

Anne Morel
Data Administrator, Info Fauna
anne.s.morel@bluewin.ch

Data Science Project

Text extraction from a herbarium label Conceptual Design Report

5 October 2024

Abstract

Herbaria preserve biodiversity information dating back 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, and name of the collector are the primary information associated to a specimen. The design developed here aims to do the data capture of the herbarium sheet, with the use of a JPG file, by reading the information written on the label and saving it in an excel file. This first approach could open the road leading to do text extraction on the most difficult handwritten labels of our collections. By automating the OCR reading of typewritten herbarium labels, the digitization of natural history collections can be significantly improved, promoting faster, more accurate data processing and increased global accessibility.

Table of Contents

<i>Abstract</i>	<i>0</i>
<i>Table of Contents</i>	<i>1</i>
<i>1 Project Objectives</i>	<i>2</i>
<i>2 Methods</i>	<i>3</i>
<i>3 Data</i>	<i>4</i>
<i>4 Metadata</i>	<i>6</i>
<i>5 Data Quality</i>	<i>6</i>
<i>6 Data Flow</i>	<i>6</i>
<i>7 Data Model</i>	<i>8</i>
<i>8 Documentation</i>	<i>8</i>
<i>9 Risks</i>	<i>8</i>
<i>10 Preliminary Studies</i>	<i>9</i>
<i>11 Conclusions</i>	<i>10</i>
<i>Acknowledgments</i>	<i>11</i>
<i>Statement</i>	<i>11</i>
<i>Appendix 1: Example of a herbarium sheet (Sheet-0043911)</i>	<i>12</i>
<i>References and Bibliography</i>	<i>13</i>

1 Project Objectives

With over 60 million Swiss natural history museum specimens collected over the last 300 years, we have a treasure trove of inestimable wealth and an inexhaustible source of knowledge, touching on many fields of science. Currently, the total amount of digitized specimens with data available in Switzerland is only 17% [1]. The digitization effort is also varying among the different disciplines. Even the biggest herbarium in Switzerland, the Conservatoire and Botanical Garden of Geneva, has only 9% of its 6 million specimens that have been digitized [2].

The longest step in the process of natural history collection digitization is the data capture of the information present on the label of the specimens. Not only do the label format change between collections, but it also changes within a single collection depending on who prepared the specimen for conservation. The focus in this project is solely on herbarium specimens, which are dry pressed plants fixed on paper sheets along with a label. Herbarium specimen labels have standard information (scientific name of the plant, collecting location, collecting dates, and collector names). Typed labels, handwritten labels, handwritten notes written directly on the sheet, nicely handwritten labels and badly handwritten ones, detailed labels and incomplete ones, etc., are only a few examples of the diversity of how the information of a collected plant specimen is kept through time.

This project aims to facilitate the data capture of herbarium sheet standardised typewritten labels by reading with OCR the JPG file of the specimens and saving the result in an Excel file that can easily be imported into a herbarium database. To achieve that, an iterative approach was developed as follows:

- a) Typewritten standardized label – one specific information
 - a. Perform the OCR code on standardized typewritten labels.
 - b. Extract and save a specific information about the label (in this case, the catalog number) in an Excel file, joined with the UID of the picture file (data matrix encoded unique number).
- b) Typewritten standardized label – whole text
 - a. Perform the OCR code on standardized typewritten labels
 - b. Extract and save the whole result in a single Excel cell (corresponding to the verbatim label field in the database), joined with the UID of the picture file (data matrix encoded unique number).

In addition to performing this project, the whole code file doing this is to be written as simple as possible in a text file that can be executed with any computer terminal. The reason to this is that most natural history institutions have protected networks and computers belonging to canton or city administrations. This work shall serve as a basis structure for the IT team of the institutions who would like to use it, so that they can easily adapt it to their own environment constraints.

2 Methods

The dataset for this project consists of digitized herbarium sheets with standardized typewritten labels. These sheets are stored as JPEG files (.jpg format). The images belong to the Herbarium of the Botanical Garden of the University of Bern (BERN Herbarium) database.

