Grant Agreement Number: 777483 / Acronym: ICEDIG
Call: H2020-INFRADEV-2017-1 / Type of Action: RIA
Start Date: 01 Jan 2018 / Duration: 27 months

REFERENCES:

Deliverable D4.2 / R / PU
Work package 4 / Lead: APM
Delivery date 13M

ICEDIG.EU

Innovation and consolidation for large scale digitisation of natural heritage

Report on New Methods for Data Quality Assurance, Verification and Enrichment

Deliverable 4.2

Authors: Sarah Phillips[1], Mathias Dillen[2], Quentin Groom[2], Laura Green[1], Marie-Hélène Weech[1], Noortje Wijkamp[3*],

1. Royal Botanic Gardens, Kew, United Kingdom
2. Meise Botanic Garden, Meise, Belgium
3. Picturae BV, Heiloo, Netherlands

*Chapter 5

Funded by the Horizon 2020 Framework of the European Union
H2020-INFRADEV-2016-2017
Grant Agreement No 777483

ICEDIG.EU

Contents

Executive Summary 2

1. Introduction 3

2. Work of the Biodiversity Information Standards Group 4

3. Transcription Pilots 5

3.1 Transcription Pilot one 5

3.1.1 Assessing data quality 7

3.1.2 Pilot One Results 13

3.1.3 Discussion Transcription Pilot 1 36

3.2 Transcription Pilot Two 37

3.2.1 Pilot Two Results 38

3.2.2 Discussion Transcription pilot 2 46

3.3 Recommendations for Transcription 49

4. An Electronic Marketplace for Transcription 50

5. Organising a Large Transcription Project: Insights from a Commercial Party 53

5.1. Starting up a large-scale transcription project 53

5.1.1. Transcription rules 53

5.1.2. Logistics of a large-scale transcription project 54

5.1.3. Training new transcribers 57

5.2. Process and workflow 57

5.3. Final stage of the project 60

5.4. Ideas on using an electronic marketplace for transcription services 61

5.5. Recommendations 62

6. Georeferencing 64

7. Recommendations for DiSSCo 69

8. References 71

9. Acknowledgements 74

Funded by the Horizon 2020 Framework of the European Union
H2020-INFRADEV-2016-2017
Grant Agreement No 777483

Executive Summary

Distributed Systems of Scientific Collections (DiSSCo) will facilitate the production of tens of millions of natural history specimen collection images along with their labels each year. The labels of these specimens contain valuable information for research studies, but their transcription can be very difficult and time consuming with often hard to read handwritten labels. Whilst accurate label transcription is only one step along the way to create a specimen record fit for different research uses, it is an extremely important one. It would be very time consuming to have to return to recheck label information for even a very small proportion of specimens. Once a specimen is transcribed correctly it becomes much easier to enhance the record with additional information from other sources, e.g. from literature or collector itineraries, determine the point of collection from the textual information on the label by a process known as georeferencing, or even to find inaccuracies within the label itself. This document discusses and compares different approaches for the efficient accurate transcription of these labels. Using Herbarium specimens as an example, the quality of transcribed data by in-house trained institute staff, outsourced to a commercial company or transcribed by the general public through online crowdsourcing platforms was compared. Key transcription data was assessed and common errors in label transcription identified. Reasons for these errors are discussed along with possible mechanisms to improve the accuracy of the transcriptions. The need for standards for transcription was identified and recommendations made.

This document also considers different options for outsourcing transcription work including the possibility of using an electronic marketplace website where tasks can be assigned to workers. The outcomes of its use in other domains has been reviewed including ethical concerns. It also discusses the issues and procedures for organising a large transcription project from the viewpoint of a commercial party including transcription rules, logistics of running a large-scale transcription project, training new transcribers and the workflow of the project itself. Finally outlining a set of recommendations based on the commercial company's experience of running a large-scale transcription project.

To maximise the use of specimens in research it is necessary to determine the point of collection from the textual information on the label by a process known as georeferencing. This process is currently extremely time-consuming and hence very costly. Therefore, methods to speed up georeferencing need to be investigated for implementation. A review of the available tools for automating some of these processes is discussed.

This report highlights several recommendations for DiSSCo to enhance the quality of specimen records, including providing tools to aid label transcription, georeferencing and data cleaning to try and maximise their fitness for use.

Funded by the Horizon 2020 Framework of the European Union
H2020-INFRADEV-2016-2017
Grant Agreement No 777483

1. Introduction

It is anticipated that Distributed Systems of Scientific Collections (DiSSCo) through its automated imaging facilities will enable the production of tens of millions of images of specimens and their labels each year and the metadata from these natural history collection labels will need to be transcribed to enable information management, retrieval and their use in scientific research. The most efficient means of transcription therefore need to be determined. However, label transcription is a time-consuming task. Historic collection labels are often written in difficult-to-read handwriting, in many different languages, which currently means that humans often need to complete transcription and this task cannot be completely automated. There are many methods of transcription including in-house by trained institute staff, outsourcing to a commercial company or asking the public to aid with transcription through crowdsourcing platforms. The sources and types of errors between these transcription methods will differ and the quality of the transcriptions returned need to be analysed and assessed to assist with the determination of the most efficient methods. This task will also consider options for the outsourcing of transcription work and the steps that would need to be taken to start a large-scale transcription service. An option that may be considered is an electronic marketplace which can assign workers with tasks to transcribe. This method is used in other domains for services and outsourcing, and the use of such a system for transcribing natural history specimens should be evaluated.

For many research needs it is necessary to know the location from where the specimen was collected. Modern collectors take GPS coordinates at the point of collection but historical label data is without these coordinates. To maximise the use of specimens in research it is necessary to determine the point of collection from the textual information on the label by a process known as georeferencing. This process is extremely time-consuming. Therefore, methods to speed up georeferencing should be investigated for implementation.

Ultimately, the process of extracting data from the labels of specimens is a long, complicated and iterative process. As a result, it is also one of the more time-consuming and costly processes in the digitisation of a collection. Even small reductions in processing time and improvements in quality can amount to significant savings when multiplied by the millions of specimens that need to be processed. Therefore, this report is of particular importance in making collection digitisation cost-effective. In this report we evaluate methods that are currently being used for transcription, but also look to the future at what could be changed and what new methods might emerge. We also point readers to the parallel ICEDIG report 4.1 Methods for automated text digitisation, that sets out in more detail automated methods for extracting data (Owen et al., 2019).

Funded by the Horizon 2020 Framework of the European Union
H2020-INFRADEV-2016-2017
Grant Agreement No 777483

2. Work of the Biodiversity Information Standards Group

A vast amount of work on data quality has been completed by the Biodiversity Information Standards (TDWG) Data Quality Interest Group (DQIG, https://www.tdwg.org/community/bdq/) established in 2014. The goal of the Interest Group, as stated on their website, has been "to discuss, determine, formalise and standardise concepts, problems, policies, metadata, methodologies and mechanisms related to biodiversity data quality, collaboratively and incrementally, and to promote best practices throughout the biodiversity informatics community".

The work of the DQIG has been split up into four task groups:

Task group 1 – Framework on data quality
Task group 2 – Data quality tests and assertions
Task group 3 – Data quality use cases
Task group 4 – Best practices for development of vocabularies of value

A formal conceptual framework for the assessment and management of the fitness for use of biodiversity data has been developed (Veiga *et al.*, 2017). The aim of the framework is to support the biodiversity informatics community, allowing for the description of the meaning of "fitness for use" from a data user's perspective in a common and standardised manner. By necessity, the framework appears quite complex to anyone not familiar with the data quality landscape and would be difficult to interpret by collections data managers without sufficient study. To aid understanding and implementation of the framework, it was evaluated by Veiga *et al.*,2017, with a case study conducted in the Museum of Comparative Zoology of Harvard University using a dataset from Arizona State University Hasbrouck Insect Collection. This proof of concept provides an example illustrating how to implement the framework, although it is still quite a complex task.

The second task group, the Data Quality Use Cases Group is collating a list of use cases where data are assessed for their suitability for that particular use case. The use case descriptions will include the data required, quality dimensions and thresholds used to assess the data or dataset. This should provide a reference set of information that can be used to inform the data user how to assess the suitability of the data for a particular purpose. Data quality use cases are currently under development https://github.com/tdwg/bdq/blob/master/tg3/README.md.

The Data Quality Tests and Assertions Task Group will provide a report of the practical tests, assertions, principles, software and key references associated with assessing data quality of biodiversity records. The aim is that an internationally agreed standard suite of core tests and resulting assertions can be used by all data providers and hopefully data collectors. As stated by James *et al.*, 2018, the standard tests being developed by the DQIG will be implemented by data collectors for use in the field; by data aggregators such as Integrated Digitized Biocollections (iDigBio; https://www.idigbio.org/), the Atlas of Living Australia (ALA; https://www.ala.org.au), and GBIF (https://www.gbif.org/); by ancillary services such as Kurator (Morris et al., 2017; http://wiki.datakurator.org/); by data users; and by collections data custodians.

Funded by the Horizon 2020 Framework of the European Union
H2020-INFRADEV-2016-2017
Grant Agreement No 777483

Rather than replicate the work of the DQIG in this task we decided to focus on the process of label transcription. Whilst accurate label transcription is only one step along the way to create a specimen record fit for different research uses it is an extremely important one. It would be extremely time consuming to have to return to recheck the label information for even a very small proportion of specimens. Once the specimen is transcribed correctly it becomes much easier to enhance the record with additional information from other sources.

3. Transcription Pilots

To compare the data quality of transcription obtained through different methods, two specimen transcription pilot projects were undertaken. The different methods to be compared were:

a) In-house transcribers
b) Expert taxonomist transcribers
c) Outsourced commercial transcription
d) Volunteer crowdsourced data

The first pilot compared all four methods but was limited to specimens from one genus and two institutions due to limited access to the uncleaned transcriptions from outsourced commercial transcription projects. The second pilot compared volunteer crowdsourced data to in-house transcribed data but covered more taxa and specimens from 7 institutions.

3.1 Transcription Pilot one

A test dataset was compiled from 200 S*olanum* specimen images from the Royal Botanic Gardens, Kew, in the UK and 200 from Meise Botanic Garden in Belgium. This particular genus was chosen as both institutes had specimens from which the label data had already been transcribed through multiple methods. Both contributing institutes had outsourced label transcription through the digitisation company Picturae (https://picturae.com/) completed by Alembo (https://alembo.nl/) and both institutes had label data transcribed by a taxonomist with expertise in the family Solanaceae. The Kew specimens had also been transcribed in-house by staff employed as digitisation officers and curators. A sample of S*olanum* images from both institutes were uploaded to two crowdsourcing platforms: DigiVol, run by the Atlas of Living Australia (https://digivol.ala.org.au/); and DoeDat (https://www.doedat.be/), the crowdsourcing platform of Meise Botanic Garden which is an adaptation and configuration of DigiVol. Kew put both sets of specimens up into one expedition to help with the signposting of the expeditions to Kew volunteers. Many other institutes can use the DigiVol platform and volunteers can choose to transcribe for any expedition. However, to encourage transcription, Kew can email volunteers that have already transcribed for Kew's expeditions, to thank them for their contribution and at the same time point them to a new expedition. Some volunteers may have loyalty to Kew through past interactions, or volunteering for Kew could be part of the motivation for completing the transcription. Meise chose to put up two separate expeditions: one for Meise specimens and one for Kew specimens.

Funded by the Horizon 2020 Framework of the European Union
H2020-INFRADEV-2016-2017
Grant Agreement No 777483

The different transcription platforms had different protocols, which makes direct comparisons of the results more complex. Differences will occur for many reasons, including but not limited to: constraints on the type of platform or software used to transcribe the data; the differences in preferences on formats for each institute, which in itself is often due to differences between the collection management systems used by the institutes; and the amount of interpretation that is being requested. Currently, standards for transcription are lacking, leading to this diversity in protocols. A summary of the differences for key fields is provided below:

Collector Name Field:
Transcription of collector names are handled differently in all the different protocols.

Volunteer crowdsourced data:
In the template used by Kew within DigiVol, each collector is entered in a separate field in the order shown on the label with a request for the format to be entered as: surname, initials (e.g. Maitland, T.D.). There is an inbuilt collector lookup in the system, but this displays previously entered values and not an authorised collector list. In DoeDat there are two collector fields, one field "Collector(s) as given" asks the transcriber to record the collector name(s) as written on the label, the other field "Collector (standard)" contains a dropdown list from which a standardised version of the primary collector can be selected if it is in the list.

Outsourced commercial transcription:
The Kew specimens were transcribed by Alembo using Brahms v7 (Botanical Research and Herbarium Management System) 7 (https://herbaria.plants.ox.ac.uk/bol) and its Rapid Data Entry form (RDE). There were two fields: a "collector" field for the primary collector and a second "addcollector" field where any additional collector names could be transcribed, each one separated by a semicolon. The format specified was: surname, initials (e.g. Maitland, T.D.). There was a lookup list available for each of the collector fields, but it was far from comprehensive. For Meise specimens, rather than using Brahms, Alembo developed an in-house MySQL database system for use by both Alembo and Picturae for transcription. In the collector field there was a lookup for collector or a group of collectors from which transcribers could select. An additional field, "Collector id", was filled in automatically when one of these was selected, the id relating to the primary collector. If the name was not on the list then the transcriber was instructed to write the name verbatim, i.e. as written on the label.

Collector number Field:
Collector number was a single field in all of the methods used. All protocols asked for the number verbatim from the label and both protocols for Alembo specified that this included any prefixes or suffixes. However, the Meise Alembo protocol stated not to include letters if they were the same as the name of the collector or his/her initials.

Collection date(s) Field:

Volunteer crowdsourced data:
In both DigiVol and DoeDat, collection date(s) were transcribed in six separate fields in the format DD MM YYYY - if a collection date range was indicated on the label a start date and an end date could be entered. Day and month were to be entered in a two-digit format and year in a four-digit format. For

Funded by the Horizon 2020 Framework of the European Union
H2020-INFRADEV-2016-2017
Grant Agreement No 777483

DoeDat if a day or month was specified on the label but no year, the DoeDat protocol indicated to add 3000 in the year field to indicate the year was not present on the label.

Outsourced commercial transcription:

For the Kew specimens, collection dates were also transcribed by Alembo in six different fields, however the entries for day and month defaulted to a single digit format where the value entered was only a single digit value rather than the format DD MM. For example, 7th September on the label was entered as 7 in the day field and 9 in the month field. Year was also transcribed in the format YYYY. There was an additional field called "date text" where transcribers could note if no date was present on the label or enter incomplete or illegible dates. When a year but no century was indicated on the label, the protocol stated to check the collector list to see if the collector on the label was present in the lookup list with collection date information to see if they could infer the century from that information. However, if that provided no helpful information then the year without a century was entered in the "date text" field. For Meise specimens, rather than six date fields, only three date fields were used by Alembo. Any collection date ranges on the label were captured in an additional "Date as given" field and entered as free text exactly as they were written on the label. Transcribers were also asked not to interpret the century if it was missing from the label so a year format without the century was acceptable in the year field.

3.1.1 Assessing data quality

According to Veiga et al.'s 2017 Conceptual Framework on Biodiversity Data Quality, to assess data quality (DQ), it is first necessary to establish a DQ profile, by describing a use case, valuable information elements, and measurement, validation and enhancement policies. We attempted to follow this framework, from the perspective of data curators who are not developers. Our use case was that the people who use our data often wish to find records of specimens of a particular taxon, from a particular collector, collection event (often identified by collector, collector number and date combination) or from a particular locality. As such, the valuable information elements in this case, corresponding to a controlled vocabulary (Darwin Core) were:

Plant name: dwc:scientificName
Collector: dwc:recordedBy
Collector number: dwc:recordNumber
Collection date: dwc:eventDate
Country: dwc:country
Barcode label number: dwc:catalogNumber (not assessed but an essential key to the data)

Measuring quality in data can prove problematic particularly where no uniform data standard or data structure has been determined. Recognising the initial goals set for each of the data collection methods and the differences in the outlined data collection protocol is important when trying to assess the level of quality in each of the datasets. In order to try to address this issue, data quality can be investigated under a number of different possible DQ dimensions. [Table 1.] lists and defines the dimensions that were selected in order to assess the quality of the different datasets in this study. This corresponds the DQ measurement policy of the framework.

Funded by the Horizon 2020 Framework of the European Union
H2020-INFRADEV-2016-2017
Grant Agreement No 777483

DIMENSION	DEFINITION
Completeness	The field contains a value where there is relevant data on the label.
Accuracy	The extent to which the given value in a particular field from a particular dataset corresponds to that of the known reference value.
Similarity	The extent to which the given values in a particular field correspond across all datasets.
Validity	The value within this field is recorded in the correct location.
Format	The value within the field complies with the required data formatting.
Standardisation	The value of the field corresponds to a recommended, recognised standard vocabulary.

Table 1: Definitions of Dimensions Selected to Assess Dataset Quality

Our DQ validation policy was relatively simple and was as follows:
Plant name: dwc:scientificName has to be accurate, complete, valid and standardised in each single record to be compliant.
Collector: dwc:recordedBy has to be accurate, complete, valid and standardised in each single record to be compliant.
Collector number: dwc:recordNumber has to be accurate, complete, and valid in each single record to be compliant.
Collection date: dwc:eventDate has to be accurate, complete, valid and standardised in each single record to be compliant.
Country: dwc:country has to be accurate, complete, valid and standardised in each single record to be compliant.

We did not assess multi-record records or check for consistency between fields within single records, though this could be possible with further investigation.

We applied the following enhancement policy:
Recommendations for plant names, collectors, collection dates and countries were made for each of these fields based on services providing list matching services such as IPNI and Catalogue of Life, Harvard List of Botanists, Excel (ISDATE function) and the ISO-3116 standard.

Dimensions such as completeness and format allow for relatively simple evaluation, comparing absence and presence of data, or comparing the format of the selected dataset field with that required in the protocol. Other dimensions such as accuracy and similarity prove to be more complex. Accuracy is often recognised as one of the key measures of quality in data, however there are a great many other dimensions that can prove just as useful for evaluating data quality (Sidi, F. et al. 2013). In addition, whilst measurement of accuracy does provide important information regarding data quality, in order to ascertain a measurement of accuracy, the data must be compared to a known reference

Funded by the Horizon 2020 Framework of the European Union
H2020-INFRADEV-2016-2017
Grant Agreement No 777483

value (Cai and Zhu 2015). The task of measuring accuracy of data collected from specimen labels, where no previous known reference value exists, is therefore likely to be both desired but also unachievable.
In the absence of a known reference value, the possibility that similarity (to what extent each of the datasets agree with each other) may be used as a suitable proxy for measuring accuracy could be tested. Where all datasets agree, it may be that an assumption of complete accuracy can be made. In order to investigate the data under the different dimensions, a number of different approaches were followed.

The analysis for the data sets in pilot one was split into two initial workflows. The first analysis was performed by taking a matching approach, comparing several of the datasets directly against each other. The datasets chosen for this task were the Solanum records crowdsourced on DigiVol and DoeDat and outsourced to Alembo as both Meise and Kew possessed data from these sources allowing direct comparisons. This approach tested the dimension of similarity. It is important to acknowledge that some of the data quality dimensions are likely to be dependent on others (ref) and so this approach addressed the dimension of format due to differences in protocol and factored out certain differences before addressing similarity.

The second approach taken was to compare the transcriptions based on the work completed by the TDWG Data Quality Interest Group. In order to assess the suitability of differing transcription approaches, an agreed outcome on expected transcriptions must be ascertained. In order to retain consistency and avoid subjectivity or re-deciphering of the same labels on multiple occasions, it was decided that a data gold standard should be first agreed for each specimen in order to provide a known reference value. In order to do this, parameters were agreed for each data field found on a specimen label, outlining the expected format of the data captured (discussed below). The creation of the gold standard datasets requires data collection of the whole specimen, taking into account true transcription and also format and so is a very time-consuming task. Because of this, it was decided that only a subset of the *Solanum* data would be selected for capture. As there was also a dataset held that was created by in-house staff for the Kew specimens, we decided that the gold standard data would be created for these specimens as it would allow for additional analysis of the quality level created by staff working with the collections. The analysis of the data using the gold standard approach evaluated the different data quality dimensions accuracy, format, validity and standardisation independently of each other.

Matching approach method

Three principal fields: collection date(s), collector(s) and collector number, were identified for comparison and used to investigate whether the values transcribed were the same across all the platforms.

Collection dates were compared by looking for exact matches, while taking into account trivial differences, which were the result of differences in formats and protocols. Collector names were considered as free text as transcribers were not always selecting from a dropdown list in every case. Therefore, an analysis was completed using fuzzy matching metrics and the Levenshein distance that strongly suggest similar transcription values. Levenshtein distance (LD) is a measure of the similarity between two strings, the source string (*s*) and the target string (*t*). The distance is the number of

Funded by the Horizon 2020 Framework of the European Union
H2020-INFRADEV-2016-2017
Grant Agreement No 777483

deletions, insertions, or substitutions required to transform *s* into *t*. The greater the Levenshtein distance, the more different the strings are. The workflow for the analysis is outlined in Fig. 1.

