Herbarium Specimen Label Interpretation and Transcription: First steps used to clean digitized data

Henry Engledow ‡

‡ Botanic Garden Meise, Meise, Belgium

Corresponding author: Henry Engledow (henry.engledow@plantentuinmeise.be)

Abstract

At present many herbaria and musea around the globe are digitising their natural history collections. Capturing the label information on these specimens is crucial to finding these specimens and using their data. To this end, Meise Botanic Garden chose to record minimal data as part of its second mass digitization project. The collections digitized, dating back to the beginning of the 19th century, are diverse and poorly curated. Diversity of collectors, geography, languages and conventions increase the complexity of label interpretation and transcription. Examples from these records serve to illustrate both the problems and solutions to producing clean data.

Label transcription was outsourced to the commercial company Picturae, who subcontracted Alembo to do the transcription. Quality control on random subsets of the data was regularly carried out by Alembo, Pictuare and Meise Botanic Garden. Despite the data being delivered at a high standard, extensive data cleaning was required (12-60% of the fields needed adjustment depending on the field). The amount of data cleaning was a function of the type of field, whether it was standardised or free text, as well as the length of the string (the longer the string the greater the variance).

The main issue associated with data quality was legibility. Two factors were at work here: 1) orthography of the collectors was to some degree illegible; 2) the label information was obscured by something (often plant material). As transcribers only get to see a single specimen image at a time, this does not allow them to compare similar labels by the same collector. As a result the information transcribed is the best interpretation of the specimen label. The advantage of mass digitization and transcription, is that one can compare and analyse the data once the project has finished. Collectors often collect multiple specimens at the same collecting site, and herbaria tend to have many specimens from the same collector. These two factors allow information to be grouped and sorted, allowing for mass editing and cleaning of the data.

The major transcription errors include:

- Misinterpretation of certain characters e.g. a & o; 1 & 7; u & n; w & m; z & s, etc.;
- Different versions of 'same' characters e.g. º vs °; «» vs " " vs “ ” vs ‘ ’; ß vs ss; æ vs ae; etc.;
- Inconsistent use of accented characters e.g. o ò ó ô õ ö ǒ ø; ij ÿ; l ł; s š; u ü ū ũ ǔ; etc.;
- The use of varying punctuation results in divergent strings;
- Non-visible characters e.g. tabs, invisible spaces, line feed, etc.;
- The switching of characters or numbers inadvertently;
- Data entered into the wrong field due to confusion amongst taxa, collectors & places;
- Data on the herbarium label that has been previously incorrectly transcribed from original label data resulting in wrong data (not technically a transcription error in this project);
- The order in which the information was entered is variable (problem for long strings e.g. verbatim locality);
- Transcriber not familiar with language on the label versus those who are.

First steps taken in data cleaning

- Explore your data set, followed by some basic analysis;
- Clean obvious mistakes:
    - Sort data, many things will group naturally;
    - Group data using 'keywords';
    - Trim fields;
    - Remove hidden characters;
    - Correct obvious spelling errors;
    - Standardise common symbols, accents & characters.
- Standardise & normalise data where possible
- Regroup or sort partially cleaned data

- Use the data to clean the data e.g.
    - Standardised collector & country code fields help to clean incomplete date fields;
    - Country code can help correct the collector field & vice versa;
    - Collector birth & death dates can be used to inform interpretation of the collection year.
- Repeat the above steps until most of the “noise” is reduced.

When one first looks at the data it appears messy and full of errors, like an unsharp image. The first phase of data cleaning is to reduce as much of the noise as possible, at this stage the image starts to come into focus. The second phase is data validation, where inconsistencies or incongruencies in the data are identified and corrected. The third phase is augmentation and linking of data; in this phase the specimen data is expanded to link to other sources of related information thereby giving added value to the specimens. In reality, these processes can happen simultaneously, but the bulk of the changes follow these stages. The first phase is almost entirely a manual process, but as these changes are often done in batches it proceeds at a reasonable pace. The remaining phases were corrected using automated and manual processes. Data cleaning is not a one-off process, it takes more time than is often allocated to this phase in projects. In time, cleaning data errors, will eventually bring the picture into focus.

Keywords

mass digitization, transcribed data, data cleaning, transcription, standardise, normalise, linked data

Presenting author

Henry Engledow

Presented at

TDWG 2022

Acknowledgements

I would like to thank Sofie De Smedt & Ann Bogaerts for quality control & data cleaning; the DOE! Team for quality control; Picturae & Alembo for transcription & quality control; and the Flemish Government for financing the project.

Conflicts of interest