This project requires two main elements:

- The text file containing the OCR python code
- A folder with JPG pictures of herbarium sheets

The number of pictures is not important, because the code is executed in serial. For an easier management of the pictures and result file, they should be named with the catalog number of the institution. In our case: BERN-0000001, BERN-0000002, etc.

For the extraction of a specific information on the label, the python code does the following steps. The code line shown serve as an illustration but do not encompass the entire code used.

- | | |
|---|--|
| 1. List all JPEG pictures present in the chosen folder | <pre>> liste_files = os.listdir(main_dir) > new_rows = [] > jpg_files = [file for file in liste_files if file.lower().endswith('.jpg')]</pre> |
| 2. For each element of the list | <pre>> for i in range(len(jpg_files)): > image_path = os.path.join(main_dir, jpg_files[i]) > results = reader.readtext(image_path) > text = ' '.join([result[1] for result in results])</pre> |
| a. Read the text on the picture with the library <i>easyOCR</i> | |
| b. Find a certain string of a length of 2 to 5 digits located after the string "Beleg Nr" or "Probst Nr" | <pre>> posBelegNr = text.find("Beleg Nr") > posProbstNr = text.find("Probst Nr") > if posBelegNr > 0 : > substring1 = text[posBelegNr:posBelegNr + 16] > othercatnum = re.search(r'\d{2,5}', substring1) > if posProbstNr > 0 : > substring2 = text[posProbstNr:posProbstNr + 16] > othercatnum = re.search(r'\d{2,6}', substring2) > if (posBelegNr == -1) & (posProbstNr == -1): > othercatnum = "Not Found"</pre> |
| 3. Save the text read (list of all results) in an Excel file, joined with the corresponding matrix code number. | <pre>> new_row = {'otherCatalogNumber': othercatnum.group(0), "catalogNumber" : jpg_files[i][-4]} > result_file = pd.concat([pd.read_excel(excel_file_path), pd.DataFrame(new_rows)], ignore_index=True) > with pd.ExcelWriter(excel_file_path, engine='openpyxl', mode='a', if_sheet_exists='replace') as writer: > result_file.to_excel(writer, index=False, sheet_name='Sheet1')</pre> |

For the extraction of the whole text of the label, step 2.b. is ignored and the new_row first feature is named "verbatimLabel".

3 Data

The data for this project consists of around 56'000 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. The pictures were not available online at the start of this project in 2023, but are today partially available on the Herbarium Bernense official website (<https://herbarium-bernense.ch/> [accessed on 05.10.2024]).

Each picture corresponds to one 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 is its first large-scale inventory mission.

The collection 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 found in it (Figures 1 to 4).

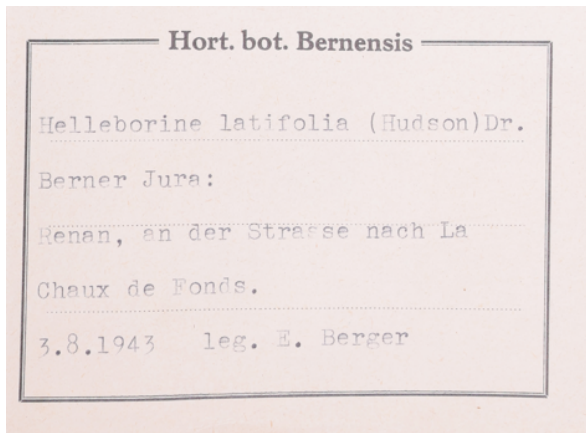


Figure 1: Typewritten label example (Sheet-0061001).