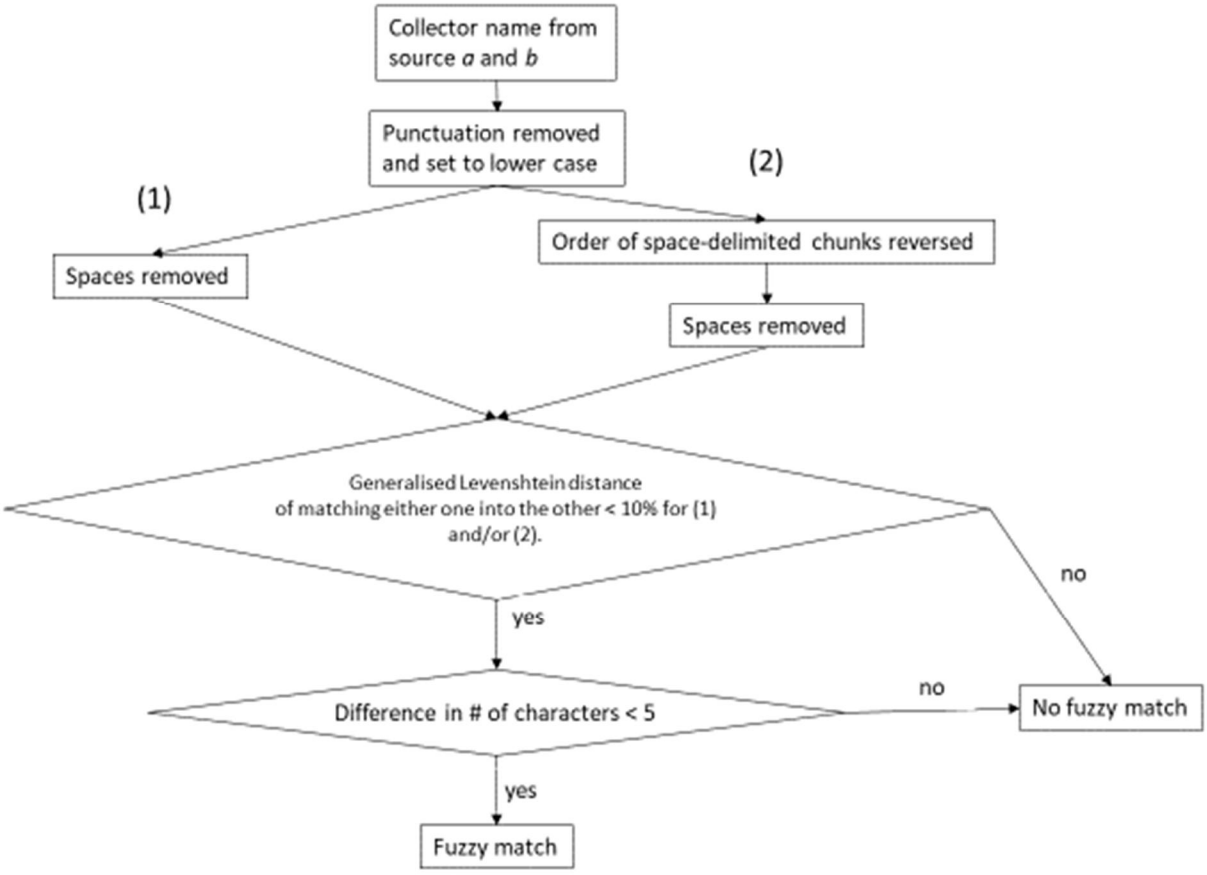


Fig. 1. Workflow diagram for fuzzy matching for collector strings.

The first step in the workflow was to remove all punctuation and set all characters to lowercase from the collector string, this was completed on the assumption that we are not so concerned with errors with punctuation as this can be more easily corrected in any cleaning stage. Then three different matching conditions are applied.

1. For the first condition all spaces were removed from the collector string and then one string was matched into the other string and vice versa. If the Levenshtein distance was 10% or less when the string from source a was matched into the string from source b or vice versa then the strings were considered to match. Positive matches would include exact matches, some typos, the inclusion of a middle name in one string but not the other or some encoding errors. There may be some false positive matches for collectors with almost matching names or for very short generic strings such as “de” or “dr” which would match together.

2. For the second condition the order of the space-limited sections of text were reversed before the spaces were removed. If the Levenshtein distance was 10% or less when the string from source a was matched into the string from source b or vice versa then the strings were considered to match. Inversion between first name and surname can often occur depending on the format

Funded by the Horizon 2020 Framework of the European Union
H2020-INFRADEV-2016-2017
Grant Agreement No 777483

specified in the transcription protocol, how the name is written on the label, or the standard used in the source of the data. Collector names with double barrelled surnames or with initials separated by spaces are still not matched this way.

3. If a positive match is given for both or either of condition one or condition two then condition three is applied. For condition three the difference in character length between both transcriptions, excluding punctuation and spaces, must be 4 or less to be considered a match. This captures mismatches when names of additional collectors are missing or erroneously added. However, it may also cause false negatives in the case of very long first names, but these are often abbreviated. This condition also avoids the issue present in the first condition where very short generic transcriptions e.g. "dr" or "SC" match to many other terms.

If comparing collector strings from three data sources, A, B and C. Then the following matches would apply – A to B, B to A, A to C, C to A and B to C and finally C to B.

Gold Standard approach method

It is important to note that the initial assessment of these data is gauging levels of direct transcription, reproducing information from the label verbatim, without correcting inaccuracies.

We randomly selected 20 records (10 percent) from the Kew *Solanum* dataset (200 records), for three assessors to transcribe independently. The initial aim for this subset of transcriptions was to assess the consistency of data transcription achieved between the three assessors. A suitable level of consistency would allow the remaining 180 records to be allocated between the assessors and the gold standard data to be transcribed independently. If the level of transcription for these 20 records proved inconsistent between the assessors, then the remaining records could not be reliably transcribed by just one person alone.

Statistical tests such as the kappa test (McHugh 2012, Kirilenko & Stepchenkova 2016) and Krippendorff's alpha (Krippendorff 2011) were investigated for suitability in evaluating inter-rater reliability (IRR). Both of these tests are appropriate for analysing categorical data and the probability of random agreement must be ascertained. However, the nature of herbarium specimen labels is such that it can contain an unlimited scope of data, which renders the calculation of random agreement probability extremely problematic to calculate. For this reason, a more holistic approach was taken when producing the gold standard data.

The results of the initial 20 records were compared, any inconsistencies were flagged and an agreement was made on the correct entry. Based on this, a protocol for the transcription was agreed. The remaining 180 records were split between assessors for transcription, however each data record created was checked independently by the other two assessors, to ensure agreement on the gold standard. Any further discrepancies were discussed as above.

Once we had ascertained the gold standard data for these 200 specimens, five principal fields were identified for comparison: collection date(s); collector(s); collector number; country and plant names. As these fields were to be used for analysing several data quality dimensions, the data were recorded both exactly as they appeared on the label, but also in an agreed protocol format for transcription.

Funded by the Horizon 2020 Framework of the European Union
H2020-INFRADEV-2016-2017
Grant Agreement No 777483

This allowed for the data to be compared for accuracy, but also to be adapted to display in the expected format for each of the different datasets. Each dataset was independently compared to the gold standard dataset and relevant dimensions analysed for each of the principal fields.

A .csv file created for each principal field, containing values for each record from the gold standard dataset. This was also done for each of the five other datasets. Each file contained the record id and the corresponding principal field value. The online tool Diff.text was used to run a comparison for each dataset's principal fields against those of the gold standard data and analysed for additions, deletions, substitutions and movements at character level. This tool was chosen for its ability to identify text movement within the field, which would otherwise be identified as an insertion and or deletion. The Levenshtein distance was then calculated using the stringdist R package. The inclusion of movement in the analysis using the Damerau-Levenshtein distance algorithm is possible, however it only allows for adjacent character transposition rather than full string movement and so movements were not used in the calculations but were noted when present. Any values that displayed differences from the expected gold standard values were analysed and categorised based on the type of difference.

Accuracy

The following categories were used to measure accuracy: Match, when the value matched the gold standard value exactly; Added when the value was complete but wholly absent from the sheet, and therefore the gold standard value; Typo/Misinterpretation, when the value differed from the gold standard in only a few characters (mostly substitutions or movement); Additional characters, when the value contained additional characters not present in the gold standard value (such as a prefix or suffix in collection numbers); and Missing characters, when the value had many, but not all, characters present in the gold standard.

Completeness

Absence or presence of a data entry in each of the principal fields for each record was compared against the gold standard data fields for each of the different datasets. Values were classed as Incomplete when the field was left blank where there was a value present in the golden standard dataset. By subtracting the number of values classed as "incomplete" from the total number of values in the field, using the results of the method above, we could ascertain both when fields were left empty when there were data present on the label and when a recording of no data was incorrectly made.

Validity

As above, values were classed as Incorrect, when the value was complete but wholly different from the gold standard value. It is possible to ascertain the validity levels for each platform by subtracting the number of values classed as "incorrect" from the total number of values in the field, using the results of the method above. This shows how many values in the dataset were transcribed with the transcriber looking at the correct piece of information on the label for that field.

Format

Values were classed in the Format category when the value had been atomised differently from the gold standard (e.g. in collector name - a surname initial placed with other initials or with surname, or in collector number – punctuation included or left out) The provided data collection protocol for each of the datasets was checked and an expected format for each principal field was ascertained. Due to

Funded by the Horizon 2020 Framework of the European Union
H2020-INFRADEV-2016-2017
Grant Agreement No 777483

the variation in format requirements for each of the platforms as well as the export formats, several of the fields had to be reformatted in order to assess other dimensions.

Standardisation
Finally, standardisation of data was checked against certain lookup lists such as the Harvard List of Botanists, by using their REST service, fetching URLs including collector names as query parameters in OpenRefine, and the International Plant Names Index (IPNI) values, using the IPNI names reconciliation tool in OpenRefine. Some of the dataset protocols requested that collector names and plant names be transcribed as they appeared on the label, so levels of measured accuracy might have been lower if a transcriber used some interpretation. Measuring standardisation allows for the added value of interpretation to be acknowledged in data quality.

3.1.2 Pilot One Results

The joint Kew and Meise *Solanum* expedition on DigiVol was completed in 15 days with 23 different volunteers. Following the typical pattern of transcriptions in crowdsourcing projects two contributors completed 34.5% of the expedition and the top five contributors completed 79.8% of the records.

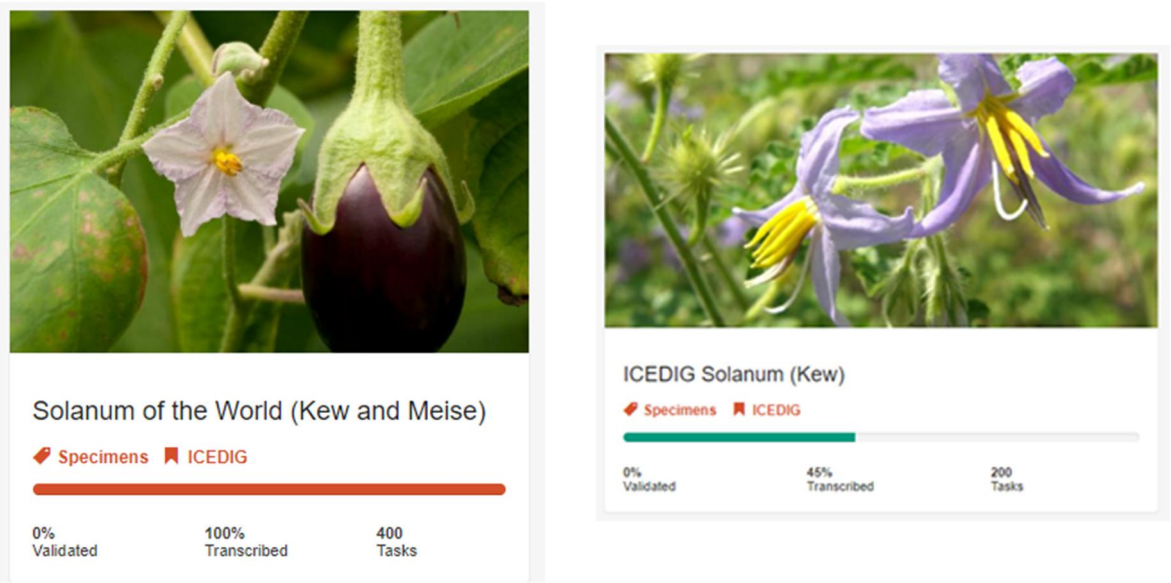


Fig. 2. Solanum crowdsourcing expeditions on DigiVol and DoeDat.

The DoeDat Meise specimens *Solanum* expedition was completed in just 3 days by 12 volunteers but this quick turnaround was due to an onsite transcription event. One transcriber contributed 35.5% of transcriptions. The Kew specimen *Solanum* expedition took much longer 172 days with 16 volunteers transcribing over this time. The top transcriber for the Meise *Solanum* specimens was also the top transcriber for the Kew specimens contributing 46.5% or records. Out of the 200 Meise specimen records it was discovered that only 131 were transcribed across all three methods as 69 specimens from the set had not been transcribed by Alembo.

Funded by the Horizon 2020 Framework of the European Union
H2020-INFRADEV-2016-2017
Grant Agreement No 777483

Comparison using Matching Approach

1) Collection date

a) Kew Specimens

Collection date was not transcribed for 12 specimen records across all three of the platforms allowing us to infer that the collection date was not on the label. Of the remaining specimens 81, or 43%, did not match across all three platforms.
A common reason for this mismatch was that the transcription protocol used by Alembo asked transcribers not to interpret the century for a year where it was not written on the specimen label, whereas crowdsourcing volunteers could interpret this information. The number of mismatched records due to missing century information was 24, or 13%.

Barcode	DoeDat Date 1	DoeDat Date 2	DigiVol Date 1	DigiVol Date 2	Alembo Date 1	Alembo Date 2
K000028583	1970-03-15		1970-03-15		70-03-15	
K000190188	1992-08-17		1992-08-17		92-08-17	

Table 2.: Example of Kew specimen records with mismatched century data for Collection date between Doedat, DigiVol and Alembo transcriptions.

Another reason for the mismatch was due to the fact the format of the date entered was not the same across all platforms. For example, days and months were not always transcribed as two digits (e.g. 1971-10-1 or 1971-10-01). 3% of mismatches were due to this reason.

Barcode	DoeDat Date 1	DoeDat Date 2	DigiVol Date 1	DigiVol Date 2	Alembo Date 1	Alembo Date 2
K000064006	1902-9-11		1902-09-11		1902-09-11	
K000339162	1996-6-10		1996-06-10		1996-06-10	

Table 3. Example of Kew specimen records data with mismatched collection date values between DoeDat, DigiVol and Alembo transcriptions due to difference in date format used in transcription

This leaves 28% specimens with mismatches between the three platforms. We investigated these mismatches in more detail by comparing the results to the actual values on the labels.

- 10 specimens, or 5%, did not match at all.

Funded by the Horizon 2020 Framework of the European Union
H2020-INFRADEV-2016-2017
Grant Agreement No 777483

When these 10 non-matching records were compared to the original image, we could see that Alembo transcribers were correct in six cases, DigiVol volunteer transcribers in two cases and DoeDat volunteer transcribers in one. For the remaining records we determined that in fact both Alembo and DoeDat transcribers were correct; only the month was present on the original label but in the protocol for DoeDat, transcribers were asked to enter the year as 3000 if it was not present on the label, so even though both Alembo and DoeDat were correct, the date did not match. The transcriber for DigiVol left the date entirely blank.

The main error transcribers using the crowdsourcing platforms made compared to those using Alembo was that they often transcribed a received date as a collection date when there was no collection date on the label, or they made the assumption that the specimen was collected in the same year it was received and added that year as the collection year.

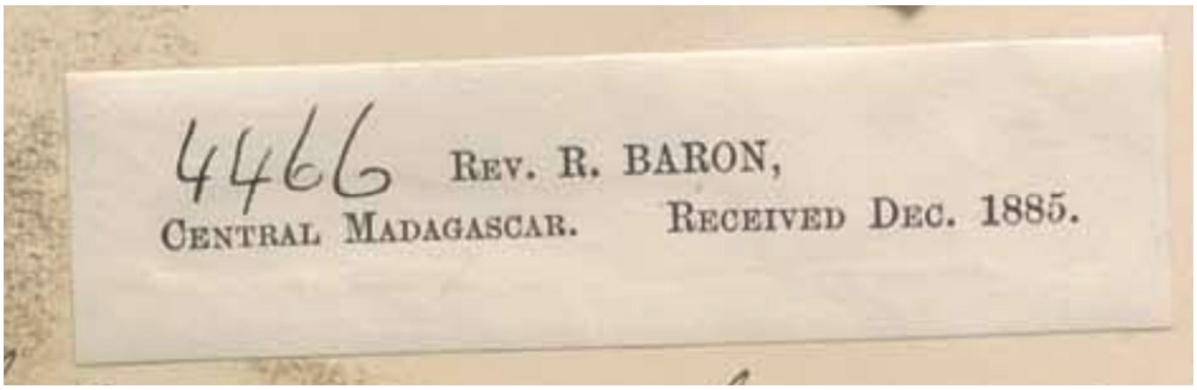


Fig. 3. Example of a label with no collection date but a received data from specimen K000212291 which was wrongly transcribed as the collection date on the crowdsourcing platforms

- 8 specimens, or 4%, matched between DoeDat and DigiVol transcriptions

When these records were checked against the original images, four of the entered records were entered correctly in DoeDat and DigiVol. However, three were actually correctly entered by Alembo as there was no collection date on the label but a received or communicated date which the volunteers on the crowdsourcing platforms had added as a collection date. One of the labels was ambiguous as it had a cultivation date and another which was possibly a collection date. Alembo used the cultivation date whereas the crowdsourcing platform transcribers used the other date which could have been the collection date of the cultivated specimen.

- 19 specimens, or 10%, matched between Alembo and DigiVol transcriptions

Of these 19 records, we confirmed that 17 were entered correctly by both DigiVol volunteers and Alembo transcribers. In 10 instances the DoeDat transcriber recorded other dates as a collection date when there was no collection date obviously present on the label e.g. the communicated date, received date, date from a stamp or a date from another specimen on the same sheet. In two instances the DoeDat transcriber added only a partial date where a full date was available on the label (Fig 4.). Four instances seemed to be a case of mistyping the date. Finally, one DoeDat date should

Funded by the Horizon 2020 Framework of the European Union
H2020-INFRADEV-2016-2017
Grant Agreement No 777483

have matched but the transcriber put the date in the end date rather than the start date fields, although this could be easily rectified in a clean-up process.

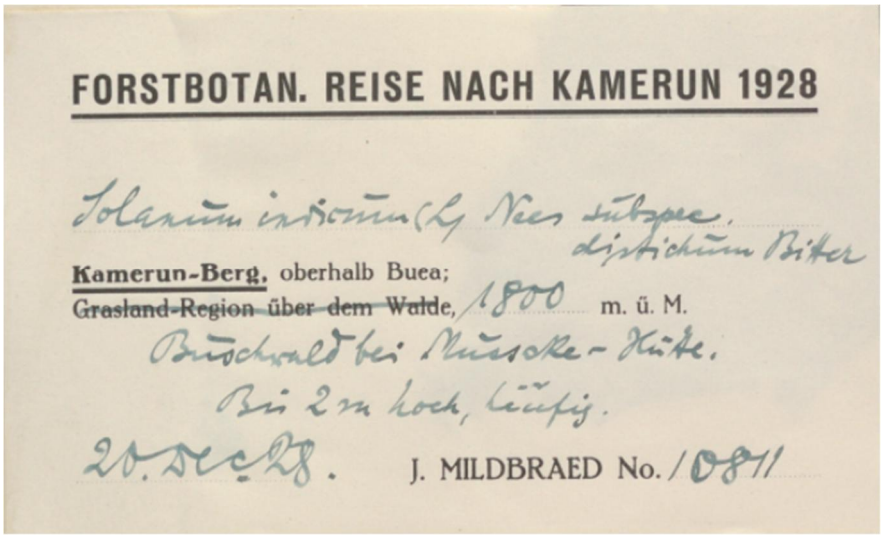


Fig 4. Example of specimen label from Kew specimen [K000028624](#) where only the printed collection year was transcribed by the DoeDat transcriber but a full handwritten date was available further down the label

However, we noted that in two instances the DoeDat transcriber transcribed the correct date, in one case picking up day information from a handwritten label where the month and year only was written on a typed label. This was missed by the other two transcribers. In the second instance, a *fide* date was left blank by the DoeDat transcriber but transcribed as a collector date by both DigiVol volunteers and Alembo staff.

- 15 specimens or 8% matched between DoeDat and Alembo.

All records by DigiVol were found to be incorrect with either a wrong date entered or dates on the specimen label missed. In seven instances a date was transcribed that was not the collector date. This included communicated date, received by date (Fig. 5), determiner date or collector number (Fig. 6.).

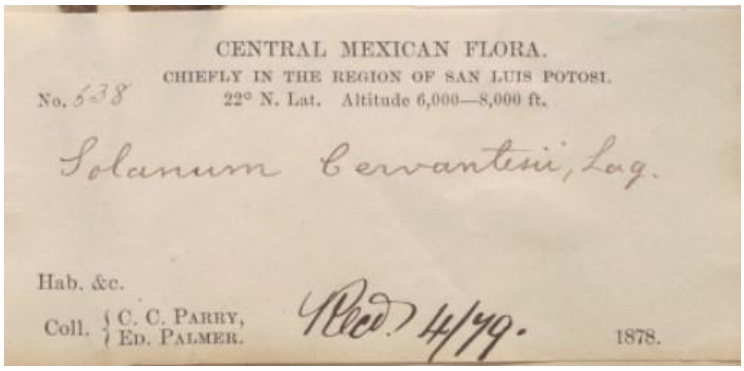


Fig.5. An example of a label where the Recd date was wrongly transcribed as a collector date [K000063939](#)

Funded by the Horizon 2020 Framework of the European Union
H2020-INFRADEV-2016-2017
Grant Agreement No 777483

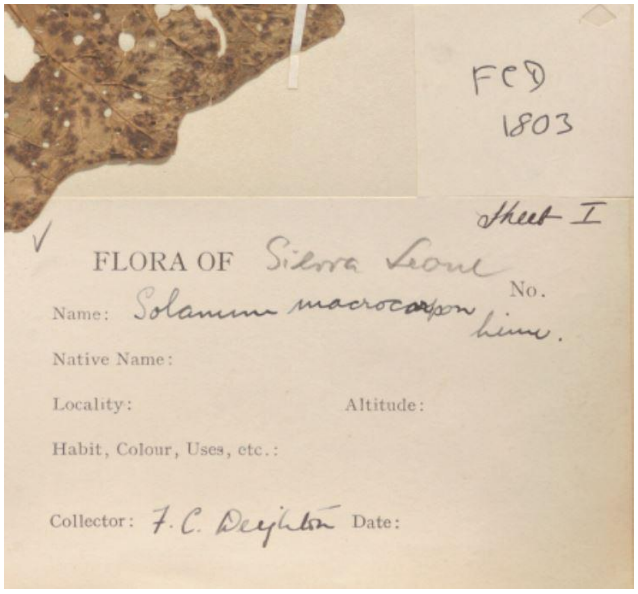


Fig. 6. Example of a label where the collector number was wrongly transcribed as a collector date. K000212445

In three instances the transcriber had failed to spot a collection date and in three instances only the year was transcribed although month information was also giving on the label. One instance seemed to be a typo. In the final instance, the date was transcribed as it appeared on the label, however for that particular collector rather than day, month year the labels are written month, day, year. The transcribers for DoeDat and Alembo must have learnt this but the transcriber for DigiVol was not aware of this difference.

b) Meise Specimens

Collection date was not transcribed for 12 specimen records across all three of the methods, allowing us to infer that the collection date was not on the label. Of the remaining specimens, 81, or 43%, did not match across all three methods.
The number of mismatched records due to missing century information was 16 or 14%.
There were also differences due to how date ranges were transcribed, however these differences only accounted for 4 records, or 3%.

The remaining mismatches represented 28 records or 24%, of all records

- 5 records, or 4%, did not match at all

As expected, some of these were particularly hard labels to transcribe with the dates written in very difficult handwriting.

