☐
Do the entries in the Wizard match those in the **data sharing agreement***?	☐	☐	☐
Is the **data-set title*** (e.g. project, collection, report name, article title) unique and self-explanatory? [1]	☐	☐	

2. Control of the Import Wizard's quality checks

The Import Wizard provides a list of remaining ambiguities that was saved as a text file and sent to the data provider for correction. Here you must check whether the reported errors have been corrected and whether any new terms or taxa ("new items") should be included in Edaphobase.

	Yes	No	N/A
Were new terms or taxa ("new items") requested? If so, which?[2] _____ _____	☐	☐	
Are the "new items" understandable and has their inclusion in Edaphobase been approved?[3]	☐	☐	☐
Are there 'invalid' values or units which the Import Wizard could not interpret, e.g. kg/ha?	☐	☐	
If yes, are these units and values meaningful?	☐	☐	

[1] E.g.: „Projectname_Studyarea_Animalgroup_Samplingyear". If necessary, the Edaphobase team can assign this title after submission.
[2] If only a few terms, list here; otherwise refer to the corresponding file (with file name).
[3] The taxonomist responsible for the taxonomic group checks new taxa. Name here only other terms.

3. General quality control for errors in content

Geographical site/location

	Yes	No	N/A
Is the **name of the study site*** precise, complete and unambiguous [4]?	☐	☐	
Is the **name of the study area** precise, complete and unambiguous [5]?	☐	☐	
Is indicated how the geographical coordinates were determined and which system was used?	☐	☐	
Is a radius of uncertainty ("precision") specified for the geographical coordinates?	☐	☐	

Sampling Event

Is the **sampler's name** specified and written in full (no initials; applies to all names in the data set)?	☐	☐	☐
Do the **dates** make sense? (e.g. consistent, not in the future or the deep past)	☐	☐	
For **pitfall traps**: are **sampling duration** and **trapping liquid** clearly indicated?	☐	☐	☐
For **soil samples** and **chemical extraction**: Is the sampled **surface area** indicated or can it be calculated from the available data?	☐	☐	☐
For **soil samples**: Are **surface area** and **depth** indicated or can they be calculated from the available data?	☐	☐	☐
If a **sampling-event name** [6] has been assigned, is it unique and date- and method-specific?	☐	☐	
Are the **sample numbers** unique (especially within a site & sampling date)?	☐	☐	
If individual samples are subdivided into **subsamples** (i.e., depths): assignment to the individual sample specified (is it recognizable which belong together, for example, via a unique sample number)?	☐	☐	☐
Is it specified whether the samples represent **pooled samples**?[7] If yes, is given from how many samples it consists?	☐	☐	☐
Is the **Biotope type** specified? (if not, is it recognizable from the original citation of the site description and can be added?)	☐	☐	
Is indicated whether **(anthropogenic) Influence(s)** are present?	☐	☐	☐
If it is an **experimental study site**, is that specified??	☐	☐	☐

[4] At least one of the two fields "study area" or "study site", preferably the latter, must be filled with information; "plot" *can* be additionally indicated, if available. Place names must be written out, i.e. NOT abbreviated.

[5] Real toponyms, no descriptions like "cows grazing, soil quite wet"! Check the spelling of, i.e., German place names (München statt Muenchen oder Munich).

[6] If several sampling events have taken place in one area, these should be given a unique designation (especially for different dates & different methods/animal groups). If no number was assigned by the data provider, this is done automatically from site/plot x date x collection method.

[7] If not, number of samples >1 may indicate a composite sample. Then ask!

In the case of **arable fields**, is the current **crop(rotation)** indicated within the sampling event??

Is information on **vegetation** given, e.g. species of tree, shrub, herb, moss layer? [With **% cover**?]

Are **climate data** given?

Are **weather data** given?

Are data on **soil properties** given?

For collection objects: Is the **collection name*** specified?
- and the **object number**?
- Are the required Nagoya documents (**PIC and MAK forms**) stored? (only relevant for Seckenberg-internal data)

Quantities

Are zero values included? These must be removed for Edaphobase. If removed, the data provider must be informed.

Are whole numbers specified for "Number in sample" and "Number in collection"?

Taxonomy - Basic control

For **new taxa**, did the data provider specify the describing author, year (and bracket) and systematic classification?

For assessment of Determination Reliability:
- Is the **determination literature(*)** used indicated??
- Is it specified whether, and if so where, **voucher material** is stored?
- Is the **determinator** (species identifier) specified?
- Is it indicated whether, and if so by whom, the determination was checked?

