BISS Biodiversity Information Science and Standards

Biodiversity Information Science and Standards 2: e26514
doi: 10.3897/biss.2.26514

OPEN ACCESS

Conference Abstract

An Evaluation of *In-house* versus *Out-sourced* Data Capture at the Meise Botanic Garden (BR)

Henry Engledow‡, Sofie De Smedt‡, Ann Bogaerts‡, Quentin Groom‡

‡ Meise Botanic Garden, Meise, Belgium

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

Received: 08 May 2018 | Published: 21 May 2018

Citation: Engledow H, De Smedt S, Bogaerts A, Groom Q (2018) An Evaluation of *In-house* versus *Out-sourced* Data Capture at the Meise Botanic Garden (BR). Biodiversity Information Science and Standards 2: e26514. https://doi.org/10.3897/biss.2.26514

Abstract

There are many ways to capture data from herbarium specimen labels. Here we compare the results of in-house verses out-sourced data transcription with the aim of evaluating the pros and cons of each approach and guiding future projects that want to do the same.

In 2014 Meise Botanic Garden (BR) embarked on a mass digitization project. We digitally imaged of some 1.2 million herbarium specimens from our African and Belgian Herbaria. The minimal data for a third of these images was transcribed in-house, while the remainder was out-sourced to a commercial company. The minimal data comprised the fields: specimen's herbarium location, barcode, filing name, family, collector, collector number, country code and phytoregion (for the Democratic Republic of Congo, Rwanda & Burundi). The out-sourced data capture consisted of three types:

1. additional label information for central African specimens having minimal data;
2. complete data for the remaining African specimens; and,
3. species filing name information for African and Belgian specimens without minimal data. As part of the preparation for out-sourcing, a strict protocol had to be established as to the criteria for acceptable data quality levels.

Also, the creation of several lookup tables for data entry was necessary to improve data quality. During the start-up phase all the data were checked, feedback given, compromises

made and the protocol amended. After this phase, an agreed upon subsample was quality controlled. If the error score exceeded the agreed level, the batch was returned for retyping. The data had three quality control checks during the process, by the data capturers, the contractor's project managers and ourselves.

Data quality was analysed and compared in-house versus out-sourced modes of data capture. The error rate by our staff versus the external company was comparable. The types of error that occurred were often linked to the specific field in question. These errors include problems of interpretation, legibility, foreign languages, typographic errors, etc. A significant amount of data cleaning and post-capture processing was required prior to import into our database, despite the data being of good quality according to protocol (error < 1%). By improving the workflow and field definitions a notable improvement could be made in the "data cleaning" phase.

The initial motivation for capturing some data in-house was financial. However, after analysis, this may not have been the most cost effective approach. Many lessons have been learned from this first mass digitisation project that will implemented in similar projects in the future.

Keywords

mass digitisation, crowd sourcing, citizen science, data quality control

Presenting author

Henry Engledow

Acknowledgements

Elke Scheers; Kathy Peeters & Anne De Groote