Funded by the Horizon 2020 Framework of the European Union
H2020-INFRADEV-2016-2017
Grant Agreement No 777483

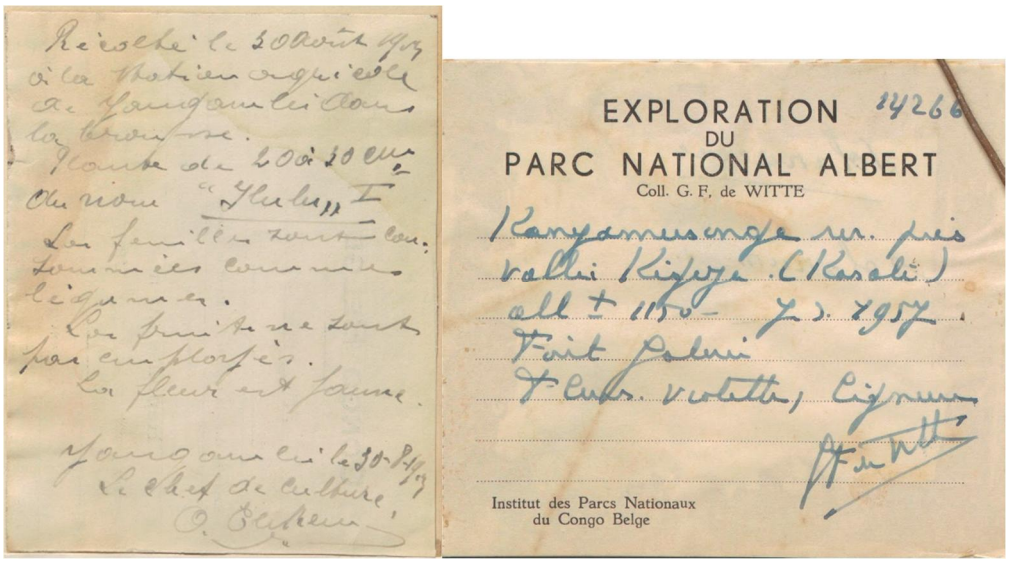


Fig 7. Examples of hard to read specimen labels which were transcribed differently by transcribers of DoeDat, Digivol and Alembo.

- 5 records, or 4%, matched between DoeDat and DigiVol transcriptions only

In two out of the five records all transcribers were actually correct however the collection date range was expressed differently by Alembo due to differences in the protocol previously described. For one specimen record all transcribers had deciphered the year correctly but the Alembo transcriber added a month range from a very difficult to read handwritten label which is difficult to confirm if correct. In the remaining two cases the Alembo transcriber did not detect the date on the label.

- 6 records, or 5%, matched between Alembo and DigiVol transcriptions only

In all these cases the Alembo and Digivol transcribers were determined to be correct. The DoeDat transcriber added the accessioned date of the specimen for two records, missed the collection date for one record, misinterpreted roman numerals for one record and transcribed the month incorrectly for the final record on a hard to read label.

- 12 records, or 10%, matched between DoeDat and Alembo transcriptions only

The DigiVol transcriber was incorrect for these 12 records, the main error again was misinterpreting accession dates or other printed dates on the label for collection dates.

Funded by the Horizon 2020 Framework of the European Union
H2020-INFRADEV-2016-2017
Grant Agreement No 777483

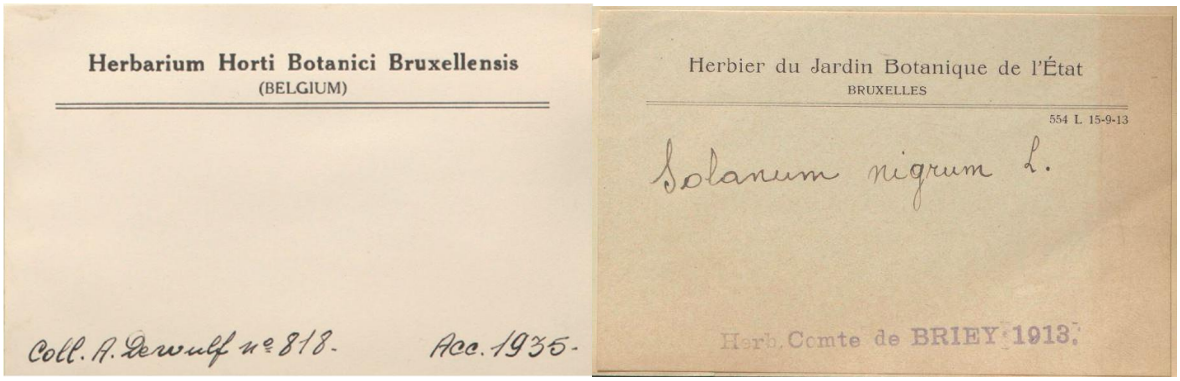


Fig. 8.: Examples of specimen labels where the DigiVol transcribed misinterpreted accession dates for the collector date.

2. Collector

a) Kew Specimens

The number of records which did not match for collector across all transcriptions methods after the matching conditions were applied was 103, or 51.5%.

Barcode	DoeDat Collector	DigiVol Collector	Alembo Collector
K000028556	[Unreadable]	Maitland, T.D.	[Mamailund, H.]
K000028561	F. W. H. Migeod	Migeod, F.W.H.	Migeod, J.W.A.
K000028563	C. D. Adams	Boughey, A.S.	Adams, G.C.
K000028567	Hutch. Metcalfe	[Hutch.] [Metcalfe]	[Mekalde, M.]
K000028583	P.J. Baver	Bauer, P.J.	Bauer, P.I.
K000028601	J. Hutchinson and C.R. Metcalfe		Hutchinson, J. Metcalfe, C.R.
K000028615	J. Hutchinson and C. R. Metcalfe	Hutchinson J. Metcalfe C.R.	Hutchinson, J. Metcalfe, C.R.
K000028624	Mildbraed J.	Mildbraed, J.	Mildbread, J.
K000028637	Maitland		Maitland
K000028663	Stephen D. Manning	Manning, S.D.	Manning, S.D.

Table 4: Example of Kew specimen records with mismatched Collector values between DoeDat, DigiVol and Alembo transcriptions after automated matching conditions were applied.

Funded by the Horizon 2020 Framework of the European Union
H2020-INFRADEV-2016-2017
Grant Agreement No 777483

Barcode	DoeDat Collector	DigiVol Collector	Alembo Collector
K000028557	P. Bamps	Bamps, P.	Bamps, P.
K000028559	J. Olorunfemi	Olorunfemi J.	Olorunfemi, J.
K000028565	D. Dunlap	Dunlap, Dr.	Dunlap, D.
K000028580	J. N. Ngomba 2	Ngomba J.N.	Ngomba, J.N.
K000028581	F. W. H. Migeod	Migeod, F.W.H.	Migeod, F.W.H.
K000028585	Binuyo & Daramola	Binuyo Daramola	Binuyo Daramola
K000028587	T.D. Maitland	Maitland, T.D.	Maitland, T.D.
K000028588	Maitland T.D.	Maitland T. D.	Maitland, T.D.
K000028625	TDM	T.D.M	TDM
K000028626	Dx Dunlap	Dunlap, Dr.	Dunlap

Table 5.: Example of Kew specimen records with mismatched collector values between DoeDat, DigiVol and Alembo that have been accepted after automatic matching criteria have been applied.

Breaking down the remaining mismatches is not straightforward. This is because fuzzy matching allows for relations to be non-transitive (e.g. if a collector name on a DoeDat transcription matches the collector name transcribed by Alembo transcribers, and the name transcribed by Alembo transcribers matches a DigiVol transcription, this does not necessarily mean that the DoeDat transcription will match the DigiVol transcriber's).

The remaining mismatched records were compared to the specimen labels in the images to determine if the collector(s) had been transcribed correctly. The collector(s)'s name(s) transcribed was considered correct if the collector entered was written as it was found on the label or in the format asked for in the protocol. Any issues with using wrong punctuation were ignored under the assumption it could be adjusted to conform to the correct standard during the data-cleaning stage. The records were also checked for the presence of additional collectors up to a maximum of four collectors, the maximum allowed in the protocols. Records with single transcription errors in the name, e.g. one wrong letter in the surname or one wrong initial, were also noted under the assumption that a single error would be easier to match and resolve to the correct collector name format in a later data clean-up stage. Any record with an incorrect collector, i.e. a collector with more than one error in the name or a missing first collector, was considered an incorrect transcription.

Funded by the Horizon 2020 Framework of the European Union
H2020-INFRADEV-2016-2017
Grant Agreement No 777483

	Percentage of Specimen Records		
	DoeDat	DigiVol	Alembo
Correct Collector(s) value	60	72	72
Additional Collector(s) missing	17	2	4
Single error in Collector(s) Field	7	10	11
Incorrect Collector(s)	17	16	13

Table 6.: Percentage of specimen records in which collector values were recorded correctly or otherwise, on each Platform for the mismatched specimens.

When comparing values transcribed to the values on the images of specimen labels, we could see that the non-matching values were often not incorrect but did not match due to differences in the protocols between the different platforms and the various ways to enter a collector(s)'s name. For these 103 mismatched records, DoeDat collector values were correct for 60% of the records and Digvol and Alembo values were correct for 72% of the records. DoeDat had a higher percentage of additional collectors missing when compared to the other platforms. Again, this was down to protocol differences. DoeDat gave a dropdown list for the principal collector only and when selected, the additional collectors were then not transcribed by the volunteer transcriber.

The number of mismatched records that were considered incorrect and could not be matched to collector(s) values on the other platforms were 17% for DoeDat transcriptions, 16% for DigiVol and 13% for Alembo. This equates to only 9% of the total 200 records for DoeDat and DigiVol, and 7% for Alembo.

b) Meise Specimens

The number of specimens that did not match for collector across all transcription methods after the fuzzy matching conditions was applied, was 51, or 39%.

Funded by the Horizon 2020 Framework of the European Union
H2020-INFRADEV-2016-2017
Grant Agreement No 777483

Barcode	DoeDat Collector	DigiVol Collector	Alembo Collector
BR0000009300460	M.Pignal	Pignal, M. Pibot, A. Mas, C.	M. Pignal
BR0000013872946	L. Pauwels		L. Pauwels
BR0000013874742	Schimper	de Bunge Al. Cosson E.	E. Cosson
BR0000013875541	Breyne H.	Nlandu	Breyne H.
BR0000013877774	P.A.M. De Graer	De Graer, P.A.M.	P.A.M. De Graer, O.P.
BR0000013877859	G.F. de Witte	de Witte, G.F.	G.F. de Witte
BR0000014574788	S.C.	Dewèvre, A.	Alfr. Dewevre
BR0000014575082	A. Sapin	Sapin, Adolphe	A. Sapin
BR0000014575419	R.B.Drummond	Drummond, R.B. Hemslay, J.H.	R.B. Drummond and J.H. Hemsley
BR0000014575549	Manuel Fidalgo de Carvalho	Carvalho, M.F.	Manuel Fidalgo de Carvalho

Table 7.: Example of Meise specimen records with mismatched Collector values between DoeDat, DigiVol and Alembo transcriptions after automated matching conditions were applied.

Funded by the Horizon 2020 Framework of the European Union
H2020-INFRADEV-2016-2017
Grant Agreement No 777483

Barcode	DoeDat Collector	DigiVol Collector	Alembo Collector
BR0000005148042	A.R. Torre & M.F. Correira	Torre, A.R. Correia, M.F.	A.R. Torre & M.F. Correia
BR0000005563746	P.C.M. Jansen	Jansen, P.C.M.	P.C.M. Jansen
BR0000008312716	J.Vali?re	Valiere, J.	J. Valière
BR0000009472792	Luke Q, Bytebier B, Butynski T, Ehart C, Perkins A, Kimaro G	Luke Q; Bytebier B; Butynski T; Ehart C; Perkins A; Kimaro G	Luke Q., Bytebier B., Butynski T., Ehart C., Perkins A., Kimaro G.
BR0000009691537	M.G. Bashonga	Bashonga, M.G.	M.G. Bashonga
BR0000013872007	P.M. Daniel	Daniel, P.M.	Daniel P.M.
BR0000013872311	Ph. Gerard	Gerard, P.	Ph. Gerard
BR0000013872526	H. Callens S.J.	Callens, H.	H. Callens s.j.
BR0000013872571	I. Friis, V. Alstrup, A. Michelsen	Friis, I. Alstrup, V. Michelsen, A.	I. Friis, V. Alstrup & A. Michelsen
BR0000013872601	J.B. Gillett	Gillett, J.B.	J.B. Gillett

Table 8.: Example of Meise specimen records with mismatched collector values between DoeDat, DigiVol and Alembo that have been accepted after automatic matching criteria have been applied.

The remaining mismatched records were compared to the specimen labels in the images to determine if the collector(s) had been transcribed correctly. The collector(s)'s name transcribed was considered correct if the collector entered was written as it was found on the label or in the format asked for in the protocol. Any issues with using wrong punctuation was ignored under the assumption it could be adjusted to conform to the correct standard during the data-cleaning stage. The records were also checked for the presence of additional collectors up to a maximum of four collectors, the maximum allowed in the protocols. Records with single transcription errors in the name e.g. one wrong letter in the surname or one wrong initial, were also noted under the assumption that a single error would be easier to match and resolve to the correct collector name format in a later data clean-up stage. Any record with an incorrect collector, i.e. a collector with more than one error in the name or a missing first collector, was considered an incorrect transcription.

Funded by the Horizon 2020 Framework of the European Union
H2020-INFRADEV-2016-2017
Grant Agreement No 777483

	Percentage of Specimen Records		
	DoeDat	DigiVol	Alembo
Correct Collector(s) value	63	69	82
Additional Collector(s) missing	14	2	10
Single error in Collector(s) Field	8	6	4
Incorrect Collector(s)	16	24	4

Table 9. Percentage of specimen records in which collector values were recorded correctly or otherwise, on each platform for the non-matching specimens after fuzzy matching was applied.

When comparing values transcribed to the values on the images of specimen labels, we could see that the non-matching values were often not incorrect but did not match due to differences in the protocols between the different platforms and the various ways to enter a collector(s)'s name. For these 51 mismatched records, DoeDat collector values were correct for 63% of the records, DigiVol 69%, and Alembo values were correct for 82% of the records.
The number of mismatched records that were considered incorrect and could not be easily matched to collector(s) values were 8% for DoeDat transcriptions, 6% for DigiVol and 4% for Alembo. This equates to only 6% of the total 131 records for DoeDat, and 9% for DigiVol, and 2% for Alembo.

3. Collector Number

a) Kew Specimens

75 specimens (38%) did not match between platforms and 125 specimens matched exactly. The mismatches were investigated in more detail by comparing transcribed values to the actual values on the labels. As illustrated in Table 10., some of the mismatches were due to inconsistencies in the inclusion or omission of collection number prefixes between the transcription protocols. Due to these differences in protocol a value was considered correct if the prefix was transcribed correctly or if it was omitted and the rest of number transcribed correctly.

Barcode	DoeDat Collector No.	DigiVol Collector No.	Alembo Collector No.

Funded by the Horizon 2020 Framework of the European Union
H2020-INFRADEV-2016-2017
Grant Agreement No 777483

K000028559	30589		30589
K000028563	6768	GC 6768	6768
K000028567	24	124	124
K000028580	S.N.		2
K000028585	35550	FHJ35550	35550
K000028601	138		138
K000028615	16		16
K000028624	81	10811	10811
K000028626	31		31
K000028640	S.N.	SCA 673B	1

Table 10: Example of Kew specimen records with mismatched Collector numbers between DoeDat, DigiVol and Alembo transcriptions.

Of these non-matching records only 9% or 17 specimens did not match across any of the transcription platforms. Alembo transcriptions were correct in 13 cases, DigiVol in one transcription and DoeDat in six. Three of the records did not match between Alembo values and DoeDat because of the inclusion or omission of prefixes but were considered correct values.

- 31 (16%) specimen records matched between DoeDat and Alembo transcriptions

Of these, 29 specimens were found to be transcribed correctly by DoeDat and Alembo transcribers. In two cases it was not clear without further investigation and research if the collector number was correct, because the handwritten numbers on the label are unclear. DigiVol transcriptions were also found to be correct for seven specimens but did not match DoeDat and Alembo due to null values not matching to s.n. values or due to the inclusion or omission of the collector number prefixes. The main error DigiVol transcribers made was failing to add a collector number when it was present on the sheet. In these instances, the transcriber was perhaps taking a cautious approach and not adding any values for collection number unless there was a clear indicator on the specimen label that the number present was a collector number. E.g. “coll no.” was written on the sheet.

- 18 (9.0%) specimen records matched between DigiVol and Alembo transcriptions

We found these 18 specimen records were correctly entered by DigiVol and Alembo transcribers. DoeDat transcribers entered 5 specimen records correctly. However, either additional information was written in the collector number field e.g. (sheet i) or the prefix was omitted or transcribed, so the values did not match.

Funded by the Horizon 2020 Framework of the European Union
H2020-INFRADEV-2016-2017
Grant Agreement No 777483

- 9 (5%) specimen records matched between DigiVol and DoeDat transcriptions

We found 7 specimens records were transcribed correctly by DigiVol and DoeDat transcribers, one transcribed incorrectly and one specimen which was ambiguous. In the case of the incorrect record, DigiVol and DoeDat transcribers added two collector numbers, the correct number and the number of a different specimen on the same sheet. In the ambiguous case, there were two specimen labels on the sheet but only one plant and it was not clear which was the correct specimen label. However, the additional label seems to be of a specimen in spirit (Fig 9.)

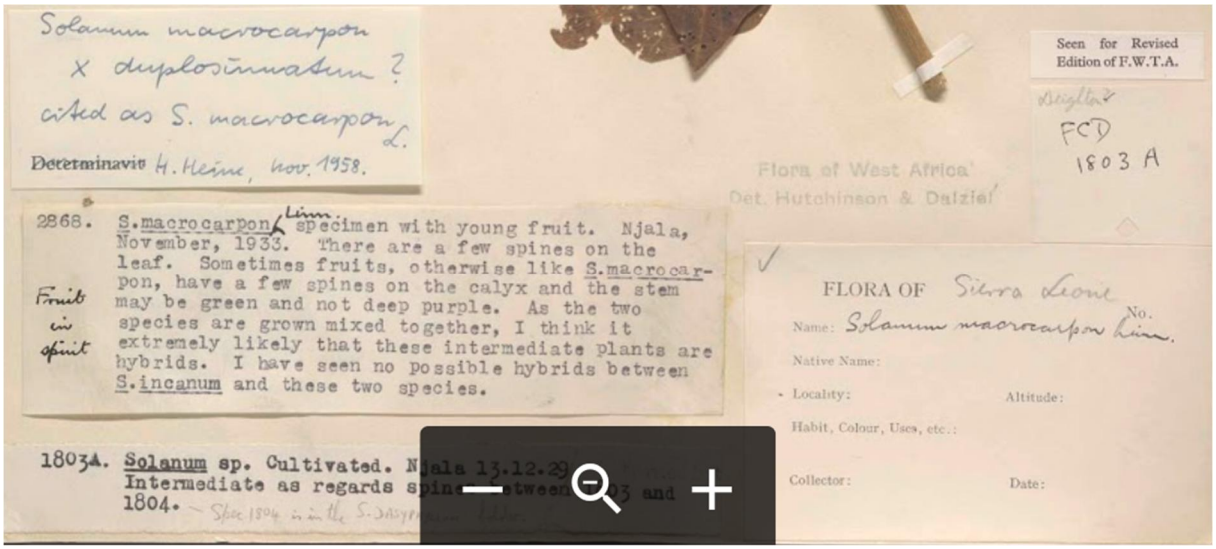


Fig. 9. Labels from Kew specimen [K000212446] of two specimen labels on sheet with only one plant specimen.

Overall, 10 non-matching specimen records were found to match when punctuation was removed (e.g. commas and dashes within a number) and case was ignored, thus “S.N.” and “s.n.” were considered to match.

b) Meise Specimens

31 (24%) specimen records did not match across all platforms

Funded by the Horizon 2020 Framework of the European Union
H2020-INFRADEV-2016-2017
Grant Agreement No 777483

Barcode	DoeDat Collector No.	DigiVol Collector No.	Alembo Collector No.
BR0000008312716	S.N.	831 271	S.N.
BR0000009300460	S.N.	1234	1234
BR0000013872946	5303	5353	5303
BR0000013874735	8042 Bot.	8042	8042Bot
BR0000013874742	I181	181	I181
BR0000013875886	1039	1039	1030
BR0000013877859	10976	10076	10976
BR0000014574788	S.N.	344	344
BR0000014575082	S.N.		S.N.
BR0000014576966	S.N. 2e exempl.	s.n.	S.N.

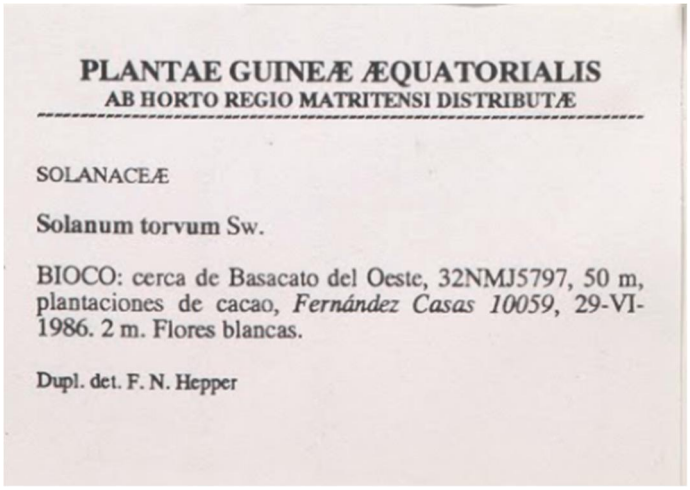
Table 11: Example of Meise specimen records with mismatched Collector number values between DoeDat, DigiVol and Alembo transcriptions.

Of these non-matching records:

- Only 2 (2%) specimen records did not match across any of the platforms.

One of these specimen sheets had a plant part covering the collector number so it could not be determined from the image. The collector number for BR0000014798580 was transcribed correctly by Alembo but missed by the DigiVol and DoeDat transcribers who entered S.N. and Null This may have been because the label although while clearly printed presented the collector information between locality information and plant description. For this same specimen the same DigiVol and DoeDat transcriber wrongly identified the determiner of a duplicate as the collector.

Funded by the Horizon 2020 Framework of the European Union
H2020-INFRADEV-2016-2017
Grant Agreement No 777483


PLANTAE GUINEÆ ÆQUATORIALIS
AB HORTO REGIO MATRITENSI DISTRIBUTÆ

SOLANACEÆ

Solanum torvum Sw.