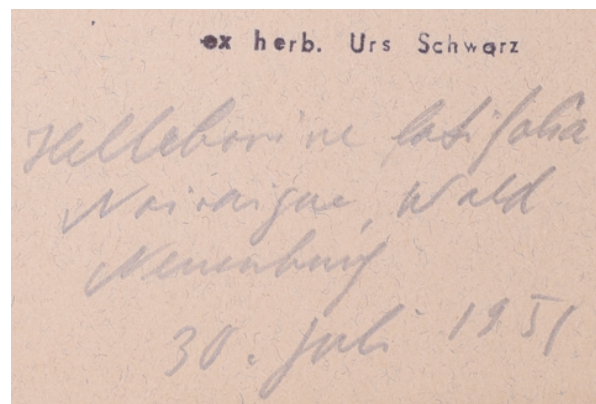


Figure 2: Handwritten label information example, written directly on the sheet of paper (Sheet-0061002)

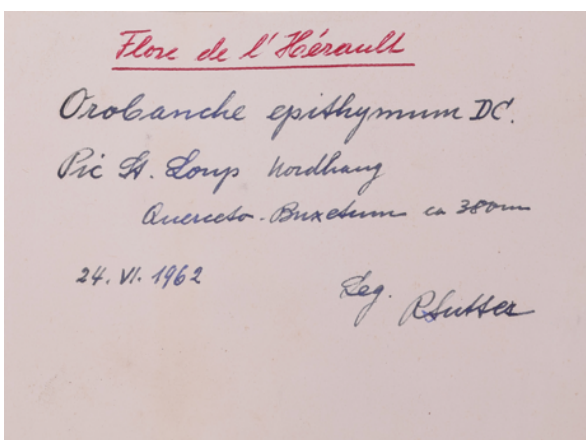


Figure 3: Handwritten label of an important collector in the collection, easily-read handwriting (BERN-0066800)

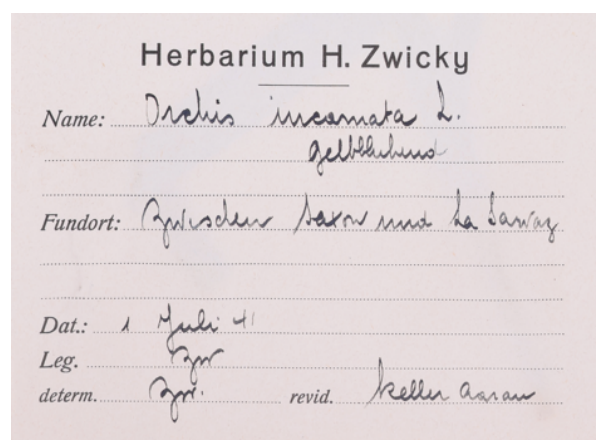


Figure 4: Handwritten label of an important collector in the collection, difficult-to-read handwriting (Sheet-0059007)

For this project, only the herbarium specimen with a certain standardized typewritten label have been used. In the digitization process by Picturae, each specimen had received a new catalog

number in the form of a matrix code sticker (see Appendix 1 for the whole picture). The database of these specimens had to be updated with this new catalog number, by finding the previous catalog number ("Beleg Nr." on the label) and adding the data matrix encoded number (also corresponding to the name of the jpg file) in the corresponding field of the database (Figure 5).



OCR result

In ; ; 8 Herbarium Rob. Streun; Bern & 5 882]
Olquimanin ounxfaz 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 (Gonon)
Miller, Wohlriechender Odermenning
FamilielOrdnung: Rosaceae, Rosales LandlOrt:
CH, Bern BE Flurname: Bern Datum: 10.7.1924
Koord: 601 / 198 Höhe: 540 m.ü.M. 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 Jw!

Figure 5: Sheet-0043911, blue highlight: text of the standardised typewritten label read by the OCR code and data matrix code; orange highlight: target text to be extracted by the code.

