Anne Morel

Herbarium of the University of Bern

anne.morel@unibe.ch

**Data Science Project**

Text extraction from an herbarium label

Conceptual Design Report

30 November 2023

# Abstract

Herbarium preserve biodiversity information dating up to 5 centuries. An herbarium sheet consists of an acid-free paper sheet on which a dried pressed plant specimen has been fixed. An affixed label with the most important information concerning the plant is often handwritten next to it: scientific name, location of collect, date of collect, name of the collector. The design developed here aims to read the information present on an herbarium label and paste them in the correct excel field, in order to facilitate the label deciphering and optimize the digitization process of natural history collections.

# Table of Contents

[Abstract 0](#_Toc151136446)

[Table of Contents 1](#_Toc151136447)

[1 Project Objectives 2](#_Toc151136448)

[2 Methods 3](#_Toc151136449)

[3 Data 3](#_Toc151136450)

[4 Metadata 5](#_Toc151136451)

[5 Data Quality 6](#_Toc151136452)

[6 Data Flow 6](#_Toc151136453)

[7 Data Model 8](#_Toc151136454)

[9 Risks 8](#_Toc151136455)

[10 Preliminary Studies 9](#_Toc151136456)

[11 Conclusions 9](#_Toc151136457)

[Acknowledgements 9](#_Toc151136458)

[Statement 10](#_Toc151136459)

[Appendix 2: Example of an herbarium sheet (Sheet-0043911) 12](#_Toc151136460)

[References and Bibliography 13](#_Toc151136461)

# 1 Project Objectives

0.5-1.0 page ; Formulate the goal and purpose of your project here. It is very important to be very specific on your goals. What do you want to find out or predict? What are the numbers and plots you need to do that?

In the process of herbarium digitization, the longest step is the data capture of the information present on the label. Herbarium sheet labels may have standard information (scientific name of the plant, collecting place, collecting date and collector name), but the format of the labels vary greatly among the various sheets present in a collection and often prepared by different people: typed labels, handwritten labels, handwritten notes directly on the sheet, nicely handwritten labels and badly handwritten ones, detailed labels and incomplete ones, etc.

This project aims to facilitate the data capture by reading the written information on a label and copying it in an excel file that can easily be imported into the herbarium database. The ultimate goal of this process would be to read the most difficult handwritten labels (e.g. the Sütterlin handwriting). To achieve that, the workflow has been divided into several steps.

* 1. On a specific collection (code written and working)
     1. Perform the text reading code on standardized typewritten labels. Extract and save a specific information of the label (e.g. the catalog number) in an excel file, joined with the name of the picture file (corresponding to a data matrix encoded unique number).
  2. On the main collection (code to be written)
     1. Categorize the label between typewritten and handwritten for the whole collection.
        1. On the typewritten label of the main collection
           1. Perform the text reading code

Extract all the information in a single excel cell (corresponding to the verbatim label information field in the database)

Separate all the information in multiple cells depending on their category (scientific name, date of collect, collector, family, location of collect, …)

Compare the read information with datasets containing the categories and their possible values

* + - 1. On the handwritten label of the main collection
         1. Categorize the label between clear handwriting (similar to typed) and unclear handwriting

Extract all the information in a single excel cell (corresponding to the verbatim label information field in the database)

Separate all the information in multiple cells depending on their category (scientific name, date of collect, collector, family, location of collect, …)

Compare the read information with datasets containing the categories and their possible values

# 2 Methods

0.5-1.0 page Which infrastructure, tools, software libraries, statistical methods etc do you intend to use. It is clear that you may not know all this at this stage, but try to make yourself some thoughts, even if it is going to change during the CAS.

The code is to be performed in Python with libraries able to:

* Find and open a picture, specifically in the JPG format.
  + if possible, with the possibility to choose between local files, files in a server or in an external disc.
* Categorize the elements present on the picture (plant, label(s), plain text, matrix code) (e.g. Segment Anything Model).
* Read text on the picture (e.g. *easyOCR*), if possible only in the part containing text.
  + if possible, use a supervised machine learning algorithm that would use the label handwriting already identified to perform it on undeciphered labels and improve the quality of the text extraction code (e.g. MNIST database of handwritten digits).
* Save the text read in an excel file, joined with the corresponding matrix code number.
* Do all the previous steps in series and perform them on a high number of pictures.

At the end of the process, a human verification of the data still needs to be done.