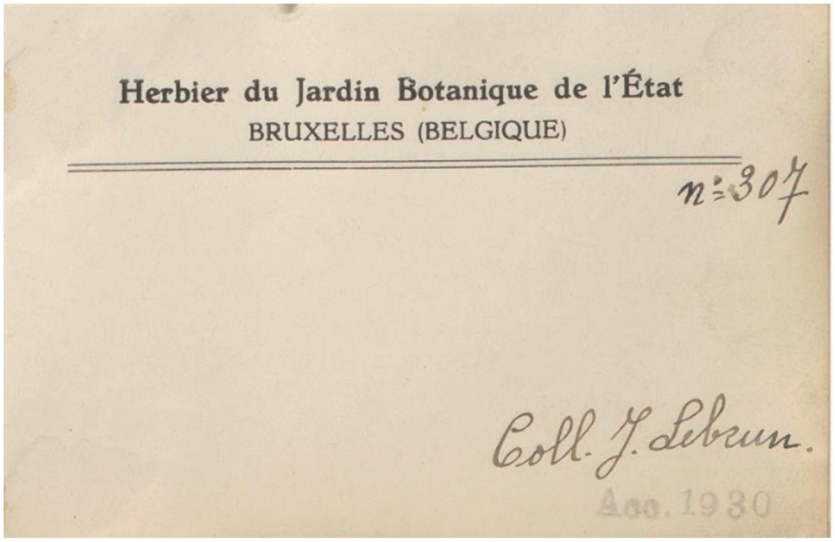
BIOCO: cerca de Basacato del Oeste, 32NMJ5797, 50 m, plantaciones de cacao, *Fernández Casas 10059*, 29-VI-1986. 2 m. Flores blancas.

Dupl. det. F. N. Hepper

Fig.10. Example of a specimen label where the collector number was missed and the determiner of a duplicate specimen was added as the collector by DigiVol and DoeDat transcribers.

- 22 (17%) specimens records matched between DoeDat and Alembo transcriptions.

DoeDat and Alembo transcribers were found to be correct in all but one case where the ink of the printed collection number was unclear but had been handwritten on a determination label. The DigiVol transcriber was also found to be correct for seven specimens but did not match DoeDat and Alembo due to null values not matching to s.n. values or due to the inclusion or omission of the collector number prefixes. The main mistranscriptions by DigiVol volunteers was due to missing the collector number on the sheet (5 cases) and misinterpreting other numbers on the sheet such as an accession date for the collector number (7 cases).


Herbier du Jardin Botanique de l'État
BRUXELLES (BELGIQUE)

n:307

Coll. J. Lebrun.

Acc. 1930

Fig. 11.: Example of specimen label [BR0000014801372](BR0000014801372) where the Accession stamp was misinterpreted as the collection year by the DigiVol Volunteers

Funded by the Horizon 2020 Framework of the European Union
H2020-INFRADEV-2016-2017
Grant Agreement No 777483

- 4 (3%) specimens records matched between DigiVol and Alembo transcriptions.

For all four of these specimens DigiVol and Alembo were found to be correct. The DoeDat transcriber had missed the collector number for three of the specimens but had identified correctly that the fourth specimen did not have a collector number but entered the collector number as S.N. 2e exempl. , as "(2e exampl.)" was written on the label after the collector name, so the transcriptions did not match.

- 3 (2%) specimens records matched between DigiVol and DoeDat transcriptions.

For these three specimens DoeDat and DigiVol transcribers were found to be correct. However, in one case Alembo also matched but added the suffix "bot" to the collector number. In the two other specimens where Alembo transcribers did not match the collector number was difficult to translate but could be confirmed by other labels on the specimen.

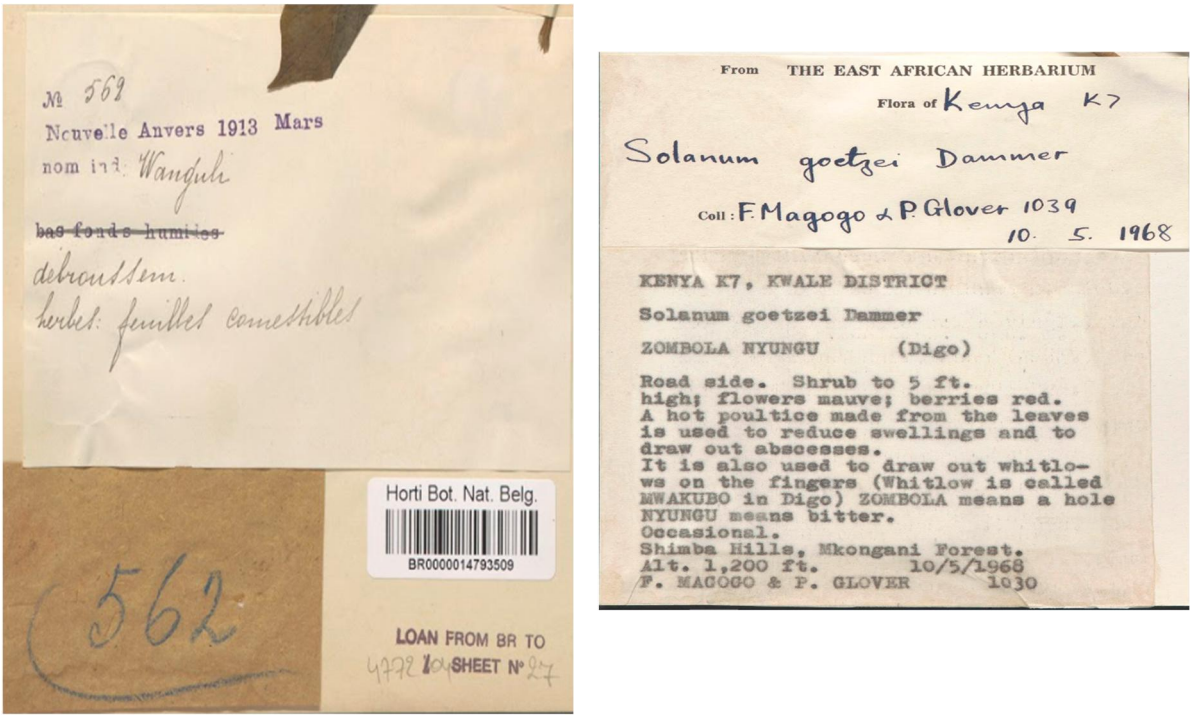


Fig. 12: Examples of the two labels transcribed incorrectly by Alembo for collector number

10 non-matching specimens were found to match when punctuation was removed (i.e. commas, spaces and dashes within a number) and case was ignored, e.g. 1057a and 1057A.

Funded by the Horizon 2020 Framework of the European Union
H2020-INFRADEV-2016-2017
Grant Agreement No 777483

Comparison using Gold Standard Approach

1) Collection date

Collector Date values categorised for each dataset					
	Alembo	HerbCat	DoeDat	DigiVol	Researcher
Match	157	162	169	172	173
Incorrect	2	18	19	16	18 (5)
Typo/Misinterpretation	4	7	7	4	2
Additional characters	0	1	1	0	3
Missing characters	32	12	3	5	3
Format	0	0	0	0	0
Incomplete	5	0	1	3	1
Total	200	200	200	200	200

Table 12: Number of Collector Date values in each category for all datasets

Accuracy of Collection Date:

The percentage of collection dates that were recorded with complete accuracy were highest in the dataset created by the independent researcher at 87% and lowest in the dataset provided by Alembo. However, the Alembo dataset included the least number of values that were totally incorrect at only 1% of the records, whereas incorrect dates from the DoeDat dataset came out the highest at 10%. Incidence of additional characters occurring in the date fields was low across all datasets. There was a higher incidence of missing characters within the Alembo dataset, however this was mostly regarding the year, where protocol instructed a recording of partial year information where that is how it appeared on the label.

Completeness of Collection Date:

Values classed as complete in most cases across all datasets and completeness scores did not differ much between datasets (98% -100%), with the lowest score belonging to the Alembo dataset and the highest to the HerbCat dataset. Perhaps, however, given that five of the ten dates entered by the researcher that would have been classed as incorrect as they were completely different from the information on the sheet were probably from another source, it may be that the researcher's dataset would have attained a higher completeness score if another verification step was part of the completeness test.

Funded by the Horizon 2020 Framework of the European Union
H2020-INFRADEV-2016-2017
Grant Agreement No 777483

Validity of Collection Date:

Five of the ten collection dates categorised as incorrect in the researcher's data were received dates and were therefore considered invalid information for this field. This was consistent across the other datasets.

Standardisation of Collection Date:

Collection dates were found, somewhat surprisingly, to be valid according to international standard ISO 8601 *Data elements and interchange formats – Information interchange – Representation of dates and times* in all six datasets. This is unusual because aside from transcription errors resulting in dates from the future, e.g. 9183 instead of 1983, labels with erroneous dates, e.g. the 31st day of a month with only 30 days, usually occur from time to time, but none was present in this dataset.

Similarity of Collection Date across datasets:

Of all 200 records, 104 (52%) had a collection date value which matched the gold standard in at least one dataset. Of the remaining 96 records, 17 did not match the gold standard in either three or four of the five datasets. The most common reason for this was invalid collection dates as they were received and communicated dates (as in figure 3). Some were difficult to read and one which was incorrectly transcribed in four of the five datasets comprised the month of November written in shorthand as "9br", transcribed as September.

2) Collector name

Collector name values categorised for each dataset					
	Alembo	HerbCat	DoeDat	DigiVol	Researcher
Match	123	141	115	123	123
Incorrect	9	0	14	7	2
Typo/Misinterpretation	23	10	13	10	11
Additional characters	2	10	6	4	26
Missing characters	5	6	6	9	4
Format	29	29	40	43	30
Incomplete	4	4	6	4	4
Uncertainty	4	0	0	0	0
Total	200	200	200	200	200

Table 13: Collector name values categorised for each dataset

Funded by the Horizon 2020 Framework of the European Union
H2020-INFRADEV-2016-2017
Grant Agreement No 777483

Accuracy of collector names:

The percentage of collector names that were recorded with complete accuracy were highest in the dataset exported from Kew's in-house database HerbCat at 71%. Data provided by Alembo, DigiVol and the researcher that completely matched that of the Gold Standard values were all equally accurate at 62%. DoeDat produced the lowest at 58%. None of the values provided in the HerbCat data were completely incorrect, but 2% were incomplete, 3% were missing characters and 5% contained and 10% either contained additional characters or typos. Format issues were relatively high across all datasets (this is likely to be due to different requirements highlighted within the protocols). Data from the independent researcher contained a much higher instance of additional characters than any other dataset. The Alembo dataset also included values marked with uncertainty, this is where uncertain values are entered in square brackets. This dataset also presented the highest number of typos or misinterpretation of characters at 12%.

Completeness of collector names:

All datasets presented very high completeness scores for collector name. Four of the five were 98% complete except DoeDat which was 97% complete.

Validity of collector names:

The dataset presenting the highest rate of invalid collector names was DoeDat, with 14 names that were not that of a collector, the HerbCat dataset had no invalid collector names. All other datasets were somewhere in between.

Standardisation of collector names:

The percentage of records matching at least one collector in HUH, versus an error when there was no match, was highest in the independent researcher's dataset at 85% and lowest in the Alembo dataset at 32%. The Gold Standard dataset, among the other four datasets which all had matches in HUH for above 71% of first collector names, had only the second highest percentage of matches.

Number of first collectors matching at least one standard entry in HUH for each dataset						
Dataset	Alembo	Researcher	HerbCat	DigiVol	DoeDat	Gold Standard
Number with at least one match in HUH	145	170	166	141	63	168

Table 14: Number of first collectors matching at least one standard entry in HUH for each dataset

Similarity of collector name across datasets:

Of all 200 records, 79 (39%) had a collector name value which matched the gold standard in at least one dataset. Of the remaining records, 35 did not match the gold standard in three or more of the five datasets.

Funded by the Horizon 2020 Framework of the European Union
H2020-INFRADEV-2016-2017
Grant Agreement No 777483

Though standardisation depends heavily on accuracy, the difference in results between accuracy and standardisation of collector names can partly be explained by the fact that an accurate transcription from a label does not necessarily correspond with a collector's standard name or abbreviation. This explains why, for example the gold standard data matched fewer collectors than the independent researcher's. This is also indicative of the fitness for use case. I.e., for the researcher, it was more important to record a standard name than to have an accurate transcription. For example, in one case, the collector R. Baron was recorded on the label as "Barron" but was recorded in the researcher's dataset under the standard name. In the case of an independent researcher, who is familiar with collectors pertaining to his or her specimens, this interpretation can most likely be assumed to be correct. In the opposite case, volunteers and paid transcribers may not have the confidence, the expertise, or permission to make these assumptions.

One limitation of this test is that collectors were not checked in HUH for correctness. I.e. a positive match does not necessarily include the individual collector in question on the label, nor did we test for this. Also, different values for one record in two datasets could both bring up their own matching results. Though it would be possible to ascertain the correct HUH identifier for each collector name in this dataset, this would be very time consuming and would not be a realistic prospect for large datasets. Also, searches in HUH require specific search formats; the person's name must be entered "either as *lastname, firstname*, e.g. *jones, david* or by the standard abbreviation, e.g. *d r jones".* Datasets which contained names which were already atomised into surname, firstname or initials, as well as names entered in a consistent format were easy to parse and query in HUH, whereas names entered in a variety of formats, such as in DoeDat, were not as easy to parse and therefore scored lower on the standardisation test, regardless of accuracy or completeness.

3) Collector Number

Collector Number values categorised for each dataset					
	Alembo	HerbCat	DoeDat	DigiVol	Researcher
Match	174	177	150	149	179
Incorrect	3	3	15	8	4
Typo/Misinterpretation	5	3	8	4	4
Additional characters	7	1	10	5	9
Missing characters	6	16	5	7	0
Format	2	0	4	3	4
Incomplete	3	0	8	24	0
Total	200	200	200	200	200

Table 15: Collector Number values Categorised for each dataset

Funded by the Horizon 2020 Framework of the European Union
H2020-INFRADEV-2016-2017
Grant Agreement No 777483

Accuracy of collector numbers:

All datasets presented levels of complete accuracy of 75% or higher. HerbCat produced a much higher incidence of collector numbers that were missing characters (8%). DoeDat produced a larger number of records that had incorrect collector numbers or numbers with superfluous characters.

Completeness of collector numbers:

The Researcher's and the HerbCat datasets performed well on completeness with no empty values. There was a much higher instance of incomplete values, where the field was left blank despite a number present on the label, in the DigiVol dataset.

Validity of collector numbers:

DoeDat presented the highest percentage (8%) of incorrect values while the Alembo and HerbCat datasets both had the lowest with less than 2%.

Standardisation of collector numbers:

We did not test for standardisation on the collector number field as there is no standard on collector number notation.

Similarity of collector numbers across datasets:

Collector numbers which were well transcribed across all datasets accounted for 58% of values in all records. The remaining values presenting differences across datasets had prefixes and suffixes and other characters, such as dashes and punctuation, in common. Multiple numbers on sheets (such as in figure 9) also caused variation across datasets with transcribers either entering the multiple numbers in one field or choosing different ones.

4) Country

Country values categorised for each dataset					
	Alembo	HerbCat	DoeDat	DigiVol	Researcher
Match	166	197	180	190	192
Incorrect	6	1	8	3	3
Historical	2	1	0	2	1
Format	4	1	4	4	4
Incomplete	2	0	0	1	0
Unknown	20	0	8	0	0
Total	200	200	200	200	200

Table 16: Country values Categorised for each dataset

Funded by the Horizon 2020 Framework of the European Union
H2020-INFRADEV-2016-2017
Grant Agreement No 777483

Accuracy of countries:

There was an extremely high percentage of records within the HerbCat dataset that displayed the correct country. Instances of incorrect country values, incomplete values or values in a different format was low across all datasets. Some values that did not match the gold standard could be explained due to old terminology or boundary changes and were categorised as “Historical”. Alembo had the highest instance of “Unknown” values accounted for 10% of the dataset’s values.

Completeness of countries:

Completeness scored very highly for all datasets, ranging from 99% to 100%. There was not much difference between datasets.

Validity of countries:

The values entered in the country field were always countries, and never any other kind of information. Therefore, validity was 100% for all datasets.

Standardisation of countries:

Though not all country names used were currently recognised country names, all could be mapped one to one with present-day countries. Therefore, all could be considered to conform to the *Codes for the representation of names of countries and their subdivisions – Part 1: Country codes*, ISO 3166-1 standard after an enhancement step.

Similarity of countries across datasets:

We found that 80% of records matched across all datasets on country values. The remaining records with values which differed from the gold standard had labels which contained information such as expedition titles (e.g. Niger Expedition) that did not correspond to the country the specimen was collected from as indicated from the rest of the information on the label, or historical names for countries, regions where there had been a boundary change resulting in a different present-day country, as mentioned above, and regions with the same name present in more than one country, causing ambiguity, e.g. Gold Coast.

5) Plant Names

Standardisation of Plant Names:

Dataset	Alembo	HerbCat	DoeDat	DigiVol	Researcher
Number of values corresponding to a unique IPNI identifier	193	182	103	184	193
Number of values with a match in Catalogue of Life’s List Matching Service	199	186	87	189	199

Table 17: Number of Plant Name values corresponding to standard values in IPNI and Catalogue of Life in each dataset

Funded by the Horizon 2020 Framework of the European Union
H2020-INFRADEV-2016-2017
Grant Agreement No 777483

Datasets presenting the highest numbers of plant names which corresponded to a standard value present in IPNI were those of Alembo and the independent researcher, both at 97%. The DigiVol and HerbCat datasets were not far behind at 92% and 91%, respectively. The DoeDat dataset displayed the lowest, with only 52% of names matching a standard value in IPNI. The most common reasons for not matching was duplication of names in IPNI, misspelling of names and transcribers omitting species epithets, so just the genus was given.

Fig. 13 summarises the results of the gold standard approach for each data quality dimension across each dataset. Comparing the scores for each data quality dimension in the gold standard approach method revealed no obvious overall leader among the datasets, given the limitations we describe above.

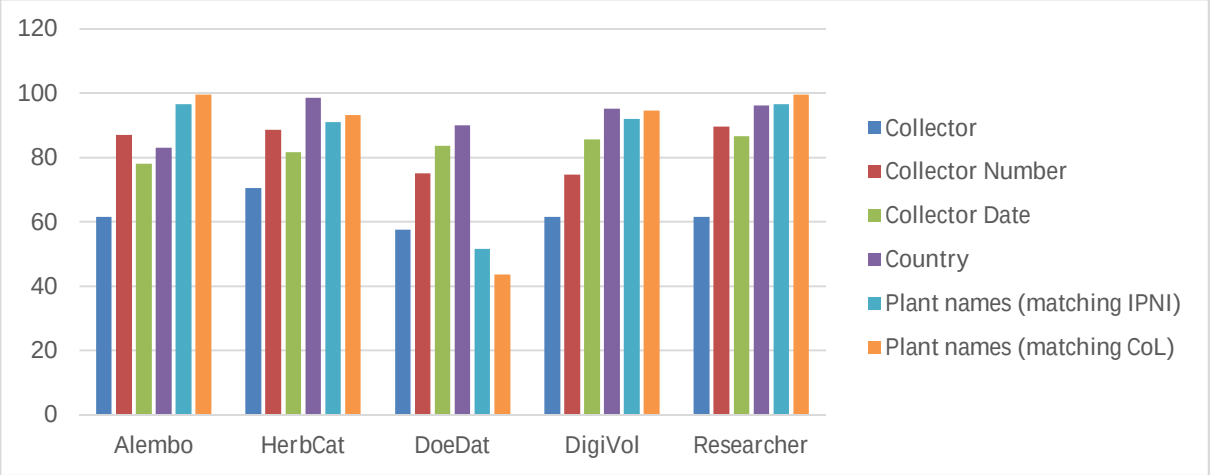


Fig. 13. Percentage of values matching the gold standard for each field by dataset on Kew Solanum records

3.1.3 Discussion Transcription Pilot 1

Both approaches to assessing DQ lead to similar results and similar data issues emerged from both sets of results. Here, we highlight the most common issues in each of the fields we looked at.

Collection Date:

In both DQ assessment approaches, the same data issues emerged for collection date. Data present on labels which caused difficulties for transcribers were: centuries not present on labels, partial dates (such as month only), notation (such as roman numerals and short-hand), difficult-to-read handwriting and other dates which were not transcription dates. Data entry set-up which caused errors and inconsistencies included: Date fields not atomised (e.g. leading to mix-ups of dates and months), lack of validation allowing typos or misuse of the start and end date fields. Differences between protocols lead to inconsistency between data entry methods such as entering centuries absent on label as blank, adding filler numbers or asking people to deduce information instead.

Funded by the Horizon 2020 Framework of the European Union
H2020-INFRADEV-2016-2017
Grant Agreement No 777483

Collector name:

The most common DQ issues affecting collector names transcribed from labels were names which could be interpreted in different ways. Issues caused by data entry set-up included encoding issues and lack of atomised fields for effective parsing. Differences in transcription protocols between methods also lead to inconsistencies in collector name transcription.

Collection number:

The DQ issues shared across all methods of transcription for the collector number field, arose mostly from the different formats in which collector numbers are present on labels. Prefixes and suffixes as well as various punctuation and other symbols were handled differently by different data-entry set-ups as well as by different transcription protocols. Atomisation, or not, of fields for capturing collector numbers also lead to differences from label data. Handwriting, non-collector numbers on labels and lack of validation in data entry forms were also issues, we address in our recommendations below.

Country:

Historical names and boundary changes, misleading expedition or Flora titles and regions from different countries sharing the same name were all issues present on specimen labels.

Plant name:

Because of the number of plant names often present on labels and lack of consistency between protocols, we did not assess most DQ dimensions for plant names. However, we measured standardisation for the names that were entered, as every determination recorded validly from the labels is of value (although we did not test for presence on the sheet in this study). The most common reason for non-matches was multiple records containing the same values which had duplicates in IPNI and therefore could not match a unique ID in IPNI. Misspellings accounted for some name mismatches. The dataset with the lowest number of matches also had the highest number of missing species epithets.

From our findings, we think the greatest enhancements in DQ in future data capture will come from improvements to atomisation, validation and standardisation of data capture fields as well as more granularity, clarity and standardisation in data entry protocols. There are inherent ambiguities and inconsistencies present on specimen label data. We provide recommendations for data entry set-up, protocols and dealing with inherent issues on labels further down.

3.2 Transcription Pilot Two

Multiple European institutions holding botanical collections were approached to provide a sample of their digitally imaged herbarium sheet specimens to upload to multiple crowdsourcing platforms (Dillen *et al.*, 2019). Specimens were chosen to ensure a representative cross-section of herbarium specimen sheets. The chosen sheets were a mixture of different plant families, collectors, collection dates and were collected from different countries. The aim was to provide representative specimens that covered different languages, typed and handwritten labels, different collectors and labels with

Funded by the Horizon 2020 Framework of the European Union
H2020-INFRADEV-2016-2017
Grant Agreement No 777483

rich data compared to labels with very little information. Volunteers would therefore come across the many different types of challenges typically involved with label transcription.