The result of the text extraction is a table in an excel file (Table 1), consisting of two columns:

- *otherCatalogNumber*, text extracted on the picture.
- *catalogNumber*, name of the jpg file and new unique catalog number.

lines	otherCatalogNumber	catalogNumber
1	3537	Sheet-0043938
2	3536	Sheet-0043939
3	3535	Sheet-0043940
4	35;	Sheet-0043941
5	3533	Sheet-0043942
...
627	3576	Sheet-0044569
628	8 g0	Sheet-0044570
629	3574	Sheet-0044571
630	3573	Sheet-0044572
631	357~	Sheet-0044573
632	3571	Sheet-0044574
...

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

4 Metadata

The goal of this code is to be performed on any natural history museum specimen picture by focusing first on herbarium specimens. 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 online pictures can only be seen one-to-one. The code developed here is planned to be share among the institutions and adapted to their environment.

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

5 Data Quality

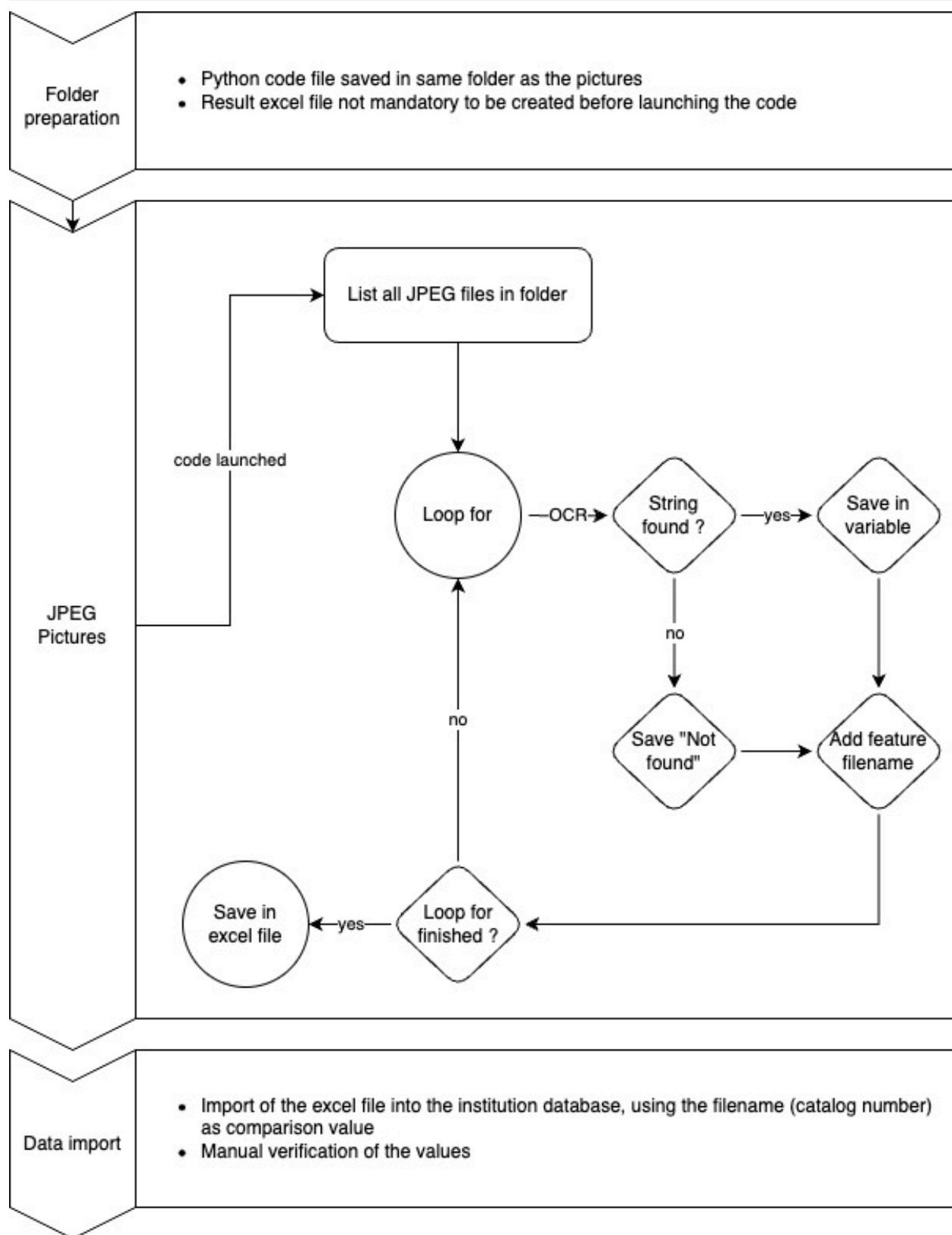
The data quality of the OCR extracted text is checked first by selecting only strings with digits of a length between 2 and 6 digits. Then by making sure this string is located after the fix strings of “Beleg Nr” and “Probst Nr”.