# 3 Data

Which data will be used (with references) 0.5-1.0 page ; A couple of plots, maybe some histograms of the columns ; A couple table row as an example maybe ; Security issues etc (see data management plan, you may attach a SNSF data management plan for your data)

The data used is not published yet, and is not available online. It is at the moment around 66’000 pictures in JPG format of herbarium sheets belonging to the Herbarium of the Botanical Garden of the University of Bern (official acronym: BERN), recently taken by Picturae, a company specialized in herbarium digitization, and by Anne Morel, the digitization collaborator of the BERN Herbarium.

Each picture corresponds to an herbarium sheet of the BERN herbarium. The herbarium has only recently started this digitization process. Around 500’000 herbarium sheets are assumed to constitute this collection, and the current digitization process would be its first large-scale inventory mission.

The collection in itself is not homogenous but more of a complex assemblage of professional and amateur botanists’ collections through centuries. Hence the high diversity in label writing and quality that found in it (Figures 1 to 4).

|  |  |
| --- | --- |
|  |  |
| Figure 1: Typewritten label example (Sheet-0061001). | Figure 2: Handwritten label information example, written directly on the sheet of paper (Sheet-0061002) |
|  |  |
| Figure 3: Handwritten label of a common collector in the collection, easily-read handwriting (BERN-0066800) | Figure 4: Handwritten label of a common collector in the collection, difficult-to-read handwriting (Sheet-0059007) |

The first step of the code has been performed on a portion of the pictures’ dataset, which all had a standardized typewritten label. Since the matrix code sticker corresponding to the new unique catalog number of the digitization had been placed by the Picturae company, the database of these sheets had to be manually updated with it, by finding the previous catalog number (“Beleg Nr.” on the label) and adding the matrix code number (also corresponding to the name of the jpg file) in the corresponding field of the database (Figure 5).

|  |  |
| --- | --- |
|  | In ; ; 8 Herbarium Rob. Streun; Bern &5 882] Olquimanin ounxnfaz Htle @; 6; Zeuzznzz\_ 727SZ 7 8g Jnnl 55464] gg Jecnn L 10 VIL 4q44 888 gs Dalum ; Ao 714q?4 leg : Tdezy 88 g8 g\*; ggg 8g ~ 8\*g 1 8g2 pg 8 5 0 80" 8 8 9g 68 g9g/8 Herbarium Bernense PROB 58/358 [Agrimonia procera Wallr. Synonym: A. odorata (Gonan) Miller, Wohlriechender Odermenning FamilielOrdnung: Rosaceae, Rosales LandIOrt: CH, Bern BE Flurname: Bern Datum: 10.7.1924 Koord: 601 / 198 Höhe: 540 m.üM. Atlas Nr: 251,263,311 Standort/Soziologie: leg Jdet: R. 128 verif : R. Gerber, 2002 Sheet-0043911 unweit der Halenbrücke in lichtem Gebüsch und Junwald Aus Herbar: R. Probst Beleg Nr. 3310 8 er . 1877/2002 ; @4hu7 Jwt' Streun |
| Figure 5: Sheet-0043911, blue highlight: text almost correctly read by the OCR code; orange highlight: target text to be extracted for the first step of the code. | |

The result of the text extraction is a table in an excel file. Four columns were created in the code:

* *otherCatalogNumber*, text to be extracted from the picture.
* *catalogNumber*, name of the jpg file and new unique catalog number.
* *BOXNumber*, folder name where the picture is stored in the server.
* *QualityMeasure*, evaluation result of the correctness of the text extracted.

| *lines* | **otherCatalogNumber** | **catalogNumber** | **BoxNummer** | **QualityMeasure** |
| --- | --- | --- | --- | --- |
| *1* | 3537 | Sheet-0043938 | BOX-000471 | Correct |
| *2* | 3536 | Sheet-0043939 | BOX-000471 | Correct |
| *3* | 3535 | Sheet-0043940 | BOX-000471 | Correct |
| *4* | 35; | Sheet-0043941 | BOX-000471 | To check |
| *5* | 3533 | Sheet-0043942 | BOX-000471 | Correct |
| *…* | … | … | … | … |
| *627* | 3576 | Sheet-0044569 | BOX-000478 | Correct |
| *628* | 8 g0 | Sheet-0044570 | BOX-000478 | To check |
| *629* | 3574 | Sheet-0044571 | BOX-000478 | Correct |
| *630* | 3573 | Sheet-0044572 | BOX-000478 | Correct |
| *631* | 357~ | Sheet-0044573 | BOX-000478 | To check |
| *632* | 3571 | Sheet-0044574 | BOX-000478 | Correct |
| *…* | … | … | … | … |

Table 1: Result table of the OCR reading code performed on 34’208 herbarium sheets of the BERN herbarium.