200 images each from:

- P/PC - Muséum national d'Histoire naturelle (Paris - France)
- BR - Plantentuin Meise, Belgium
- B - Botanischer Garten und Botanisches Museum Berlin-Dahlem (Berlin, Germany)
- BM - The Natural History Museum (London, UK)
- K - Royal Botanic Gardens Kew, UK
- E - Royal Botanic Garden Edinburgh, UK
- TU - Tartu Ülikool, Estonia.

These 1400 specimen images were uploaded for transcription to 5 different crowdsourcing platforms. DigiVol (https://digivol.ala.org.au/), DoeDat (https://www.doedat.be/), Die Herbonauten (https://www.herbonauten.de/), Les Herbonautes (http://lesherbonautes.mnhn.fr/) and Notes from Nature (https://www.notesfromnature.org). Only English labels were uploaded to the Notes from Nature platform as specified by the platform administrators, giving a total of 579 available for the test.

Label data already available corresponding to these images were downloaded from The Global Biodiversity Information Facility (https://www.gbif.org/) for all institutes except the Botanical Garden and Botanical Museum, Berlin, where the data were downloaded from JACQ (https://herbarium.univie.ac.at/database/), which is a joint specimen data management system of over 30 European and Asian Herbaria (Rainer and Vitek, 2009). There is no indication in the data to determine if the specimen is completely or partially transcribed and they are likely to be of diverse quality. However, they are data that have been publicly released and available for research use.

Currently two of the transcriptions projects are still ongoing on DoeDat and Notes from Nature and the results from Die Herbonauten returned too late for analysis. Therefore, results are currently restricted to DigiVol and Les Herbonautes.

To assess the quality of the data was to investigate which records matched across the published data (GBIF/JACQ) and the crowdsourced data from DigiVol and Les Herbonautes.

3.2.1 Pilot Two Results

The final number of specimens compared was 1393 as a few images were dropped from the final dataset. In one instance an image did not go through the tiling process in the Les Herbonautes platform so was not available for transcription.

Funded by the Horizon 2020 Framework of the European Union
H2020-INFRADEV-2016-2017
Grant Agreement No 777483

Les Herbonautes mission lasted from 22/06/2018 to 10/10/2018 (100 days). In total 41 people transcribed some of the label data. As is common with crowdsourcing missions, over 50% of the contributions (not specimen records) was from three major users, over 75% by 5 major users and 12 users made 10 or fewer contributions. 21 specimens had already gone through the crowdsourcing platform so these specimens could not be transcribed again. Therefore, the data from the previous mission were used for comparison. In Les Herbonautes users start with simple transcription of one field (country) and are tested in a tutorial before progressing to more challenging fields. There is validation of individual fields by other participants (2 to 3) until consensus is reached.

Fig. 14. Transcription pilot 2 crowdsourcing expedition on Les Herbonautes

The DigiVol expedition lasted from 13/06/2018 to 12/09/2018 (91 days). In total, 30 people transcribed specimen labels. Over 26% were completed by one major user and 79% were completed by five major users with 17 users transcribing fewer than 10 specimens. In DigiVol each individual transcribes one specimen record. The records can then be validated by another user who has been granted validator status by individuals with administrative rights. For the purpose of this test the records did not go through this second validation step and the raw individual record data were used.

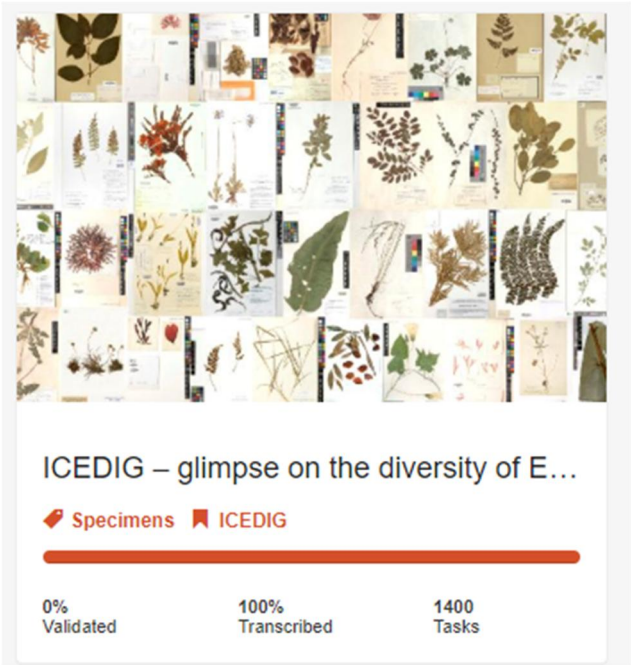


Fig. 15. Transcriptions pilot 2 crowdsourcing expedition on DigiVol.

Funded by the Horizon 2020 Framework of the European Union
H2020-INFRADEV-2016-2017
Grant Agreement No 777483

More information about the comparison of the different platform features can be found in the milestone report MS26 – evaluation of existing transcription systems.

1) Collection date

In total there were 96 specimen records (7%) with no date transcribed by either platform so these were removed from further analysis. Of the remaining records present in GBIF/JACQ, there were 225 records (17%) with a missing date. As we do not know how complete the data records sent to GBIF or JACQ are, we cannot assume that because there is no collection date in the data that this means a collection date is not present on the label. It might be that this date field was not transcribed or the data in this field were not sent to GBIF/JACQ. Therefore, these records were excluded from further analysis leaving 1072 specimen records (77%) for further analysis. Of these remaining specimens, 686 dates were found to match with 386 mismatching.

Barcode	DigiVol Date	GBIF Date	Les Herbonautes Date
B 10 0000390	1917-06-09	1914-06-09	1914-06-09
B 10 0002801	1856	1856-00-00	1856-01-01/1856-12-31
BM000521577	1819	1819-01-01	1819-01-01/1819-12-31
BM000521755	1973-08-06	1973-11-06	1973-08-06
BR0000005849017		1833-01-01	1833-01-01/1833-12-31
BR0000005857975	1874-10	1874-10-01	1874-10-01/1874-10-31
E00001162	1919	1919-01-01	1919-01-01/1919-12-31
E00016420	1917-07-23	1917-06-23	1917-07-23
K000049450	1977-3-10	1977-03-10	1977-03-10
K000232911	1905	1907-03-22	1907-03-22
P00036507	1979-07-30	1978-08-03	
P00067887	1901	1901-06-01	1901-01-01/1901-12-31
TU250942	1970-07-12	1973-07-21	1973-07-21
TU251269		1927-06-26	

Table 18.: Example of specimen records with mismatched collection date transcribed values between DigiVol, GBIF and Les Herbonautes

Funded by the Horizon 2020 Framework of the European Union
H2020-INFRADEV-2016-2017
Grant Agreement No 777483

The next step was to correct some issues due to differences caused by missing digits, so days and months were zero-padded if required (e.g. 2017-1-1 was changed to 2017-01-01). We also fixed partial dates by removing 00 values for day and month. If date ranges could be specified as partial dates, we also enabled this comparison (e.g. 2017-10 vs 2017-10-01/2017-10-31). However, this does not work if a date range spans more than one month. These changes allowed another 17 records to match leaving 369 mismatches.

Breakdown of mismatches:

- 80 records (7%) did not match across all three platforms.

These results were looked at in more detail by comparing the values against the specimen image. Many of these mismatches were due to how collection date ranges were transcribed or represented in the data. Within GBIF only the first collection date of a range is specified as a collector date and where there is no day or month the value defaults to 01 giving a false indication of precision in both cases. For example, in Fig.15. the label shows that the specimen has been collected by Schimper between 1863-1866. This label was transcribed by DigiVol as 1863/1869, by Les Herbonautes as 1863-01-01/1868-12-31 and within the GBIF DwC-A as 1863-01-01. On the Edinburgh Herbarium Specimen catalogue the date is represented as 1863 only.

Fig. 16. Example of a specimen label (E00513937) which was transcribed in DigiVol as 1863/1869, by Les Herbonautes as 1863-01-01/1868-12-31 and appeared within the GBIF DwC-A as 1863-01-01

When there is only a year on the label, Les Herbonautes transcribers will still translate this to a date range for example 1854 will be transcribed as 1854-01-01/1854-12-31 whereas in DigiVol the date would be 1854. Similarly, if there is a collection month and year but no day e.g. August 1883, Les Herbonautes transcribers will translate this to a date range e.g. 1883-08-01/1883-08-31 whereas DigiVol transcribers will transcribe as 1883-08 with no day information. Within the GBIF DwC-A the date will be shown as 1883-08-01

Funded by the Horizon 2020 Framework of the European Union
H2020-INFRADEV-2016-2017
Grant Agreement No 777483

	Number of Specimen Records		
	DigiVol	GBIF	Les Herbonautes
Correct Collection Date Value	30	25	51
Ambiguous Collector Date Values	7	10	10
Incorrect Collector Date Values	43	7	19
Collection Date value not on label	0	4	0
Dates with False precision	0	34	0

Table 19: Breakdown of mismatched collection dates across DigiVol, GBIF and Les Herbonautes.

The images were compared with the collection date values and the data categorised into Correct Collection date values, Ambiguous collection date values, collection date values not on label and dates with false precision. Ambiguous collection dates were given when the date transcribed could have been correct due to extremely difficult to read handwriting. Within the GBIF data there were four records with collection dates that were not present on the label these were not deemed incorrect as the data might have been collated from some other data source within the institution. Thirty-four records in GBIF had either day or day and month displayed as 01 where there was no day or month on the label or only the first date of a range was shown.

Common incorrect collection date values seen were taking information from the wrong specimen on the sheet, misinterpreting the collection date as the bequeathed or communicated or accessioned date or taking the information from the wrong specimen. In general, Les Herbonautes transcribers seemed to interpret the more difficult to understand labels better than DigiVol transcribers with the DigiVol transcribers leaving more collection date values blank.

- 176 records (16%) matched between Les Herbonautes and DigiVol values but not GBIF.
- 25 records (2%) matched between DigiVol and GBIF values but not Les Herbonautes.
- 88 records (8%) matched between Les Herbonautes and GBIF values but not DigiVol.

2) Collector (s)

Collectors were compared using the same three matching criteria outlined in pilot one.

Only three specimens did not have a collector over all three methods. Of the remaining records in GBIF/JACQ, there were 108 records without collector information. Again, as we do not know how complete the data records sent to GBIF or JACQ are, we cannot assume that because there is no collector information in the data that this means a collector is not present on the label. It might be that this date field was not transcribed or the data in this field were not sent to GBIF. Therefore, these

Funded by the Horizon 2020 Framework of the European Union
H2020-INFRADEV-2016-2017
Grant Agreement No 777483

records were excluded leaving 1282 specimens for further analysis. Of these records, 51% of the records had matching collectors leaving 631 non-matching records.

Barcode	DigiVol Collector	GBIF Collector	Herbonautes Collector
B 10 0002869	G.D.H.	Haviland,G.D.	
B 10 0003913	Wiechert?	E. Royl & Wiechert	Royl-Wiechert
BM000075813	Wood, J.R.I.	John Richard Ironside Wood	Wood, J.R.I.
BM000500117	Kantvilas, G.	Gintaras Kantvilas	Kantvilas, G.
BR0000005110216	Wied, M. von	Wied M.	Wied-Neuwied, M.
BR0000005212705	Riehl, Nicholas	Wied M.	
E00001162	Forrest, G.	Forrest, George	Forrest, G.
E00008912	Forrest, G.	Forrest, George	
K000001916	Ganev, Wilson	Ganev, W.	Ganev, W.
K000013094	Taylor Zappi Eggli	Taylor, N.; Zappi, D.C.; Eggli	Taylor Zappi,Eggli
P00036507	collector unknown	Sag, G.	
P00039147	Jovet, M. P.	s.c.	Jovet, P.
TU250535	Leis, M.	Mare Leis	Lens, B.
TU250942	[Pihlapun, W.]	L. Pihlapuu	Pihlapuu, L.

Table 20: Example of specimen records with mismatched collector transcriptions between DigiVol, GBIF and Les Herbonautes

As can be seen in Table 20, although the records do not match due to the different ways of writing a collector name, many of these values can be resolved to the same collector.

Breakdown of mismatches

- 120 records (9%) were mismatched across all methods.

Funded by the Horizon 2020 Framework of the European Union
H2020-INFRADEV-2016-2017
Grant Agreement No 777483

All three different data sources were compared pairwise, in both directions. However, due to the fuzzy matching, pairwise comparisons are not transitive. Even if the collector for DigiVol matches the one entered in Les Herbonautes, and the collector from Les Herbonates matches the one derived from GBIF, it is still possible for the GBIF value not to match the value from DigiVol.

- There was a partial mismatch in collectors for 511 records (40%).
- 72 records (6%) matched between DigiVol and GBIF.
- 180 records (14%) matched between Les Herbonautes and GBIF.
- 286 records (22%) matched between DigiVol and Les Herbonautes.

3) Collector Number

Over all three platforms 96 specimens had no collector number. Of the remaining records in GBIF/JACQ, there were 394 records without any collector number information. As explained above, we cannot assume that because there is no collector number information in the GBIF/JACQ data, that a collector number is not present on the label. It might be that this field was not transcribed or the data in this field were not sent to GBIF, due to time limitations the original image was not checked to see if this was the case. Therefore, these records were excluded from further analysis leaving 913 specimens. Out of these remaining records, 535 records (58.6%) matched when punctuation was excluded.

Barcode	DigiVol Collector No	GBIF Collector No	Herbonautes Collector No
B 10 0001199	877a	s.n.	877 a
B 10 0001200	877a	s.n.	877 a
BM000521570		s.n.	
BM000521577		s.n.	
BR0000005327270		S.N.	
BR0000005575718	s.n.	S.N.	
E00016420	2160	10.964	10964
E00026850	3218DB	1978	1978
K000018358	Harley17428	17428	17428
K000056070	4006	PCD4006	4006

Table 21: Example of specimen records with mismatched Collector number values between DigiVol, GBIF and Les Herbonautes transcriptions.

Funded by the Horizon 2020 Framework of the European Union
H2020-INFRADEV-2016-2017
Grant Agreement No 777483

Collector numbers mismatched across all platforms accounted for 378 specimen records (41%).

Breakdown of mismatches:

- 129 records (14%) matched between DigiVol and GBIF/JACQ values.
- 73 records (8%) matched between Herbonautes and GBIF/JACQ values.
- 136 records (15%) matched between DigiVol and Les Herbonautes values.

Many of these matches occurred due to differences in representing a blank value. All "S.N." and its variants, e.g. "s.n.", were changed to empty values. Once the differences were excluded overall mismatches were reduced to 215 or 24% from 378.

Barcode	DigiVol Collector No.	GBIF Collector No.	Herbonautes Collector No.
B 10 0001199	877a	s.n.	877 a
B 10 0001200	877a	s.n.	877 a
BM000547082		947	947
BM000561307	s.n.	66	66
BR0000005212705	158	S.N.	
BR0000005264018	767	S.N.	
E00016420	2160	10.964	10964
E00026850	3218DB	1978	1978
K000018358	Harley17428	17428	17428
K000056070	4006	PCD4006	4006

Table 22. Example of specimen records with mismatched Collector number values between, DigiVol, GBIF and Les Herbonautes transcriptions after excluding for differences due to variants in representing no collector number, e.g. s.n. and S.N.

Breakdown of mismatches

- 22 (2%) records did not match across all platforms
- 35 (4%) records matched between DigiVol and GBIF/JACQ.
- 82 (9%) records matched between GBIF and Les Herbonautes.
- 76 (8%) records matched between Les Herbonautes and DigiVol.

Funded by the Horizon 2020 Framework of the European Union
H2020-INFRADEV-2016-2017
Grant Agreement No 777483

3.2.2 Discussion Transcription pilot 2

These results show that accurate transcription is only one step needed to create an occurrence record of suitable quality for use in research. How this data is imported into collection management systems, ingested by aggregator sites, displayed and then presented on download for use by researchers is also important. Along this pipeline valuable data may be removed or its display altered that could change the meaning of the data value. Our results show this clearly with collection dates where date ranges can be reduced to a single date of collection and where collection dates without day and month are changed to 01 and 01 thus giving a false indication of precision within GBIF. This may be due to software limitations, but mechanisms should be put in place to avoid it.

Common errors for transcribing collection dates, like in transcription pilot one, included adding determiner, bequeathed, communicated or accessioned dates instead of collector dates or entering a collection date from another specimen on the sheet. Easier access to improved explanations and examples of these common errors would help to reduce these transcription errors.

Transcriptions from Les Herbonautes had fewer errors for collection dates than DigiVol transcriptions. This may have been helped by the fact that in Les Herbonautes each value would be transcribed a number of times and a consensus reached. However, in DigiVol each specimen is transcribed only once and then the record is validated by a second person (usually by staff at the institute or a volunteer trained in validation). For the transcription pilot, records were not validated. This enabled us to compare records transcribed through crowdsourcing alone rather than crowdsourcing and validated by institute staff. Reasons for many mismatches in collectors and collector number were the same as in transcription pilot one.

It was also seen that transcribers found multi-specimen sheets more difficult to transcribe as they could not work out which collection label went with which barcode or catalogue number. This was particularly difficult in this expedition as there were a variety of different barcode and occurrence id formats due to the number of different collections represented. In some instances, the barcode or catalogue number was not placed next to the label due to space limitations and labels were cross-referenced using numbers or letters. Improved explanations and visual examples of how to determine the correct label to transcribe is required. Multi-specimen sheets would probably merit a higher level of record quality control and validation. If labels could be segmented automatically and linked to the catalogue number or barcode it may be worth experimenting with presenting transcribers with just the label for transcription. However, for many sheets linking the labels with the correct barcode and synthesising this information back to together might be too time consuming a task to warrant this segmentation process.

Comparison with a chalcid wasp microscope slide crowdsourcing project

The analysis of data quality in our pilots has been based on transcription of Herbarium specimens, whilst many aspects of label transcription will be similar with information on who collected it, when and where it was collected, each different collection will have its own unique challenges for label transcription and the addition of different fields depending on the collection.

Funded by the Horizon 2020 Framework of the European Union
H2020-INFRADEV-2016-2017
Grant Agreement No 777483

The Natural History Museum London used the Crowdsourcing Platform Notes From Nature to transcribe Chalcid microscope slides. Chalcid wasps are within the superfamily Chalcidoidea and part of the order Hymenoptera. Most of the species are parasitoids of other insects, attacking the egg or larval stage of their host, though many other life cycles are known. The following information was taken from an unpublished data quality report produced for the Synthesys project in December 2017

Fig. 17. Chalcid wasp expedition on Notes From Nature crowdsourcing platform

Three batches or expeditions were put up on the platform. The data categories transcribed were collection date(s); country; collector(s); host Insect and host plant; Type specimen; and registration number. For all three batches each slide was transcribed three times and the data then automatically reconciled into a single set of transcriptions. Between batches some changes to the user interface was changed, such as the addition of an "unclear" option for 'Collector' and 'Host Insect and Plant', however, the information requested and layout remained unchanged between the batches.

A subset of randomly selected records was manually checked for errors in each of the three batches by the same person. Each of the data categories was scored for errors and missing data. Even if there were several errors per category the scorer would only record a maximum of one error per category. For example, if there was an error in the month and the year in the collection date category this would be recorded as one error in the collection date category.

There was no substantial difference in the number of slides with errors among the three batches of transcriptions. Batch two had the highest number of slides with errors (53%), while batch one and three had similar number of slides with errors (45 and 43%, respectively). However, the quantity of incorrect and missing data errors per slide decreased in Batch 3.

Between the three batches the category with the most errors varied between 'Host Insect and Plant' and 'Collector', which is unsurprising given the complexity of the data associated with these categories and the free-text nature of the field. 'Collection Date' also had a high number of errors in each batch. The vast majority of the errors occurred due to the issue of blank entries not being considered as data

Funded by the Horizon 2020 Framework of the European Union
H2020-INFRADEV-2016-2017
Grant Agreement No 777483

resulting in erroneous data from a single transcription being carried through to the final reconciled data. If the issue of blank fields is addressed, then 'Collection Date' could become a more robust dataset. Overall, the 'Country' level data had minimal errors that consisted of missing data rather than erroneous data, with the exception of Batch 3.

Analysing batch three in more detail, 'Host Insect and Plant' had the highest error rate (24%), followed by 'Collector', 'Collection Date' and 'Registration Number' (9%). 'Country' and 'Type Specimen' had the lowest error rates (5%;)

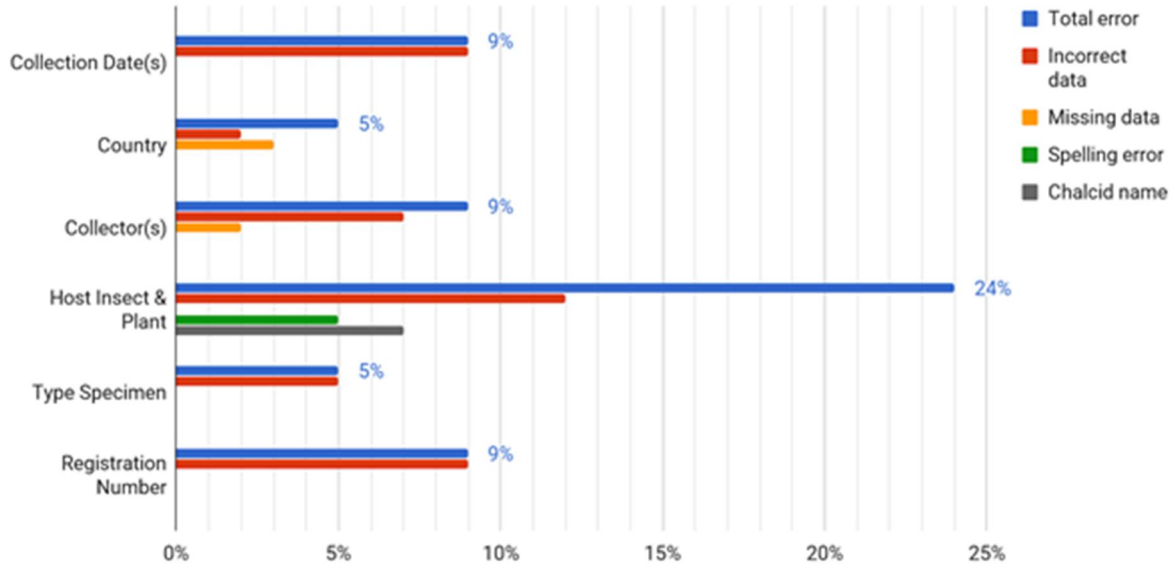


Fig. 18. Percentage of transcription errors for each of the six transcribed categories in the Chalcid wasp expedition on Notes From Nature crowdsourcing platform

The majority of incorrect data for 'Host Insect and Plant' (free-text field) was due to confusion between the insect and plant host taxa resulting in this data being inserted into the opposite field, while the second most common error was inserting the Chalcid name itself.
Errors related to the 'Collector' (free-text field) were due to species author or determiner names being transcribed instead of the collector. 'Collection Date' (drop-down list) errors were only composed of incorrect data being entered, the majority of which were present in the reconciled data from single transcription entries due to the blank entry issue stated above. The 'Country' (drop-down list) field consisted of a mixture of missing data and incorrect data. The incorrect data was a result of a historic country name being incorrectly associated with a present-day country. These slides were, however, tagged to say that the country was not explicitly stated on the slide and therefore could be checked. Errors related to the 'Registration Number' (free-text field) Barcode or project numbers were sometimes recorded. Batch cleaning of these errors will be simple as the correct format for registration numbers is easy to spot.