However, these two checkpoints could only be done by knowing the Beleg Nr. format beforehand and that sometimes the code would not read “Beleg Nr” but “Probst Nr”. This discovery has been done during one of the testing phases of the code. However, finding this out, reduced by almost a 100% the OCR mistakes.

Either being when extracting only one part of the label or the entire label, a human verification is still needed at the end of the process. Therefore, it is acceptable if the quality requirements is not 100% correct.

6 Data Flow

Even if the conceptual design is separated in two distinct parts (specific information searched or whole label saved), the data flow is the same, with just one more step to look for the specific information. The data flow presented here details this situation (Graph 1).



Graph 1: Data flow for OCR on a set of JPEG pictures to search a specific string.

7 Data Model

The data model at the conceptual level is to develop a tool that extracts and saves typewritten information on a picture in an excel file. As this tool is aimed for natural history institutions, its complexity has to be reduced, so that it can be used easily on any computer and simply adapted by the local IT team to the institution's requirements.

An OCR model is applied at the logical level, resulting in one feature, the extracted value (otherCatalogNumber), or the text targeted (verbatimLabel). A second associated feature is the unique ID value of the picture, which is obtained from the jpeg file name.

On the physical level, there are no special hardware requirements since the model is simple enough to be executed on any computer. As this code is planned to be shared among the various natural history institutions of Switzerland, and applied on images they already have, even the result file is only temporary. After using the code, the data is directly imported in the institution database system and attributed to the correct specimen metadata.

8 Documentation

The python code text file is self-explanatory, with instructions at the beginning of the file (hidden with comments #) on how to prepare the image folder, where to put the code file, and how to execute the code on the computer terminal.

To facilitate access to the file, a public GitHub repository is created and contains the files and their corresponding explanations. If more python code text files were to be developed for other goals (such as text cleaning, data mapping, etc.), a tutorial page is planned to be written with screenshots and detailed explanations.

Still, the people accessing this file would need a minimum knowledge on coding, so that they can react and adapt appropriately to error messages.

9 Risks

This project is developed to facilitate the data capture of standardized typewritten labels. The worst situation would be the uselessness of the code and the obligation to do the data capture by hand. However this is the current situation, so we can be sure that any improvement, event the smallest, is a huge help to the workflow. Adding to that, excluding the specimens for which the OCR worked perfectly, allows to focus our time and energy only on the ones that did not work, which is already a significant reduction of the total workload.

A risk that has to be avoided though is the introduction of typos in the data. Text recognition is a difficult task for a program. In the current state of the OCR model, some characters like "/" [slash] are often mistaken with the digit "7" [seven]. Also, if the picture is slightly tilted, the code has difficulty to recognize the lines in the text and put some middle portions at the end of the text result. Additionnally, the herbarium sheet have not always been prepared for a digitization process. Some parts of the plants hide text elements, and in the worst cases, even important information that is present on the label (Figure 9).

Finally, the one risk that can't be avoided unless thorough verification before launching the code, is the human limit. In the Figure 6, the person who stucked the data matrix sticker covered the one old catalog number that we are looking for.