# 4 Metadata

What metadata is required for reproducing your analysis? ; Where do you store the metadata, how can people access it?

The pictures are planned to be available online in 2024. Still, the goal of this code is to be performed on any herbarium sheet picture. At the moment, a huge amount of herbarium sheet pictures is available online (e.g. <https://plants.jstor.org/> or on specific herbarium web catalogs, e.g. <https://www.ville-ge.ch/musinfo/bd/cjb/chg/?lang=en>). However, most of these pictures can only be seen one-to-one. In order to perform this code on a greater amount, the herbarium curators have to be contacted personally and asked for a deposit folder of their pictures.

Concerning the herbarium of Bern, the 66’000 pictures are available only for intern use.

# 5 Data Quality

What are the quality requirements you need in order to meet your project objectives (data size, data precision missing values, ….)? Are they met? If not, do you expect a significant impact on your result?

Any measures to improve the data quality?

To check the data quality of the code performed, a specific column (*DataQuality*) has been created in the code. The value results of testing if the *otherCatalogNumber* value is only made of 6 digits. The code has been performed on a total of 34’208 herbarium sheets with the standardized label model presented in Figure 5.

|  |  |  |
| --- | --- | --- |
| **DataQuality** | **Total count** | **Percentage** |
| Correct | 30,096 | 87,98% |
| To check | 4112 | 12,02% |
| **Total** | **34,208** | **100,00%** |

Table 2: Data quality results after running the code. The data labeled as “Correct” corresponded to an only-digit extracted value. The data labeled as “To check” corresponded to a different result, which could be either wrong or incomplete data extraction.

A human verification is still needed at the end of the process, so the quality requirements must not be 100% correct.

Concerning the next steps of the code, a verification by comparison of the previous entries is planned to be computed. For example, the scientific name of a species (Genus + specific epithet + Author) is completely standardized and would allow a 100% correctness when the algorithm has access to a nomenclatural database. The dates have various format (e.g. 11/17/23, 11/17/2023, 17 November 2023, 17 nov. 23, etc.) but the list is still short and can easily be integrated into the code basis of reference. The most difficult information’s quality to optimize is the locality of collect. Names of location have sometimes changed through time, or are locally known but not nationally. A subdivision of the label into the various parts could help reduce this imprecision by saying “What is not <the rest> is locality and has to be paste in the verbatim locality field”.

# 6 Data Flow

Explain with a figure and words how the data flow of your project will be, from the data source to the final plots and numbers.

This conceptual design is separated into three distinct steps. The first one being the extraction of a specific information of an herbarium label (Figure 6). The second one being the extraction of the whole text present on the label (Figure 7). Finally the third one being the categorization of the label information after the extraction and then separation into the corresponding fields of the result excel file (Figure 8).

Figure 6: Data flow of the first code, extraction of a specific information of an herbarium label.

Figure 7: Data flow of the second code, extraction of the whole text present on the label.

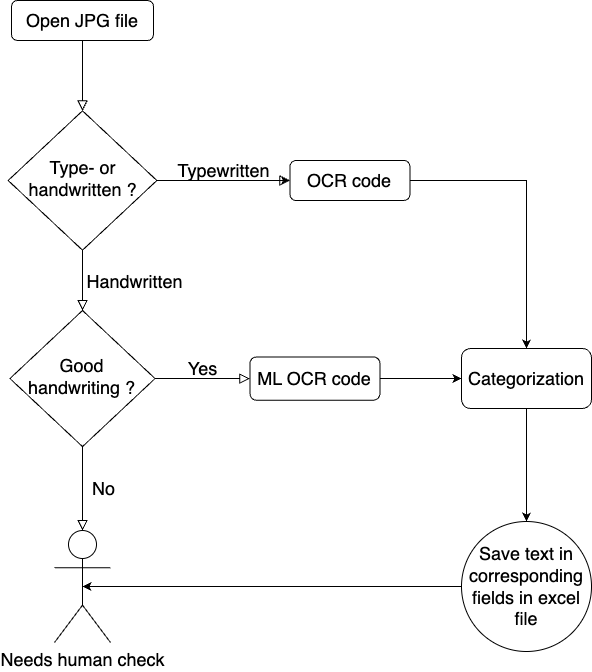


Figure 8: Flow chart of the third code, categorization of the label information after the extraction and then separation into the corresponding fields of the result excel file

# 7 Data Model

Draw and explain your data model at the conceptual level, the logical level and the physical level.