These results show similar challenges to the herbarium specimen transcription with transcribers making errors on identifying the collector on the label getting confused with the determiner or species

Funded by the Horizon 2020 Framework of the European Union
H2020-INFRADEV-2016-2017
Grant Agreement No 777483

author. However, the transcribers had more of a challenge with the taxonomic names on the label itself determining the plant, from the host from the parasitic wasp itself. Providing clear links to names resources is vital to assist with this identification, where possible dropdown names list would also help. Country level information had the lowest % of errors and this was also seen in transcription pilot one where instances of incorrect country values, incomplete values or values in a different format was also seen to be low across all of the datasets.

3.3 Recommendations for Transcription

From the results of the transcription pilots we suggest a list of recommendations to improve the accuracy of label transcriptions as a step towards improved data quality. These recommendations are valid for all methods of data transcription.

- Implement methods for transcribers to indicate when the information is not present on the label. This will enable users to know if the information is not on the label or has not been transcribed.
- The establishment of a standard for Collector names. We recommend the presence of at least two fields for collector name capture - surname and given names where possible, as well as strongly desirable additional fields including title. This will give the transcriber maximum flexibility when transcribing the name. Titles are important for distinguishing individual collectors, especially for historic collections when women often collected under their husbands' name with only their title to tell them apart.
- The establishment of a standard collector identifier is recommended. The transcriber could then enter the name as given on the label but in addition link to a standard name. If the look-up list provides extra information e.g. known collecting localities, collection dates then this information will help transcribers select the correct standard name.
- There should be separate fields for each collector allowing for a look-up list in each case.
- Collector Date should allow for the use of ranges with the presence of six fields. The issue of substituting blank values with a default of 01 should be addressed as it gives a false indication of precision.
- Additional date fields such as received or communicated dates could be added for records to be comprehensive and protocols improved with instructions on what to do when these are found.
- A standard should be adopted for handling missing century data, or consistency checks (using other information such as collector life span), implemented at the data-entry or data-cleaning steps.
- There establishment of a standard for collector number addressing how to transcribe prefixes or suffixes. We recommend separate fields to assist with searching.
- Where possible enable the download and use of the DiSSCo taxonomic backbone for use in transcription to reduce taxon name errors.
- Transcription uncertainty flags are recommended for all fields to attract review without dirtying the data itself.
- A taxonomic names resolution system so names can be checked before occurrence data is ingested into the DiSSCo infrastructure. A mechanism should be implemented to feedback suggested changes to the taxonomic backbone.

Funded by the Horizon 2020 Framework of the European Union
H2020-INFRADEV-2016-2017
Grant Agreement No 777483

- Transcription platforms should provide more examples and images of common label transcription errors to avoid replication of these errors.

4. An Electronic Marketplace for Transcription

The rise in crowdsourcing has led to a large number of online websites which offer an online marketplace where workers complete atomised tasks to more complex multi-faceted tasks for a small amount of money. There are a large number of websites available in a quickly changing landscape - examples include:

RapidWorkers - www.rapidworkers.com

Samasource - www.samasource.org

Clickworker - https://www.clickworker.com/download-app/

Crowdguru - http://www.crowdguru.de

MicroWorkers - https://www.microworkers.com/

CrowdSource - https://www.crowdsource.com/

Prolific Academic (ProA)- https://www.prolific.ac

Figure Eight - https://www.figure-eight.com/ - formerly crowdflower

One of the first and best known is Amazon Mechanical Turk (MTurk) - https://www.mturk.com/ where users known as "Workers" can select to complete small tasks referred to as Human Intelligence Tasks or "HITs" for a monetary reward. The type of tasks available include identifying objects in an image and tagging them, categorising images, transcribing audio recording or completing surveys. To post tasks and become a "Requester", you must be over 18 years of age, create an account and need to prepay the cost of the work with a credit or debit card. Tasks can be submitted through an online interface or through the using the MTurk web services API or command line tools. When creating a task it is necessary to decide how much to pay for each task completed, how long Workers are allowed to work on that task, and the number of people you want to complete the same task. The maximum number of HITs in a batch is currently 250,000.

The total price that Requesters need to pay for each task posted comprises the price to the Worker for completing the task and a fee for Amazon Mechanical Turk. The minimum fee is $0.01 per assignment and the Amazon Mechanical Turk fee is an additional 20% of the reward Requesters pay workers. Typical costs for a task are between $0.02 and $0.5 per task. There are also some additional fees if Requesters want only workers that have been assigned a master's degree or "Premium Qualification" to complete their tasks, or submit HITs with 10 or more assignments, for example. As Workers complete tasks they will have statistics associated with them based on how they have completed those tasks. For example, "Master Workers" will have demonstrated the ability to

Funded by the Horizon 2020 Framework of the European Union
H2020-INFRADEV-2016-2017
Grant Agreement No 777483

complete specific types of tasks for multiple Requesters. It is also possible to assign a custom "Qualification Type" tag, such as for Portuguese speakers, for example.

Requesters have the opportunity to approve completed HITs before paying for them. They can manually approve, auto-approve without looking or auto-approve when their answers to a HIT are the same. However, if the tasks are not approved within the specified deadline set during the set-up of the task, then workers will be automatically paid by MTurk. It is possible to reject Workers' tasks so they do not get paid but there must be a valid reason for this as it may affect the reputation of the Worker and the Requester.

MTurk has been used to gather research data by thousands of social scientists from many different fields (Burhmester *et al.*, 2018). A Google Scholar search suggests that approximately 15,000 papers containing the phrase "Mechanical Turk" were published between 2006 and 2014, including hundreds of papers published in top-ranked social science journals using data collected from MTurk (Chandler and Shapiro, 2016). MTurk was found by researchers conducting psychology experiments to be efficient, relatively inexpensive and provided data that met or exceeded the psychometric standards associated with the published research (Buhrmester *et al.*, 2011). Merged multiple transcriptions of non-native speech from MTurk were found to be as accurate as an individual expert transcriber when transcribing the speech of speakers reading a paragraph out loud and only slightly less accurate for the spontaneous responses where speakers were talking about a particular topic but not following a script (Evanini *et al.*, 2011).

A recent evaluation by Buhrmester *et al.*, 2018 states that since 2011 there have been over 40 articles in which some form of evaluation of MTurk's suitability has been conducted. Although MTurk seems to be the most widely used there have also been a few studies evaluating similar platforms. Peer *et al.*, 2016, states that MTurk exhibits slowing rates of population replenishment and a growing number of participants who complete experimental tasks and questionnaires more than once so they investigated alternative platforms. They compared Crowdflower (CF) and Prolific Academic (ProA) against MTurk and found that both platforms participants were less dishonest compared with MTurk. CF showed the best response rate but failed more attention-check questions. Their comparison methods also showed that ProA produced data quality that was higher than CF's and comparable to MTurk's.

So, could electronic marketplace platforms like MTurk be a mechanism for enriching or transcribing label data from Natural History Collections?

Wilis *et al.* 2017, assessed the reliability of using non-expert workers (i.e. MTurk) against expert workers to collect phenological data from herbarium specimens. The 'non-expert' pool comprised 270 anonymous workers hired through the Amazon's MTurk platform and the expert workers consisted of four curatorial staff members. Non-expert workers were paid at the rate of $0.10 per image plus a $0.15 base participation rate for each set of 10 images. Before beginning their data collection, expert and non-expert workers were required to watch a short (c. 1 min) video giving instructions on how to collect data on the crowdsourcing platform CrowdCurio (http://crowdcurio.com/). Both sets of workers were also presented with example images for each phenological trait, for each of the species. For both experts and non-experts, the time duration required to score each image was recorded. From analysis of the results it was found that there was no difference in the data quality of non-experts and experts, however non-experts were a more efficient way of collecting more data at lower cost.

Funded by the Horizon 2020 Framework of the European Union
H2020-INFRADEV-2016-2017
Grant Agreement No 777483

Even though experts were significantly more efficient at counting phenological traits on a per specimen basis, processing specimens, on average, c. 0.80 min faster than non-experts. The total MTurk costs were significantly lower with a per image cost of $0.16 compared to $0.80 for experts.

A	Total number of images scored on MTurk	4197
B	Cost paid to MTurk worker per scored Image	$0.10
C	Number of MTurk assignments	270
D	MTurk fee per assignment	$0.86
E	Number of MTurk workers	270
F	MTurk worker baseline fee	$0.15
G	Total of images scored by experts	2,560
H	Total expert work hours	59.6
I	Expert Hourly Cost Rate	$34.37
J	Total MTurk costs (A*B) + (C*D) + (E*F)	$692.40
K	Total Expert costs (G*F)	$2,048.45
	MTurk cost per Image (J/A)	$0.16
	Expert cost per Image (K/G)	$0.80

Table 23.: Comparison of costs for scoring Herbarium specimen Images for phenological traits by expert workers and non expert workers on MTurk. Data from Wilis et al. (2017)

The use of online websites like MTurk has raised ethical questions. Wilis *et al.* (2017) acknowledged this in their study noting that non-experts earned a salary of $3.14h below the federal minimum wage of $7.25 h. They also ensured that the use and collection of data by non-experts was reviewed and approved by an ethics review committee at the University of Waterloo. Others have also raised the issue of fair pay (Chandler and Shapiro, 2016; Anderson and Lau 2018; Sheehan 2018; Pittman and Sheehan 2016). Gleibs (2017) calls for a need to develop a best practice for researchers using crowdsourcing tools.

Funded by the Horizon 2020 Framework of the European Union
H2020-INFRADEV-2016-2017
Grant Agreement No 777483

Mankar *et al.* (2017) posted a dummy task on MTurk which informed Workers that they could not work on the task because it paid less that their local minimum wage. Instead of the task, Workers were asked for feedback which they were then paid for. Mankar *et al.* found that for those Workers expressing an opinion, two-thirds of those in India favoured the policy while two-thirds of American Workers opposed it. When surveyed, the majority of Requesters also supported a minimum wage pay, but only 20% would enforce it.

In conclusion, there is evidence from its use in other scientific domains, particularly the social sciences, that platforms like MTurk can produce useful quality data for scientific study and, as such, warrants a pilot to see if this can be replicated for the natural history collections community. However, there are ethical considerations to take into account including minimal pay and the simultaneous use of volunteer non-paid crowdsourcing sites collecting the same type of data.

5. Organising a Large Transcription Project: Insights from a Commercial Party

Over the past five years, Picturae, a company specialised in digitising, storing and hosting digital heritage collections, has gained a lot of experience in the execution of large-scale herbarium digitisation projects, including the transcription of collection data on herbarium labels. The first and largest project with ca. 3 million labels, was for the Naturalis herbarium collections, Leiden, The Netherlands. Later, other projects followed: for collections within the herbaria of the National History Museum and the Royal Botanic Gardens, Kew, both in London, United Kingdom, the University of Oslo, Norway, the Smithsonian Institution in Washington, the Muséum National d'Histoire Naturelle and the Université Claude Bernard in Lyon, both in France and for Meise Botanic Garden in Belgium.

From the start of the Naturalis project in 2013, Picturae has worked with the Suriname-based company Alembo. In this partnership the actual transcription of the labels is in the hands of Alembo, whereas Picturae focuses on overall project management, communication with all parties and quality control (QC).

This chapter discusses how, both Alembo and Picturae make large herbarium digitisation projects successful and describes the advantages and disadvantages of outsourcing large numbers of transcription work to a commercial party. Furthermore, the following more practical aspects will be described: setting the rules for clear and unambiguous transcription in a protocol, setting-up workflows, look-up lists and time frames. Finally, some ideas will be shared on the use of an electronic marketplace for transcription work.

5.1. Starting up a large-scale transcription project

5.1.1. Transcription rules

Clear and unambiguous transcription rules are critical for the success of any project. For the first projects completed by Picturae those rules were set by the herbarium staff, which resulted in all kinds

Funded by the Horizon 2020 Framework of the European Union
H2020-INFRADEV-2016-2017
Grant Agreement No 777483

of problems. Rules were too elaborate and therefore difficult to understand for non-herbarium specialists. Furthermore, some rules in the protocol contradicted themselves or each other, confusing the transcribers. There were also errors in the protocols. Correcting and clarifying those rules took a lot of time and discussion from all parties involved. The changing rules also required a lot of work on the part of the transcription and QC staff. Once transcribers learn certain rules, it takes more time to unlearn those rules and learn new ones, than it did to learn the rules the first time.

Therefore, it is advisable to develop the protocol together with the actual transcribers. They have a different focus from the herbarium staff and can complement them. It does help if the party to which the work is outsourced has some understanding of natural history collections, to avoid problems in the process. Their understanding, and at the same time, their distance from the collection, helps to create rules that work best for both the transcribers and the client. For example, information that is written in the body of the label or in the title can differ. If this case is not specified, rules can be contradictory. If the transcription partner has delivered on projects like this before, they know which questions to ask, to make sure both client and transcribers are interpreting the rules in the same way.

A few years ago, Picturae created a general protocol based on experience from previous projects. This protocol contains general guidelines for recognizing the correct data and describes the transcription rules for the most common fields. As every collection has its own particularities, it functions as an example and starting point for discussion with the herbarium staff about the final rules. To make sure the transcribers meet the collection requirements and will work according to the wishes of the collection staff, the rules are adapted during the start-up of the project and the testing of the transcription.

Still, a lot of time and communication is needed from all parties involved. Generally, the protocol goes back and forth for revisions several times before being approved and multiple meetings are needed to discuss changes. In practice, after transcription starts, new questions and problems arise. Therefore, though a protocol which is generally accepted by all parties is required before transcription starts, a definitive version of the protocol is only accepted after four to six weeks of transcription, allowing some last minor changes to be done to optimize it. In practice, by then about 10 – 20,000 records have already been transcribed.

During the first project for Naturalis, it took half a year before the protocol reached its final state. Even then, after an extra half year, another last change was made. This was very confusing, slowed down the process and increased the risk of errors as transcribers started to doubt themselves. Lessons learned: rules cannot be changed after the protocol has reached a definitive state, this to enable the transcribers to reach maximum speed and quality. To make sure the protocol is following the wishes of the client, its settlement is top priority during the start-up of the project.

5.1.2. Logistics of a large-scale transcription project

The logistics for a large-scale transcription project can be quite complex. Communication is an important part for any project to be a success. With all the current possible methods of communication there are plenty of ways to get in contact. However, it helps if e-mails are answered quickly. Often, especially during the start-up of the project, it is easier to use Skype or telephone. A visit by the transcription partner to the natural history institute or vice versa is recommended. The better the parties know each other, the easier it is to work together. It can be difficult to keep track

Funded by the Horizon 2020 Framework of the European Union
H2020-INFRADEV-2016-2017
Grant Agreement No 777483

of all e-mails and other conversations, so we recommend recording the most important agreements in official documents with version control, instead of using e-mail.

An example of such a document is the transcription protocol, but also the planning document that can be updated every week. A third document that may be helpful is one for label interpretation. During the project, but especially in the beginning, new insights and exceptions can result in questions on how labels should be transcribed. For those cases a label-interpretation document can be created and shared with all participants, for example, in Google drive. This document contains examples of labels which raise questions with the transcribers and data control. Furthermore, the collection staff can insert examples of exceptions or general feedback on labels.

Setting up data-flows and software tools

There are two data-flows to be taken into account:

1. The images and metadata to be sent to the transcription party.
2. The complete transcription project, from the start of the transcription until the final import into the client's database.

Often large-scale transcription projects are simultaneous with, or follow, a large-scale digitisation project. The following graph gives an idea of what the complete process from scanning until final delivery of transcription might look like once all the rules are clear and the workflow is working properly:

The complete workflow including imaging also depends on what is agreed with the client and may differ per project. For example, often there is a larger time span between imaging and transcription, allowing the transcription partner a larger backlog to make sure they don't run out of work in case of any issues with imaging. It is recommended to plan extra time at the beginning of the project, to make sure the imaging process is running smoothly first. Only then is transcription begun and if required the delay can often be caught up during the rest of the project. Another option is to completely separate the transcription process from the imaging process, beginning only after imaging has (almost) finished.

Generally, the image files created during the imaging process are very large. As a result, the files are very slow to open and not very suitable for the transcription process. Furthermore, large numbers of resources are required to send and store the full-sized images for the complete duration of the project. Therefore, as only the labels need to be legible, it is advisable to generate small-sized JPEG derivatives of the scanned images, especially for transcription and QC. However, the overall quality of the image is not an issue for transcription; generally an image only 1MB in size is large enough. Of course, these kinds of actions require a lot of computing power to keep up with the imaging process. Picturae has workflows in place which allow large numbers of images to be processed and sent to Alembo every day during a digitisation and transcription project. Once the correct scripts are installed, the process works automatically, only needing a restart if a network problem shows up.

The second data flow mainly deals with the files resulting from transcription. For the Naturalis project an institutional application was used for transcription: BRAHMS. Though it is probably a good system for the average work that is being done in an herbarium, it was not efficient as a basic transcription tool. Mainly, the program was slow and crashed several times a day, requiring a restart after each crash. This was partly because the full institutional database was used as a the look-up, including much information that was not actually needed for the look-up itself, like previously transcribed records.

Funded by the Horizon 2020 Framework of the European Union
H2020-INFRADEV-2016-2017
Grant Agreement No 777483

This database grew significantly during the project as transcribed files were imported into the system. As the database increased in size, the look-up needed more time to load. The last issue arose because every BRAHMS file consisted of two separate files. If the two were separated somewhere during the process, BRAHMS threw an error.

For every day of imaging, several batches were created requiring transcription. For the sheets alone, in total almost 5,000 batches were created and transcribed, every batch generating two complementary transcription files. The bug-tracking tool called Mantis was adapted as a workflow tool. This tool kept track of all batches and files, and where they were in the transcription process. It also registered the numbers of records per batch, status changes etc. both for creating reports and for invoicing purposes. The use of those two tools involved downloading the two transcription files from Mantis, opening them in BRAHMS for transcription or quality checks, saving them and then uploading them again into Mantis for every step of the process. Needless to say, a lot could (and did) go wrong in this process.

Therefore, after the first two projects, Alembo developed a tool called DETA. This was a new tool especially aimed at the transcription of herbarium specimens, though it can be used for other transcription projects as well, e.g. natural history collections. All work is done in a central database which is stored on one of Alembo's FTP-servers which are accessible only through a secured VPN connection. Once access is granted, all participants who work with the software can access the same data. The application is used both for transcription and QC, and as a workflow tool. This means that not only is transcription done in DETA: batches can be assigned to transcribers; after transcription they are assigned to the QC team; after which a QC team member performs a check and assigns the batch to the Picturae QC team. As a result, it is no longer necessary to export and import files from one system to the other between different steps in the process.

JPEG-derivatives of the images, as well as a list of barcodes, or other codes corresponding to the file names of the images requiring transcription, can be imported into the system. For every barcode or file name a record is created. As it was in BRAHMS, upon editing the record, the corresponding image is opened automatically to make sure the correct image is always transcribed. After transcription and quality checks, a CSV file (or other format) can be exported from the tool and sent to the herbarium staff. For this part of the workflow, Mantis is still often used, but as in this situation there is only one file to consider and only two parties use the system, any problems which arise are less severe or frequent.

For many fields it is possible to import pre-existing authority lists with information into DETA to help the transcribers to interpret and transcribe the labels correctly, for example taxon lists, collector lists and country lists. Per field, those lists are available through a look-up window. It is also possible to automatically add information to other fields associated with the value in the field for which the look-up was originally used, for example, a collector ID with the collector, or an Author with a taxon name. The lists can be updated during the course of the project, when more information is added to the client's database. However, as only the lists with names are entered into the system, and not the complete database, which contains many duplicates (e.g. in collector names), the size remains relatively small and manageable.

Over the past few years, with every new project, new functions were added and the original program expanded and improved. In 2017 an extra module was added to allow the collection staff to do their

Funded by the Horizon 2020 Framework of the European Union
H2020-INFRADEV-2016-2017
Grant Agreement No 777483

checks in DETA as well. This allows a much easier workflow and direct access for the herbarium staff to the feedback system. An easy way to give and receive feedback is very important for a quick learning curve with the transcribers. A QC tool was built in to DETA. With this tool it is possible to give feedback directly to the transcriber of the record. The transcribers can see the feedback they receive from any QC team that works in DETA, directly in their own feedback overview, including the transcription they did and the image that belongs with the record.

5.1.3. Training new transcribers

Training new transcribers is more complicated than just having them read the transcription protocol and getting started. The average transcriber will understand little of the protocol until they have actually begun transcribing labels. The largest issue with herbarium collections, and natural history collections in general, is that over the centuries many people have been working on them with very few standards detailing which data are relevant and how to document them. As a result, there are as many types of labels as there are collectors and not all labels are straightforward. It takes time and training even to recognize every type of data on different labels. For example, there are many labels that clearly state coll., or collector, or leg. or rec., or recolté. Those terms can be quoted in the protocol to enable the transcriber to recognize the collector. However, not all collectors are marked with a term like this, so it takes training to recognize them on other labels as well.

It helps a lot if there are experienced people available for training. Also, prompt feedback is needed. The more new transcribers who need training, the more work it is to check their work and give sufficient feedback to speed up their learning process. This issue (partly) caused the delay in starting up the first transcription project with Naturalis. Naturalis simply did not have sufficient staff to give feedback to the many new transcribers needed for the large quantity of data. In combination with the protocol problems explained above, this resulted in a very long start-up period in which transcription went too slowly and many batches required corrections.

Since that first project, training has become much easier, as the basic knowledge is present now. When starting a new project now, the transcribers work with a quality and speed which was only reached after half a year of transcription during the Naturalis project. Nonetheless, training takes time. Therefore, we recommend organizing one larger project rather than carrying out several smaller projects. As large projects benefit from the longer experience and training of the transcribers, they will, in the end, bring down costs per specimen.