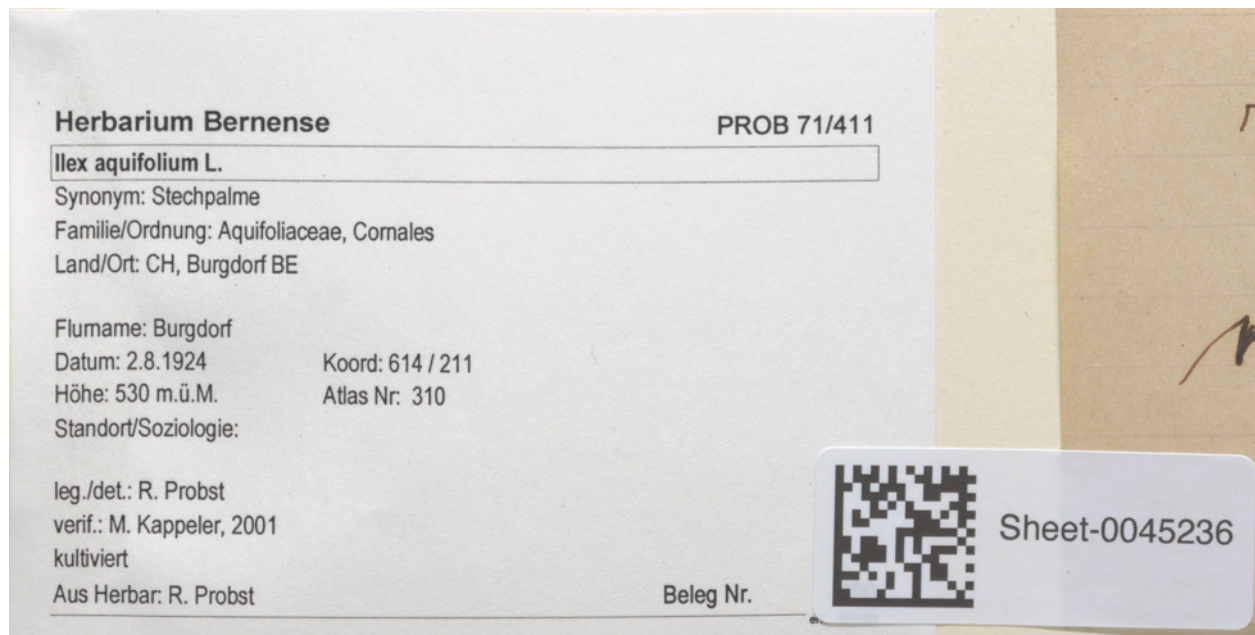
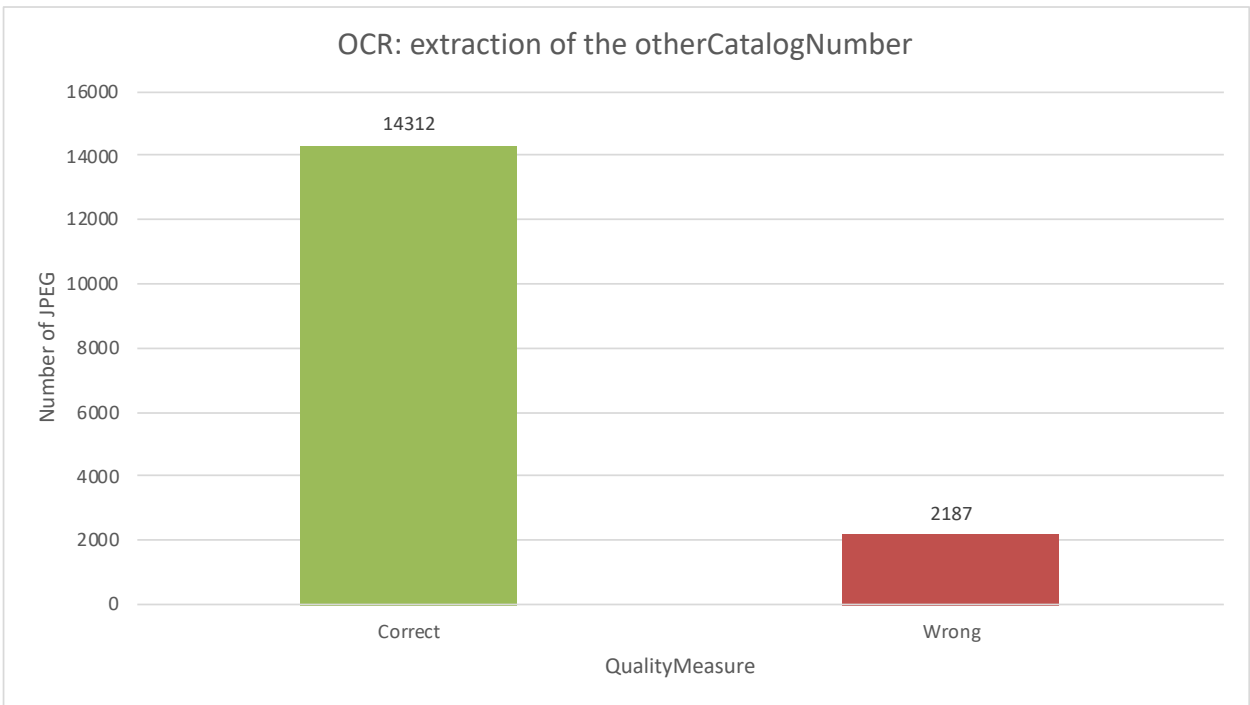