4. General control for table-typical (technical) errors

	Yes	No	N/A
Is all information given in the **correct field** (no „Data Schizophrenia“)? Cell content in incorrect columns: _____ _____ _____	☐	☐	
Is information entered in “comment” fields that can possibly be moved to "analyzable" fields?	☐	☐	

Do information fields exist that contain several data and must be **atomized** (divided into individual fields) before import?
(E.g. “longitude and latitude”, “value and mean”, "Genus species Holotypus" in one field instead of several)?

Yes, the following: _____

_____

_____

Does the **same content type exist more than once** [8/9] ? If yes, which?:

_____

_____

_____

Is the data of a column (= information field) all available in a **consistent, uniform format**? (check for different date formats, number formats [incl. comma, period] or "sp., spec."; = Data inconsistency)

_____

_____

_____

Are all units of measurement given, understandable and correct?
(see column heading or in the measurement unit columns following the numerical value)

No, missing in: _____

_____

Is it clear whether **numerical values** represent individual values, mean values (with indication of number of base values used for calculating the means), min, max, standard deviations or value ranges (min-max)?

No, information unclear in:_____

_____

_____

For *min. and max. numerical values* (or range): are **average or individual values** also available?
(Only min-max specifications without mean values or individual values are unusable for grouping analyses.)

No, missing in: _____

_____

_____

When **abbreviations or internal codes** (e.g. habitat types) are used, is the full-text information also given (e.g. area name, habitat type)?

No, not in the following: _____

_____

[8] Such double information confuses the analysis routines of EdaphoStat, for example.
[9] Examples: pH value data from >1 measurement method (KCl/CaCl2/H2O), or several soil type data. If double information is available, priority can be assigned to one method and the remaining data transferred to the comment field without loss of information (if necessary after consultation with the data provider).

other common errors [10]

Umlauts should at least be used for German names, including place names (e.g. München instead of Muenchen).

Place names from German-speaking countries should be given in German, otherwise in the original spelling of the national language.

Completion of checklist control (technical assistant)

Data Controller 1: _____

Editorial measures [11] _____

_____

_____

_____

passed on (date) _____ to (name): _____

5. Taxonomic quality control (Controller-2)

The taxonomy check is to be carried out by the taxonomist responsible for the taxonomic group.

	Yes	No	N/A
Is the spelling of new taxa (to be added to Edaphobase) correct and are they valid species?	☐	☐	☐
Does the taxonomic concept used correspond to the one used in Edaphobase? If no: Comments to the concept: _____ _____	☐	☐	
Are voucher specimens to be requested for possible subsequent verification? (For which species?) _____ _____	☐	☐	
Is it possible to contact the determiner in case of discrepancies?	☐	☐	☐
Does the occurrence of the species in the specified area/habitat correspond to expert knowledge or do discrepancies occur (= are species listed which may occur very improbably there; e.g. indicator species of unpolluted mountain streams found in lowland fields; acidity indicating species on limestone; forest species on grassland)? _____ _____	☐	☐	
Are there any unknown or very rare species listed for the studied area that should be verified? _____ _____	☐	☐	☐

[10] Please expand list of common errors if necessary

[11] Major changes to the data set should only be made by the data provider. The data provider should be informed about minor corrections, e.g. typos.

Data Controller 2: _____

Editorial measures / recommendations:

_____

_____

_____

passed on (date) _____ to (name): _____

Literature

CAG (1998): Data Quality Tools for Data Warehousing – A Small Sample Survey. Center for Technology in Government, University at Albany / SUNY, 1535 Western Avenue, Albany, NY 12203, www.ctg.albany.edu

Chapman, Arthur D. (2005 a): Principles of Data Quality, version 1.0. Report for the Global Biodiversity Information Facility, Copenhagen.

Chapman, Arthur D. (2005 b): Principles and methods of data cleaning – Primary species and species-occurrence data, version 1.0. Report for the Global Biodiversity Information Facility, Copenhagen.

DataONE Education Module: Data Quality Control and Assurance. DataONE. Retrieved Nov12, 2012. From http://www.dataone.org/sites/all/documents/L05_DataQualityControlAssurance.pptx

Hellerstein, Joseph M. (2008): Quantitative Data Cleaning for Large Databases. http://db.cs.berkeley.edu/jmh/papers/cleaning-unece.pdf, acc. 28.7.2016