Conceptual

Logical (what columns/features will you use/need)

Physical (infrastructure needs)

# 8 Documentation

How will the project be documented?

# 9 Risks

What can go wrong?

When this and that goes wrong, what counter measures do you have?

What will be the impact on the quality of the aimed output, project time schedule, project cost ?

Text recognition is a difficult task for a program. Some characters like “/” [slash] are often mistaken with the digit “7” [seven].

The herbarium plate have also not always been prepared for a digitization process. Some parts of the plants hide text portions, and in the worst cases, important information present on the label (Figure 9).

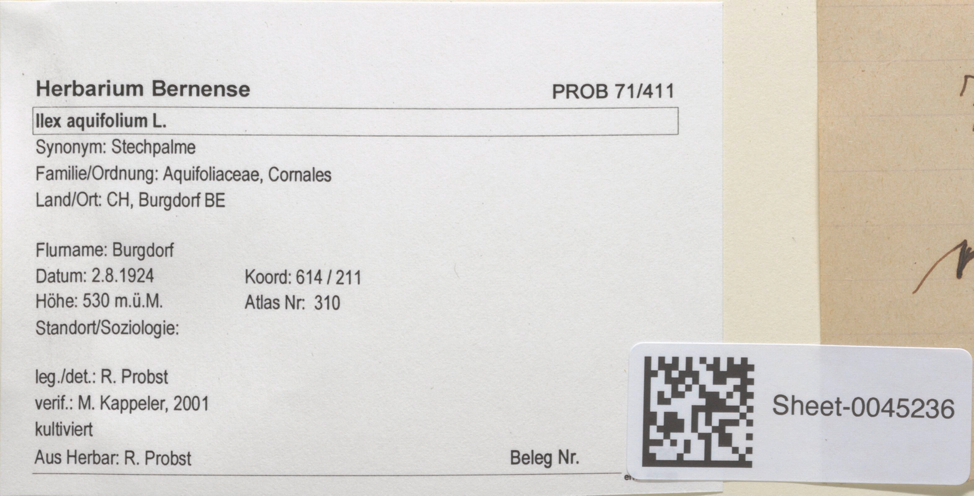


Figure 9: Example of an unsolvable case, the matrix code sticker has been wrongly placed and hides the catalog number (Beleg Nr.).

# 10 Preliminary Studies

Plots and numbers (from Module 2).

# 11 Conclusions

# Acknowledgements

Herbarium BERN

Katja and Sven

Picturae and them who worked on the Solothurner Herbar

Colleagues of the CAS with whom I talked about this project and helped me improving it

# Statement

The following part is mandatory and must be signed by the author or authors.

„Ich erkläre hiermit, dass ich diese Arbeit selbstständig verfasst und keine anderen als die angegebenen Quellen benutzt habe. Alle Stellen, die wörtlich oder sinngemäss aus Quellen entnommen wurden, habe ich als solche gekennzeichnet. Mir ist bekannt, dass andernfalls die Arbeit als nicht erfüllt bewertet wird und dass die Universitätsleitung bzw. der Senat zum Entzug des aufgrund dieser Arbeit verliehenen Abschlusses bzw. Titels berechtigt ist. Für die Zwecke der Begutachtung und der Überprüfung der Einhaltung der Selbstständigkeitserklärung bzw. der Reglemente betreffend Plagiate erteile ich der Universität Bern das Recht, die dazu erforderlichen Personendaten zu bearbeiten und Nutzungshandlungen vorzunehmen, insbesondere die schriftliche Arbeit zu vervielfältigen und dauerhaft in einer Datenbank zu speichern sowie diese zur Überprüfung von Arbeiten Dritter zu verwenden oder hierzu zur Verfügung zu stellen.“

Date: November 17, 2023 Signature(s):

# Appendix 1: Example of an herbarium sheet (Sheet-0043911)



# References and Bibliography

Please number any information source you used in the report with corresponding links here [1]:

[1] S. Haug et al., How to make a CDR, own brain, 2020 (put a weblink or DOI here)

[2]