Figure 6: Example of an unsolvable case, the matrix code sticker has been wrongly placed and hides the catalog number (Beleg Nr., bottom right corner of the label).

10 Preliminary Studies

A first version of the code was written in October 2023 and performed on around 46'000 JPEG files of herbarium specimens of the BERN Herbarium (collection: Solothurner Herbar). Unfortunately, due to the trial and error approach used, no overview of the whole result has been kept. The results presented here concern only 16'000 of the original 46'000 files. This can still be considered representative of the whole files treated.

The first version of the code did the OCR on the whole image and extracted only the otherCatalogNumber value (digits located after the "Beleg Nr." String). On a total of 16499 JPEG files considered, 87% were correctly read (14'312 files) and 13% had to be checked because the code could not find the correct string in the OCR result (2187 files) (Graph 2).



Graph 2: Number of otherCatalogNumber values correctly extracted (green) and wrongly extracted (red) in a set of 16'499 JPEG files of herbarium specimens.

11 Conclusions

In the natural history institutions, the use of computer tools to do parts of the collection curation workflow is not yet optimized. Most institutions still use excel files for their database and do repetitive work manually, one specimen after the other. The tool designed here will allow to open the way for more computer-based automatization, not only in the biggest institutions which have developed IT teams, but also in smaller institutions with less budgets and qualified personal. The basic OCR libraries in python allow to already perform most of the work needed, which is already of great help to improve the throughput of collection digitalization.

Acknowledgments

This conceptual design project could be developed thanks to Dr. Katja Rembold, curator, and Sven Imdorf, scientific collaborator, of the Herbarium of the Botanical garden of the University of Bern (BERN Herbarium). The images used for the preliminary analysis were part of a SwissCollNet project of my colleague Sven, and thanks to that it dramatically reduced his workload by months.


I also would like to express my thanks to my CAS ADS 2023 colleagues with whom I talked about this report and who helped me improve it and refine the subject.

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: October 5th, 2024

Signature(s): 

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

References and Bibliography

- [1] Frick, H., & Greeff, M. (2021). Handbook on natural history collections management—A collaborative Swiss perspective. *swiss academies communications*, 16(2). (<https://doi.org/10.3929/ethz-b-000479625>)
- [2] Catalogue des herbiers de Genève (CHG). Conservatoire & Jardin botaniques de la Ville de Genève. 05-10-2024 (<http://www.ville-ge.ch/musinfo/bd/cjb/chg>)