A few years back, Picturae conducted a test with another outsourcing partner. The quality of their work was at about the same level as the work Alembo delivered in the beginning of the Naturalis project. It was therefore decided too much effort would be required to train a new partner in doing the same work.

5.2. Process and workflow

As soon as the protocol is finished, all workflows are in place and the application is ready, it is wise to test the transcription. During this testing phase many questions and exceptions arise, which usually result in some changes to make to the protocol. The testing phase is also a good moment for the institute's staff to try out the import of the transcription files to their database system. After the

Funded by the Horizon 2020 Framework of the European Union
H2020-INFRADEV-2016-2017
Grant Agreement No 777483

testing phase, which generally takes 2 to 3 weeks, most of the rules should be clear and to the satisfaction of all parties. If not, it is better to do an extra test and extend this phase a bit longer.

After the testing phase, it is best to start transcription with a small group of experienced transcribers. As many more labels are transcribed now, more questions will arise. During this phase of the project it is crucial for the institute's staff to check all batches as soon as possible, preferably within a week, but at latest within two weeks, to provide prompt feedback on the work being done. Prompt feedback is essential to finding issues with the transcription rules as soon as possible and to further a steep learning curve with the transcribers. By the end of this phase, which normally lasts 4 to 5 weeks, some last minor adjustments may be made to the protocol if needed. After this period the protocol needs to be accepted by all parties to reach a definitive status and no more changes can be made.

After the definitive version of the protocol has been accepted, the main production phase starts. Depending on the size of the project, additional transcribers need to be trained during the first 6 weeks or so and the focus should be on reaching the required quality and speed for transcription. Like in the previous phase, to reach the required quality it is very important that the institute's staff checks the batches soon after delivery, again preferably within a week, two at the most. After a few weeks the quality improves and less QC should be needed. After that it is also possible to delay checking a bit, though we recommend continuing to check batches within two weeks after delivery throughout the project. Provided the quality is good, to keep up with QC, it is possible to decrease the size of the sample checks as soon as training is complete.

Start-up	•Making the protocol •Application Setup •Import of Look-up lists •Workflow installation
Week 1-2	•Testing application. •Testing transcription. Test batch delivered to client in week 2. •Adjusting protocol if needed.
Week 3-7	•Training transcribers. •Start transcription. First batches delivered to client by end of week 4. •Last adjustment of protocol if needed in week 7, definitive version.
Week 8-13	•Start of main production phase. •No changes can be made to the protocol after this point. •Start extra transcribers if needed. •Accelerating transcription.
From week 14	•Top production speed for transcription. •The number of records delivered per week depends on the total number and planned end date or dead line.
Depending on total and dead line	•Transcription is finished.

Table 24.: an overview of the full transcription process.

Funded by the Horizon 2020 Framework of the European Union
H2020-INFRADEV-2016-2017
Grant Agreement No 777483

Table 24 shows an overview of the transcription process. This table is set in context below, with two examples of large projects and the numbers that need to be achieved in order to finish the project on time. The first column contains the week number for the project, week 1 being the start of the project. The second column shows the weeks from the table above. The third column gives a short description of the work being done and the fourth shows the average number of sheets or records that need to be transcribed per week during that period. In the last column an estimate of the number of FTEs needed for transcription is given, based on a 40-hour working week and a comprehensive transcription of the sheet labels: collector, collector number, collection date, cultivated (yes/no), country, precise locality, elevation and coordinates. Taxon names are normally entered from the folder and these labels are typically processed much faster. Collection curators may spend a significant amount of resource in an imaging preparation stage ensuring the names on the folders are clearly written to ensure transcription is easier. Preparation may also include checking specimens are filed according to up to date taxonomy and checking identifications although the resources needed for this may be prohibitive for large scale digitisation.

For the first project we needed to transcribe 3 million sheets in two years, a total of 104 weeks:

Week number in project	Weeks following chart above	Description	Average number of records per week	Average FTE needed per week for transcription
1	Start up	2 months/ 9 weeks	0	
10	Week 1 of transcription	Test 2000 records	2000	2
11	Week 2	Quality check by herbarium staff and protocol adjustment	0	
12	Week 3-7	Training core team	4000	3.3
17	Week 8-13	Training additional transcribers and accelerating	22500	14
23	From week 14	Full production	40000	23
94	Week 85	Transcription finished.	0	
104	Week 95			

Table 25: Breakdown of activities into weeks for an outsourced transcription project that transcribed 3 million sheets in two years.

For the second project we needed to transcribe 500,000 sheets in one year, or 52 weeks:

Funded by the Horizon 2020 Framework of the European Union
H2020-INFRADEV-2016-2017
Grant Agreement No 777483

Week number in project	Weeks following chart above	Description	Average number of records per week	Average FTE needed per week for transcription
1	Start up	2 months/ 9 weeks	0	
10	Week 1	Test 1000 records	1000	1
11	Week 2	Quality check by herbarium staff and protocol adjustment	0	
12	Week 3-6	Training core team	4000	3.3
16	Week 7-10	Training additional transcribers and accelerating	9000	5.6
20	From week 11	Full production	14000	8
47	Week 38	Transcription finished.	0	
52	Week 43			

Table 26: Breakdown of activities into weeks for an outsourced transcription project that transcribed 500,000 sheets in one year or 52 weeks

The number of records that needs to be completed per week depends on the total number of specimens and the duration of the project. It is possible to transcribe a smaller number of records every week, which means fewer transcribers are needed, but the project will take longer to complete. Vice versa, if the project is shorter, more records need to be processed per week, which means more transcribers are needed. The last column shows only the number of FTEs needed for transcription and does not count the people needed for training, QC, project management and IT-support and development. Even when outsourcing transcription work, for large projects the natural history institute may need to hire extra people to keep up with the quality checks.

With those larger projects, problems may arise that will slow down the process temporarily. Those problems can range from technical issues to transcribers or QC personnel falling ill. Therefore, it is wise to implement some extra time at the end of the project, to make sure the work is finished before the deadline. Those extra weeks can be used to catch up in case of any issues delaying delivery.

5.3. Final stage of the project

Even once a batch has been transcribed, checked and accepted by the institute's staff, the institute's work is not yet finished. They would be well advised to begin the last part of the workflow as soon as the first transcriptions have arrived and been accepted. First, the transcription files should be imported into the database. Many content management systems appear to be very particular about

Funded by the Horizon 2020 Framework of the European Union
H2020-INFRADEV-2016-2017
Grant Agreement No 777483

file formats and the manner of importing. Normally a transcription partner can deliver any format that is needed, but it is best to trial during the testing phase whether the chosen format actually works. If not, this is the best moment to change something in the format of the delivered files, to make sure there are no unwanted surprises later in the process.

The linking of the records to the images must be tested as well. Often the images have already been imported at a previous stage, but this can be done simultaneously with the transcription as well. With the large number of new records, it is difficult to keep track of everything and though the workflow tools help, very rarely it can happen that, for example, an image batch was never entered into the transcription system, or a batch is transcribed twice. To prevent problems in the database, we recommend implementing some checks on the delivery of the records.

As a last step of the process we recommend cleaning up the data. This could be done during the project itself, or separately after transcription has completely finished. With an overall error rate of only a very small percentage, entering a million records, or more, could still mean thousands, even tens of thousands of records with an error. Records from previous projects or transcribed by the institute's staff, are known to contain some errors as well. As many new records are added during the course of the project, this is a good time to start cleaning up. With the growing size of the database, it will be possible to create checks for inconsistencies in the data (e.g. a collection date is in a year the collector was not alive, the collector numbers for one collector are increasing quite consistently over the years, but then there is one that is only 3 digits instead of 4, the country is Ethiopia, but that collector never went to Ethiopia, etc.).

A complicating factor can be that there are errors on the labels themselves, or that some label information is consequently interpreted incorrectly. It has happened that a collector, who lived in the 19th century, was assigned collection dates from the 20th century in the database, because he wrote the years with only two digits and those were sometimes interpreted as being from the latter century. Upon cleaning up the data, all years were incorrectly changed to the 20th century. It is therefore recommended to always check with an external source in case of incongruities within the dataset.

5.4. Ideas on using an electronic marketplace for transcription services

The question of whether an electronic marketplace could be used for transcription services in chapter 4 was that it warranted a pilot. But what would a natural history institute need to take into account when organizing a large-scale project through such a marketplace?

The first point to consider is the size of the project: Mechanical Turk, which is the platform described in chapter 4, has a maximum of 250,000 tasks, in this case specimens. As many natural history institutes have many more, often even millions of specimens, this would mean splitting the project up into smaller parts for transcription. This may be an advantage if the institute has limited funds and wants to organize the transcription in smaller units, but it may be disadvantageous if the institute has funds to process all or a larger part of the collection at once because of the cost-advantages for larger projects. On the other hand, other platforms may offer the possibility of a higher number being done for a single task.

Funded by the Horizon 2020 Framework of the European Union
H2020-INFRADEV-2016-2017
Grant Agreement No 777483

Another factor to take into account is the training of the workers. As explained above, training transcribers to enter the data exactly following the wishes of the client is not an easy task. This may lead to problems when using an electronic marketplace, as the transcribers don't have direct contact with the institute's staff. This problem can be counteracted by asking them to enter only one simple field, like in the example in chapter 4, where only phenological data was entered. Similarly, the workers could be trained to enter only the collector, date or country. It becomes more problematic if all information needs to be transcribed at once. Most herbaria ask for the following fields to be transcribed: taxon name, collector, collector number, collection date, cultivated (yes/no), country, precise locality, elevation and coordinates, or a selection thereof, but other fields can be added as well. Training transcribers to enter all of those fields correctly is the greatest challenge. On the other hand, when trained correctly, it is cheaper to transcribe all fields simultaneously, as the transcriber has only to open and interpret the image once. Entering every field separately would require opening and interpreting the image several times, which takes extra time, resulting in extra costs.

A possible advantage of using an electronic marketplace is the potential to give transcribers labels written in their native language. If they also transcribe labels from their own region, this has the additional advantage that the transcribers will be able to interpret the locality more accurately in the case of difficult handwritten labels. If a commercial party is preferred, it may be an option to look for several parties from different language regions, for labels in those languages.

There are a few drawbacks to this manner of transcription. First the language of the label or general region needs to be known. This would mean that, for example, the country should be entered beforehand or it may be possible to employ optical character recognition (OCR) on the labels or use software for language recognition, depending on the type of writing (see task 4.1). A second issue would be the training of transcribers. If native speakers do the transcription, it may mean training needs to be available in several languages as well. Training several parties instead of one is not very time effective, so the institute outsourcing the transcription must consider carefully whether the results make up for the extra time invested.

Cost is yet another issue. Many European commercial parties outsource transcription work to a country with lower labour costs, which would counteract the advantage of language. It may be difficult to find people who speak the language in other countries, but asking people earning European wages to do the transcription would significantly raise the overall cost for the work. If the costs are an issue and time is not, crowdsourcing, whereby transcription is done by volunteers, may be a solution. For this, the cost mainly lies in organizing and setting up the correct software for the project. Time is also required on the part of the institute's staff to answer questions from the transcribers, but the transcription itself is free and can still be of high quality. It is difficult to say how long a full project would take to complete, as it has not been tried for large natural history collections yet. However, smaller projects like the project of Naturalis about glass preparations in 2013, have been a success.

5.5. Recommendations

When thinking about running a large-scale transcription project, it would be wise to consider the overall costs both in monetary terms and in terms of time and work before deciding how to execute it. One option is to undertake the project in-house, either by assigning the work to the collection staff or by hiring additional personnel. However, though collections staff are probably well trained in

Funded by the Horizon 2020 Framework of the European Union
H2020-INFRADEV-2016-2017
Grant Agreement No 777483

interpreting labels, they are often not very efficient transcribers, probably taking a lot of time and making relatively many typing errors. In addition, this takes a lot of time and effort away from other duties for which the staff were originally hired.

Outsourcing the work to a commercial party is another option. There are three possibilities:

1. Using an electronic marketplace would mean the organization of the project and the training of the transcribers is still completely in the hands of the institute's staff. Likewise, it still requires a lot of work on their part to make the project work. Furthermore, training may be complicated in case complete labels need transcription at once, possibly decreasing quality.
2. The same is the case for crowd sourcing and letting volunteers do the actual transcription. Though the overall cost may be low, lots of time is still required from the organization including for answering of all kinds of questions from the volunteers. Furthermore, the duration of the project can be very uncertain as it depends on the enthusiasm of the volunteers. Also, the same issues for training and quality arise as with an electronic marketplace, (for a discussion on quality from crowdsourcing platforms, see also chapter 3).
3. Hiring a commercial party to organize the transcription may at first glance result in the highest overall cost, but much work from organizing the transcription itself to training the transcribers would be done by that party and the planning would be much more secure. Chapter 3 also discusses the quality of Alembo's work.

The commercial transcribers are trained to work precisely and fast. And, though a party that never did any transcription for a natural history collection will need more time in terms of training from the institute's staff than a party that has executed such projects before, after the start-up period they will have learned to interpret the full label correctly, thus furthering a good transcription quality. The cost, both for the complete project and per record, mainly depends on the number and type of data to be transcribed and on the total number of records. With more records, the overall cost will obviously increase, but the price per record may decrease. Similarly, the more fields or types of data that need to be entered per label, the more the price for the record goes up, but the average price per field drops. Therefore, the final cost may be lower than expected.

But how to make sure transcription quality is good when outsourcing the work? First, a clear transcription protocol or other medium in which the rules are explained is needed. In case of outsourcing to a commercial partner, this partner can help achieving a protocol that is clear and not too complicated, and at the same time meeting the requirements of the institute. When using an electronic marketplace or a crowd sourcing platform this may be more complicated. In such a case a forum of some sort in which the transcribers can ask questions may work. At any case, we recommend taking time in abundance to make sure the rules are clear to the transcribers, for example by performing one or more tests to check their understanding of the rules.

Secondly, it is recommended to check transcription deliveries as soon as possible after delivery. Some questions and issues only arise after transcription has started and some of them only because the institute's staff finds some exception to be incorrect, though it is transcribed correctly when following the protocol. It may even be needed to make some last minor adjustments to the rules. To make sure transcription can be checked soon, we recommend starting slowly, only speeding up when all parties are quite sure the protocol is clear to everyone and interpreted in the same manner. In this phase it is still wise to check all delivered transcription as soon as possible, enabling quick feedback, thus

Funded by the Horizon 2020 Framework of the European Union
H2020-INFRADEV-2016-2017
Grant Agreement No 777483

ensuring a steep learning curve with the transcribers and preventing them from learning incorrect rules.

As soon as quality is good, the checks can be loosened up a bit, allowing for smaller samples and a longer delay for the check being done. A bit of checking remains recommended. With human transcribers, after a while some small errors tend to enter the transcription process. If left unchecked, those may grow to decrease quality significantly. Regular quality checks and feedback on errors found even if the checked sample in general can be accepted, should prevent quality drops from happening.

Besides cost and effort considerations about the transcription itself, it is a good idea to think about how to implement the workflows and which software to use. Natural history institutes often do not have the capacity to develop software and to execute the large IT-assignments needed to install the workflows and develop the software needed for such an endeavour. Due to their complexity, large database systems used for saving all the data are often not suitable for large scale transcription projects and licences may make use by an external party impossible. Even if the work is outsourced, the institute's IT-department will be stretched to the limit, especially if the project is combined with a digitization project. And the institute's staff will have plenty of work to do, both in preparation of the project and quality control, and for importing to and cleaning up the database.

6. Georeferencing

Georeferencing natural history collection specimens significantly enhances their value; allows for more comprehensive spatial and quantitative analyses and opens up their use in many more research studies (Guralnick *et al.,* 2006; Ellwood *et al.*, 2016). Many specimen labels created prior to the 1990s do not contain details of the specimen coordinates (Beaman and Conn, 2003). Therefore, to determine the locality where the specimen was collected it is necessary to infer this from the locality details on the label. Often the text information can be vague, use historical locality names or often lack any locality detail at all. Georeferencing each specimen on a per specimen basis is extremely time-consuming and resource-intensive and not a simple task. In a survey of fitness for use of GBIF data for species distribution modelling 78% of respondents noted issues with geoferencing (Anderson *et al.*, 2016). It is therefore necessary to determine the most efficient means of achieving accurate georeferencing to increase further the value and use of these specimens, whether through automated and/or manual methods.

Work has been completed on developing a practical georeferencing protocol (Chapman & Wieczorek, 2006) and research carried out on using automation to assist in some of these steps (Guralnick, *et al.* 2006).

For this task we tried to review what batch-automated georeferencing tools/packages might be available online for collection managers or transcribers to use without needing a very high level of programming skills, but many were unavailable to test

1. R biogeo package (Robertson *et al.*, 2016) - Not tested (Unavailable)
   - Developed with the primary aim of data cleaning and data quality assessment.
   - Designed to work with datasets containing point records in a range of coordinate formats.

Funded by the Horizon 2020 Framework of the European Union
H2020-INFRADEV-2016-2017
Grant Agreement No 777483

- A key feature is being able to identify likely alternative locations for points that are obviously erroneous.
- Another highlighted function is to convert coordinates that are in text format into degrees, minutes, seconds and then decimal degrees.
- One stated function – fromGEarth – obtains coordinates from Google Earth for localities without coordinates. This function is listed here: https://doi.org/10.1111/ecog.02118.

2. R GeoNames package (Rowlingson, 2016) - Not tested (Unavailable)
   - Is a series of code functions found in the GeoNames handbook to be able to query geographic global data from the Geonames gazetteer database. (https://cran.r-project.org/web/packages/geonames/geonames.pdf). The functions in this package are mostly thin wrappers to the API calls documented at the geonames web services overview.
   - Although primarily a geographical database that lets users extract information such as weather, time-zone and postcodes, it can be used to georeference from a bank of eight million place names.
   - Able to use functions such as GNcountryCode to input north, south, east and west text values to find places a certain distance in a given direction from the locality.
   - Although it is a promising package, we were not able to install it to R 3.3.2 from either the CRAN website or GitHub. The direct use of the API was not explored.

3. R dismo package (Hijmans *et al.*, 2017) - Tested
   - Developed for species distribution modelling.
   - One function – geocode – was used to provide coordinates based on a locality description, but this required a relatively precise location. It would be unsuitable for vague locality descriptions.

4. BioGeomancer (Guralnick *et al.*, 2006) - Not tested (Unavailable). Portions of the website not available including the Workbench. This resource is reliant on further funding to make it accessible.
   - Geo-referencing toolkit that is capable of performing numerous functions.
   - Increases geo-referencing rate by focusing on automated methods and batch processing for textual parsing; estimates uncertainty associated with records' coordinates; tests for errors and assures consistency; defines and applies documented data standards; provides information about data processing steps so that results can be tested, replicated and improved.

5. Geo-referencing Calculator - Not tested
   - Does not determine a coordinate from its locality text description but is designed to calculate all the factors that contribute to the uncertainty in a georeference and adds them up. Once the coordinates have been determined it is possible to follow the steps in the calculator to factor in uncertainty from other factors.
   - Designed for the Mammal Networked Information System (MaNIS) project and has been adopted as well by HerpNet and ORNIS. The application makes calculations using the methods described in the Georeferencing Guidelines. The guidelines include examples of maximum error calculations.

Funded by the Horizon 2020 Framework of the European Union
H2020-INFRADEV-2016-2017
Grant Agreement No 777483

6. The Edinburgh Geoparser - Not tested (Unavailable to install software)
   - A system that is able to automatically recognise place names, within a text file or string, which can disambiguate them with respect to a gazetteer.
   - Can be used with gazetteers such as Unlock and GeoNames.
   - Has been used for a variety of text inputs and can display outputs with a Google Maps visualisation.
   - Can be installed only on Mac or Linux systems. Having access to a Mac, we tried to download it. However it requires knowledge of Bash coding language to be able to code into the terminal. There are risks to this as the terminal can change internal settings of the Mac and so this method was not pursued.

7. GEOLocate - Tested batch client service
   - A user-friendly web-based platform to georeference from a text string. The program will split the text string into country, county, locality, etc.
   - It is possible to type, cut and paste a single locality string, or upload a CSV file and batch process it.
   - Returns an output of latitude and longitude in decimal degrees, with an accuracy value in meters.
   - Outputs a number of points that need to be reviewed to choose the one that is the most likely.

We completed a small test using locality strings from 20 Kew specimens which had already been georeferenced by Kew staff trained in georeferencing. We used specimens from a variety of countries (India, Morocco, Bolivia, Australia, Papua New Guinea, Brazil, Indonesia, Georgia, Kenya, Cameroon, South Africa, Malaysia and Belize). The test platform often gave more than 10 points to choose from and an accuracy of up to 3608 km, but results varied widely. Results obtained might be better for certain countries, particularly the United States., GEOlocate has been integrated into several collection management systems such as Arctos (Sikes *et al*., 2017), Specify, Symbiota (Denslow *et al*., 2016, Emu 6) and Tropicos. This functionality was not tested.

8. SpeciesgeocodeR (Zizka & Antonelli 2015) -not tested used for cleaning coordinate date not determniing cordinates.

   - An R package for automatically cleaning, processing and analysing species occurrence data and to code them into discrete units.
   - The *GeoClean* function offers an automated and reproducible flagging of potentially problematic records. The function includes basic tests for coordinate validity (e.g. invalid coordinates, zero coordinates, equal latitude and longitude) and more complex tests accounting for common problems in large datasets (e.g. if occurrences fall within the right country borders, if occurrences have been assigned to the country centroid or country capital or to the GBIF headquarters

9. CoordinateCleaner (Zizka et. al, 2019) - Tool for speeding up the identification of problematic records and common problems in a data set for further verification . not tested –as paper only published 20th January 2019

   - Compares the coordinates of occurrence records to reference databases of country and province centroids, country capitals, urban areas, known natural ranges and tests for plain

Funded by the Horizon 2020 Framework of the European Union
H2020-INFRADEV-2016-2017
Grant Agreement No 777483

zeros, equal longitude/latitude, coordinates at sea, country borders and outliers in collection year. Flags potential problematic coordinate records.

- Compares against a global database of 9,691 geo-referenced biodiversity institutions to identify records that are likely escaped horticulture from the institution or specimens that have been geo-referenced to their physical location
- Algorithms to identify coordinate conversion errors and occurrence records derived from rasterized collection designs or subject to strong decimal rounding (e.g. presence/absence in 100x100km Grid squares.
- Spatio-temporal tests for fossil data. - detects problems with inaccurate or overly imprecise temporal information.

Zizka et al. tested CoordinateCleaner on occurrence records for flowering plants on GBIF and the Paleobiology Database (PBDB). The clean_coordinates wrapper function flagged 3.6% of the records in GBIF and 6.3% in PBDB for further verification.

Our survey of the existing automated tools for georeferencing suggests that complete automation is currently not possible but there are some tools/methods available that speed up the process to reduce this highly labour-intensive task. To utilise many of these tools it is necessary to have good data management and basic programming skills for example be familiar with downloading software from GitHub, use of R or be comfortable using API's. Providing good manuals and training on using these tools will be necessary to ensure the tools available are utilised more widely. However, some of the methods to speed up georeferencing may be relatively simple; such as sorting specimens by same location so that all the collections from the same collection site are grouped and the coordinates inferred for the whole group at the same time (Magdalena *et al.*, 2018; Paterson *et al.*, 2016; Hill *et al*. 2009). This method becomes more effective as the number of specimen records grows, as you expect more site duplication. Other sorting methods to speed up georeferencing include sorting specimens by collector, collector number and collection date to enable the georeferencing of a collector's trip. This is enhanced if you have access to field notebooks or collectors' itineraries. Mapping specimen points and plotting a collector's route could highlight outliers and likely errors. Again, as more and more natural history specimen collections are digitised, this method becomes more effective as more records from any particular trip are in electronic format. Sorting by collector will also lead to sorting by language, as all labels by one collector will normally be in the same language, so specimens can be sent for georeferencing or transcription to a georeferencer or transcriber proficient in that language.

Data aggregators could assist by creating groups of un-georeferenced records and enable their distribution to expert georeferencers, as well as facilitating feedback on the georeferenced coordinates to the original data providers (Anderson, 2016). However, this could be problematic because many records may not have coordinates due to lack of suitable locality information on the label. To avoid duplication of effort records could be filtered for presence of a locality string and records already found to be unsuitable for georeferencing should be flagged as such. Ideally there should be automated locality webservices where record localities could be checked against already georeferenced records to see if a location has been already georeferenced. The DiSSCO research infrastructure could consider compiling its own gazetteer. Trained georeferencers could assess whether a point or points found are correct and assign uncertainty when none is provided. Storing notes on how a geoference and uncertainty was determined could also avoid duplication of effort.

Funded by the Horizon 2020 Framework of the European Union
H2020-INFRADEV-2016-2017
Grant Agreement No 777483

GEOLocate has developed mechanisms for communities to collaboratively georeference and verify a shared dataset http://www.geo-locate.org/: The GEOLocate web-based collaborative client for reviewing and editing community records; and the Web-based data management portal for creating and managing communities. Users can create a community, upload records for georeferencing and allocate different records to different individuals. The data upload process detects potential duplicates so they can be georeferenced together. Efficiency of the collaborative tool over traditional methods was tested using the Tulane University Museum of Natural History fish collection. Out of 2100 records, 30% were identified as being similar to other records and an additional 33% were duplicate records leaving a total of 782 unique locations requiring correction and a 63% reduction in overall effort. An example protocol for using the collaborative tools is available (Biedron & Famoso 2016).

Van Erp *et al*., (2015) describe a knowledge-driven approach for automated georeferencing demonstrating a case study using database records for reptiles and amphibians. The approach consists of a pipeline with five rule-based modules which are automated and implemented in the programming language Perl. The results are then presented to a researcher to check before they are accepted. The method produces a confidence score to highlight to curators the records which need to be checked, most requiring human input. The modules include filtering the database records on relevant fields, text parsing, gazetteer lookup, offset calculation and disambiguation heuristics (used to distinguish between place-names that share the same name). The disambiguation heuristics module includes: Spatial Minimality - looks at text which mentions more than one locality and chooses the place name based on the place names that that are most cluster together; Expedition Clusters - groups records by collection dates and country and gives the place name that is close to the previous georeferenced location record a higher confidence measure; Species Occurrence Data - queries the occurrence records of the species in GBIF and retrieves the coordinates.

Luomus, the Finnish Natural History Museum, georeferences Finnish specimens from its insect digitisation line. The approach is semi-automated; the data are first homogenized by simple sorting & editing in a spreadsheet. A script, written in Python, is then run on the file(s), comparing the input data with a list of transformations (unique location description combinations -> autocompleted location data + coordinates + coord_system + c_error if applicable, etc.) which yields a partially georeferenced file. Any missed or non-unique localities are then reviewed and georeferenced (if possible) manually. Localities that are unique and likely to be repeated are added to the transformation source file. A major downside is that the method is only accessible by one person but the use of a handmade reference list gives complete control over the result; we can take into account oddities that most automated georeferencing systems would likely miss. For example, there were once three municipalities called Pyhäjärvi in Finland but just one carries that name today. One of them has been Russian territory since 1944 and, as such, easily missed by an indexer looking for Pyhäjärvi and Finland in modern databases of locality data (Jeres, pers. comm., 2018).

With many specimens to georeference, crowdsourcing this task could be an option. Whilst DigiVol provides some georeferencing facilities, many expeditions within DigiVol do not enable this option, perhaps because they consider the task too complex with the limited inbuilt tool provided. Ellwood *et al.*, 2016 compared undergraduate students with experts i.e. trained experienced technicians or local botanists on georeferencing using GEOLocate for fish and plant specimens. After outliers were removed, the range between student and expert georeferenced points was <1.0 to ca. 40.0 km for both the fish and the plant experiments, with an overall mean of 8.3 km and 4.4 km, respectively.

Funded by the Horizon 2020 Framework of the European Union
H2020-INFRADEV-2016-2017
Grant Agreement No 777483

Engaging students in the process improved results beyond GEOLocate's automatic algorithm alone. Calculation of a median point from replicate points improved results further, as did recognition of good georeferencers (e.g., creation of median points contributed by the best 50% of contributors). They call for the creation of an online citizen science georeferencing platform that could see improvements to these results with enhanced georeferencing tools.

Users of georeferenced specimen data need to assess the quality of the data, often downloaded from aggregator sites before they use it, this can be quite a lengthy process especially as the data comes from many different collections with varying degrees of data quality checks before being released. The amount of checks that a researcher will do will depend on the scientific use of the data and time constraints. The amount of time required would be reduced if records were flagged to be of suitable quality to be used in a particular DQIG use case. In addition, notes on how a specimen was georeferenced are invaluable and speed up any data quality checking or reduce duplication of effort, producing such notes should be included in any georeferencing task. It is vital for the user of the data to understand the accuracy of any georeferenced record; different precision will be needed for different use cases and the research being carried out. However, if the accuracy of the georeferenced points is clearly known by the data user then it will be easier for the user to decide whether to include the point in their research dataset. DiSSCo should display the accuracy of the georeferenced record and categories for accuracy should be clearly linked to the DQIG use cases.

A Georeferencing For Research Use (GRU) Workshop was run by Integrated Digitized Biocollections (iDigBio), UC Santa Barbara Cheadle Center for Biodiversity and Ecological Restoration (CCBER), VertNet, Denver Botanic Gardens, Yale Peabody Museum, Stanford Earth and GEOLocate which resulted in annotated set of data quality checks that participants report they use when evaluating and cleaning datasets for research use (Seltman et al., 2018). Participants included both data researchers and data providers.

Many of these issues found could be resolved by ensuring that the collection management system used for transcribing or georeferencing does not allow data to be captured with these issues in the first place (e.g. ensuring 1-12 values are the only acceptable values for month, and latitude values from 0 to 90). Another issue reported was fake precision giving a false representation of the accuracy of the georeferenced provided. This can occur for many reasons but can include the automatic conversion of data entered in Degrees, Minutes and Seconds as shown on a specimen label to Decimal Degrees within a collection management system or the automatic addition of too many decimal places. The number of decimal places shown should relate to the accuracy value of the record. Dialogue with collection management system companies is recommended to ensure validation and data quality checks are incorporated into the systems where appropriate.

7. Recommendations for DiSSCo

The DiSSCo infrastructure should implement the data quality tests as defined by the DQIG and that are currently being standardised and implemented by aggregator sites such as GBIF, ALA and iDigBio. DiSSCo should liaise with data providers to get feedback on how the results can be easily understood and investigate the barriers for data providers to use the results of the tests to improve the data

Funded by the Horizon 2020 Framework of the European Union
H2020-INFRADEV-2016-2017
Grant Agreement No 777483

quality of the records. It could be beneficial to provide a pre-ingestion tool so data providers can be alerted to the issues and problems with their dataset before it is ingested in the system.

DiSSCo should indicate which datasets/records are suitable for which uses, user stories are currently being gathered by the DQIG and fitness for use could be linked to these stories.

It would be beneficial for records in DiSSCo to be clearly marked to indicate the completeness of the transcription following the Minimal Information Standard for Digital Specimens (MIDS) that is currently being developed.

DiSSco could also consider providing tools to help with the transcription itself. It is clear that data quality is improved with the use of lookup lists but many collection managers do not have easy access to these lists. DiSSCo infrastructure could consider signposting useful resources to data managers or even provide lists itself, e.g. a download of its taxonomic backbone, a list of collectors with information on where the collector collected and over what dates, a list of country names, or old country names with its mapping to new country name(s). These lists would take effort and resources to maintain but as an aggregator of data from European natural history collections it is well placed to manage these lists or work with other aggregators.

DiSSCo could also investigate what other tools would be useful. For example, providing handwriting samples of collectors and determiners, use of social media to encourage communication and knowledge sharing to assist with transcription a good example of this is the twitter account EntomologyTranslator@EntoTranslator, tools for suggesting georeferences based on locality information and georeferences of records already in the DiSSCO infrastructure. Specifically in relation to herbarium specimens it could try to match duplicates and highlight inconsistencies or differences in the data and enable propagation of annotations, georeferences or type status between duplicates.

DiSSCo could offer a Taxon names reconciliation service allowing data providers to checking against the taxonomic backbone before ingestion. Users could then feedback any omissions or errors in the names to help with keeping the backbone up to date.

As an aggregator site, with an ambition to hold more than a billion specimen records from all over Europe, DiSSCo is in a unique position to process or visualise data to highlight possible errors. For example, it could allow data managers to map a collecting trip and visualise outliers or automatically flag up possible errors. For example, if a specimen is collected outside a collector's known collection date range or country range it could be flagged for investigation. If a taxon has been found outside its currently known expected country or locality range it could be flagged up for checking.

Finally, DiSSCo could work with data managers and collection management system providers to find out where the bottlenecks are for implanting data quality tests, implementing results for these tests and creating smooth feedback mechanisms from users of DiSSCo to data providers, back to collection management systems.

Funded by the Horizon 2020 Framework of the European Union
H2020-INFRADEV-2016-2017
Grant Agreement No 777483

8. References

Anderson, R.P., Araújo, M.B., Guisan, A., Lobo, J.M., Martínez-Meyer, E., Peterson, A.T., & Soberón, J. (2016) Are species occurrence data in global online repositories fit for modeling species distributions? The case of the Global Biodiversity Information Facility (GBIF). Final report of the task group on GBIF data fitness for use in distribution modelling. 1–27.

Andersen, D., & Lau, R. (2018) Pay Rates and Subject Performance in Social Science Experiments Using Crowdsourced Online Samples. Journal of Experimental Political Science, 5(3), 217-229. doi:10.1017/XPS.2018.7

Beaman RS, Conn BJ. (2003) Automated geoparsing and georeferencing of Malesian collection locality data. Telopea 10 (1): 43-52. doi.10.7751/telopea20035604.

Berinsky, A., Huber, G., & Lenz, G. (2012) Evaluating Online Labor Markets for Experimental Research: Amazon.com's Mechanical Turk. Political Analysis, 20(3), 351-368. doi:10.1093/pan/mpr057

Biedron, E.M. & Famoso, N.A. (2016) Using GEOLocate for Collaborative Georeferencing. https://epicc.berkeley.edu/wp-content/uploads/2015/11/UsingGeoLocateforCollaborativeGeoreferencing_2016.pdf

Buhrmester, M., Kwang, T., & Gosling, S.D. (2011) Amazon's Mechanical Turk: A New Source of Inexpensive, Yet High-Quality, Data? Perspectives on Psychological Science, *6*(1), 3–5. https://doi.org/10.1177/1745691610393980

Buhrmester, M. D., Talaifar, S., & Gosling, S.D. (2018) An Evaluation of Amazon's Mechanical Turk, Its Rapid Rise, and Its Effective Use. Perspectives on Psychological Science, *13*(2), 149–154. https://doi.org/10.1177/1745691617706516

Cai, L. and Zhu, Y., 2015. The Challenges of Data Quality and Data Quality Assessment in the Big Data Era. *Data Science Journal*, 14, p.2. DOI:http://doi.org/10.5334/dsj-2015-002

Chandler, Jesse & Shapiro, Danielle. (2016) Conducting Clinical Research Using Crowdsourced Convenience Samples. *Annual Review of Clinical Psychology*. 12. https://doi.org/10.1146/annurev-clinpsy-021815-093623

Chapman, A.D. and Wieczorek, J. (2006) Guide to Best Practices for Geo-referencing [Online]. Available from: http://www.gbif.org/resource/80536 [Accessed 18 November 2018].

Denslow, M.W., Brown, H., Gilbert, Rios, N. & Murrell, Z.E. (2016) SERNEC collaborative georeferencing; leveraging the interoperability between GEOLocate and Symbiota for a large scale digitization project. TDWG 2016 Annual Conference https://mbgocs.mobot.org/index.php/tdwg/tdwg2016/paper/view/1096

Dillen, M., Groom, Q., Chagnoux, S., Güntsch, A., Hardisty, A., Haston, E., Livermore, L., Runnel, V., Schulman, L., Willemse, L., Wu, Z. & Phillips, S. (2019) A benchmark dataset of herbarium specimen images with label data. Biodiversity Data Journal. In press.

Funded by the Horizon 2020 Framework of the European Union
H2020-INFRADEV-2016-2017
Grant Agreement No 777483

Ellwood, E.R., Bart, Jr., H.L., Doosey, M.H., Jue, D.K., Mann, J.G., Nelson, G., Rios, N. and Mast, A.R., 2016. Mapping Life – Quality Assessment of Novice vs. Expert Georeferencers. Citizen Science: Theory and Practice, 1(1), p.4. DOI: http://doi.org/10.5334/cstp.30

Evanini, K., Higgins, D., & Zechner, K. (2010) Using Amazon mechanical turk for transcription of nonnative speech. In Proceedings of NAACL HLT 2010 workshop on creating speech and language data with Amazon's mechanical turk, Los Angeles, USA (pp. 53–56).

Gleibs, I.H. (2017) Behav Res 49: 1333. https://doi.org/10.3758/s13428-016-0789-y

Guralnick, R.P., Wieczorek, J., Beaman, .R, Hijmans, R.J., the BioGeomancer Working Group (2006) BioGeomancer: Automated Georeferencing to Map the World's Biodiversity Data. PLoS Biol 4(11): e381. https://doi.org/10.1371/journal.pbio.0040381

Hijmans, R.J., Phillips, S., Leathwick, J. and Elith, J. (2017) Package 'dismo', CRAN, https://cran.r-project.org/web/packages/dismo/dismo.pdf.

Hill, A.W., Guralnick, R., Flemons, P., Beaman, R., Wieczorek, J., Ranipeta, A., Chavan, V., and Remsen, D., (2009) Location, Location, Location: Utilizing Pipelines and Services to More Effectively Georeference the World's Biodiversity Data. BMC Bioinformatics, vol. 10, no. 14, pp. 1–9

James, S.A., P. S. Soltis, L. Belbin, A. D. Chapman, G.Nelson, D. L. Paul, and Collins, M., (2018) Herbarium data: Global biodiversity and societal botanical needs for novel research. Applications in Plant Sciences 6(2): e1024. doi:10.1002/aps3.1024

Kirilenko, A. P., & Stepchenkova, S. (2016). Inter-Coder Agreement in One-to-Many Classification: Fuzzy Kappa. *PloS one*, *11*(3), e0149787. doi:10.1371/journal.pone.0149787

Krippendorff, K. (2011). Computing Krippendorff's Alpha-Reliability. Retrieved from http://repository.upenn.edu/asc_papers/4

McHugh M. L. (2012). Interrater reliability: the kappa statistic. *Biochemia medica*, *22*(3), 276-82.

Magdalena, U.R., Silva L.A.E., Lima, R.O., Bellon, E., Ribeiro, R., Oliveira, F.A., Siqueira, M.F., Forzza, R.C. (2018) A new methodology for the retrieval and evaluation of geographic coordinates within databases of scientific plant collections. Applied Geography. 96. https://doi.org/10.1016/j.apgeog.2018.05.002.

Mankar, Akash & J. Shah, Riddhi & Lease, Matthew. (2017) Design Activism for Minimum Wage Crowd Work. arXiv:1706.10097v3

Owen D., Groom, Q., Leegwater, T., van Walsum, M. Wijkamp, N. 2019. D.4.1 Methods for automated text digitisation.

Paterson, G., Albuquerque, S., Blagoderov, V., Brooks, S., Cafferty, S., Cane, E., Honey, M., Huertas, B., Howard, T., Huxley, R., Kitching, I., Ledger, S., McLaughlin, C., Martin, G., Mazzetta, G., Penn, M., Perera, J., Sadka, M., Scialabba, E., Self, A., Siebert, D., Sleep, C., Toloni, F. and Wing, P., 2016. iCollections – Digitising the British and Irish butterflies in the Natural History Museum, London. Biodiversity Data Journal, pp.1–16

Funded by the Horizon 2020 Framework of the European Union
H2020-INFRADEV-2016-2017
Grant Agreement No 777483

Paolacci, G., & Chandler, J. (2014) Inside the Turk: Understanding Mechanical Turk as a Participant Pool. Current Directions in Psychological Science, 23(3), 184–188. https://doi.org/10.1177/0963721414531598

Peer, E., Samat, S., Brandimarte, L., Acquisti, A. (2015) Beyond the Turk: An empirical comparison of alternative platforms for crowdsourcing online research. NA-Advances in Consumer Research, 43, 18-22.

Pittman, R. & Sheehan, K. (2016) Amazon's Mechanical Turk a Digital Sweatshop? Transparency and Accountability in Crowdsourced Online Research. Journal of MediaEthics, 31:4, 260-262, DOI: 10.1080/23736992.2016.1228811

Rainer, H., Vitek, E. (2009) Virtual herbaria - an open platform to join. In: Stevanović V(Ed.) Book of Abstracts, 5th Balkan Botanical Congress 2009

Robertson, M.P., V. Visser, and C. Hui. (2016) Biogeo: An R package for assessing and improving data quality of occurrence record datasets. Ecography 39:394–401. https://doi.org/10.1111/ecog.02118

Robertson, B. (2016) Package 'biogeo', CRAN:https://cran.rproject.org/web/packages/biogeo/biogeo.pdf.

Rowlingson, B. (2011) Package 'geonames', CRAN, www.geonames.org

Seltmann K, Lafia S, Paul D, James S, Bloom D, Rios N, Ellis S, Farrell U, Utrup J, Yost M, Davis E, Emery R, Motz G, Kimmig J, Shirey V, Sandall E, Park D, Tyrrell C, Thackurdeen R, Collins M, O'Leary V, Prestridge H, Evelyn C, Nyberg B (2018) *Georeferencing for Research Use* (GRU): An integrated geospatial training paradigm for biocollections researchers and data providers. Research Ideas and Outcomes 4: e32449. https://doi.org/10.3897/rio.4.e32449

Sheehan, K.B. (2018) Crowdsourcing research: Data collection with Amazon's Mechanical Turk, Communication Monographs, 85:1, 140-156, DOI: 10.1080/03637751.2017.1342043

Sidi, F.; Shariat Panahy, P.H.; Affendey, L.S.; Jabar, M.A.; Ibrahim, H. and Mustapha, A. "Data quality: A survey of data quality dimensions," *2012 International Conference on Information Retrieval & Knowledge Management*, Kuala Lumpur, 2012, pp. 300-304. doi: 10.1109/InfRKM.2012.6204995

Sikes, Bowser M, Daly K, Høye TT, Meierotto S, Mullen L, Slowik J, Stockbridge J. 2017 The value of museums in the production, sharing and use of entomological data to document hyperdiversity of the changing North. Arctic Science 33, 498– 514. (doi:10.1139/as-2016-0038)

van Erp, Marieke & Hensel, Robert & Ceolin, Davide & Meij, Marian. (2014). Georeferencing Animal Specimen Datasets. Transactions in GIS. 19. 10.1111/tgis.12110.

Veiga, A.K., Saraiva, A.M., Chapman, A.D., Morris, P.J., Gendreau, C., Schigel, D., et al. (2017) A conceptual framework for quality assessment and management of biodiversity data. PLoS ONE 12(6): e0178731. https://doi.org/10.1371/journal.pone.0178731

Willis, C.G., Law, E. , Williams, A.C., Franzone, B.F., Bernardos, R. , Bruno, L. , Hopkins, C., Schorn, C., Weber, E., Park, D.S. and Davis, C.C. (2017) *CrowdCurio*: an online crowdsourcing platform to facilitate climate change studies using herbarium specimens. New Phytol, 215: 479-488. doi:10.1111/nph.14535

Zizka A & Antonelli A (2015) speciesgeocodeR: An R package for linking species occurrences, user-defined regions and phylogenetic trees for biogeography, ecology and evolution. https://doi.org/10.1101/032755

Zizka, A. , Silvestro, D. , Andermann, T. , Azevedo, J. , Duarte Ritter, C. , Edler, D. , Farooq, H. , Herdean, A. , Ariza, M. , Scharn, R. , Svanteson, S. , Wengström, N. , Zizka, V. and Antonelli, A. (2019), CoordinateCleaner: standardized cleaning of occurrence records from biological collection databases. Methods Ecol Evol. Accepted Author Manuscript. doi:10.1111/2041-210X.13152

9. Acknowledgements

Many thanks to all that contributed to this report including:

All those institutions that provided images for our transcription pilot.

- P/PC - Muséum national d'Histoire naturelle (Paris - France)
- BR - Plantentuin Meise, Belgium
- B - Botanischer Garten und Botanisches Museum Berlin-Dahlem (Berlin, Germany)
- BM - The Natural History Museum (London, UK)
- K - Royal Botanic Gardens Kew, UK
- E - Royal Botanic Garden Edinburgh, UK
- TU - Tartu Ülikool, Estonia.

All those that assisted us putting the projects up on the crowdsourcing platforms including: Paul Flemons, Michael Denslow, Agnes Kirchhoff, Anton Güntsch, Dominik Röpert, Simon Chagneaux, Gwenaël Le Bras and Nuno Veríssimo Pereira.

Louise Allen, Margaret Gold, Laurence Livermore and Natalie Dale Sky for the Miniature Lives Magnified: Data Quality Report on the Chalcid microscope slide expedition.

Nicholas Wells and Maya Master for their contribution on the overview of Geo-Referencing tools.

Brendan Cordy for his useful advice on statistics.

All those that commented on the report at different stages: Agnes Wijers, Deborah Paul, Niels Raes and Luc Willemse.