Revista Mexicana de Biodiversidad

Revista Mexicana de Biodiversidad 97 (2026): e975781

Biogeography

Selective-OCR protocol for mass digitization of herbarium specimen labels

Protocolo de OCR selectivo para la digitalización masiva de etiquetas de ejemplares de herbario

Miguel Murguía-Romero [a, *], Diana G. Flores-Camargo [b]

[a] *Universidad Nacional Autónoma de México, Instituto de Biología, Unidad de Informática para la Biodiversidad, Tercer Circuito Exterior s/n, Ciudad Universitaria, Coyoacán, 04510 Ciudad de México, Mexico*

[b] *Universidad Nacional Autónoma de México, Instituto de Biología, Jardín Botánico, Tercer Circuito Exterior s/n, Ciudad Universitaria, Coyoacán, 04510 Ciudad de México, Mexico*

*Corresponding author: miguel.murguia@ib.unam.mx (M. Murguía-Romero)

Received: 07 October 2025; accepted: 20 March 2026

Abstract

It still remains to digitize label data of a high percentage of specimens of vascular plants in the herbariums. There are experiences of the use of the OCR technique to support the process of digitization of specimens, however, it is still necessary to explore and describe in greater detail its limitations and strengths. The digitization of some of the data of the labels from herbarium specimens can be done massively and automatically by applying optical character recognition techniques (OCR). Five target fields (geographic and taxonomic super-primary fields) were selected to obtain their information through OCR applied to 8,451 images of herbarium specimens, guided by a label information architecture and with human intervention. The information contained in the 5 target fields was identified in 70.6% of the labels, 23.7% in 1–4 of the target fields, and only in 5.7% none could be identified. Mistakes ranged from 0.8 to 3.3% depending on the field. OCR cannot automatically identify all information fields of herbarium specimen labels; however, it is possible in a high percentage to retrieve the information of the most consulted fields.

Keywords: Biodiversity informatics; Biological collections; Databases; Image processing; MEXU

Resumen

Aún falta digitalizar los datos de las etiquetas de un alto porcentaje de especímenes de plantas vasculares en los herbarios. Existen experiencias del uso de la técnica OCR para apoyar el proceso de digitalización de especímenes; sin embargo, aún es necesario explorar y describir con mayor detalle sus limitaciones y fortalezas. La digitalización de algunos de los datos de las etiquetas de ejemplares de herbario se puede realizar de forma masiva y automática

ISSN versión electrónica: 2007-8706; Universidad Nacional Autónoma de México, Instituto de Biología. https://doi.org/10.22201/ib.20078706e.2026.97.5781

mediante la aplicación de técnicas de reconocimiento óptico de caracteres (OCR). Se seleccionaron 5 campos objetivo (campos superprimarios geográficos y taxonómicos) para obtener su información mediante OCR aplicado a 8,451 imágenes de especímenes de herbario, guiado por una arquitectura de la información de las etiquetas y con intervención humana. La información contenida en los 5 campos objetivo se identificó en 70.6% de las etiquetas, en 23.7% en 1-4 de los campos objetivo y solo en 5.7% no se pudo identificar ninguna. Los errores variaron de 0.8 a 3.3% según el campo. El OCR no puede identificar automáticamente todos los campos de información de las etiquetas de organismos de herbario, sin embargo, es posible obtener la información en un alto porcentaje de los campos más consultados.

Palabras clave: Informática de la biodiversidad; Colecciones biológicas; Bases de datos; Procesamiento de imágenes; MEXU

Introduction

Having records of herbarium specimens in databases is important from 2 perspectives. First, it allows users to have information regarding global flora in different geographical scales in order to propose more effective and precise strategies for its utilization and conservation (Soltis et al., 2018). Secondly, it facilitates the administration and curation of the collection, because doing an efficient curation of a collection with thousands of specimens as well as maintaining the services that they provide to different types of users is practically impossible now without a database of its specimens.

Currently, we have a great number of herbarium specimens already digitized and available in public databases. For example, the website GBIF (www.gbif.org) has 102.5 million records of vascular plant specimens available. Nevertheless, despite the digitization efforts of the last 3 decades, a great number of specimens from a variety of collections have not been digitized yet, therefore, the information is not available in databases (Corlett, 2022). It has been estimated that at the present speed of the cataloging process, 1,500 years are needed to complete the digitization of the global specimens (Blagoderov et al., 2012). This showcases the need to generate more efficient technological and administrative strategies in order to have, in a shorter period, complete data on the global flora and build comprehensive databases that can be used for analysis.

The optical character recognition technology (OCR) has been proposed to regain the information from herbarium's labels massively, but the exercises in this regard achieve recognition rates below 50% (Takano et al., 2024). However, to increase the recognition percentage, the OCR can be combined with the appropriate use of the database technology, particularly, through the use of catalogs and SQL language, as well as an adequate definition of the information architecture of a specimen that guides the text recognition.

This work proposes a protocol that defines a strategy to fill the gap in information on herborized specimens that have not yet been digitized. The proposed protocol, the Selective-OCR protocol (Appendix 1), addresses 2 aspects in a novel manner. First, it explicitly recognizes that OCR technology fails in the recognition of some words or text. Secondly, it proposes the selective digitization of only 5 of the primary fields (Conabio, 2019), referred to here as super-primary fields. The information of these fields is useful to delineate searches of interest of specimens for users that look for the data. This, when associated with photographs, makes it possible to complete the other data capture with additional efforts based on specific needs, for example, it will be possible to locate the records of specimens collected from a particular state or province, or for a specific species, in the database.

Materials and methods

The Mexican National Herbarium of the Instituto de Biología of the Universidad Nacional Autónoma de México (MEXU) has around 1.6 million herborized specimens of vascular plants, 1.3 million have been digitized. It is estimated that 300,000 specimens from this collection have yet to be digitized (Murguía-Romero et al., 2024). We have selected 8,451 images from herborized specimens to process them via OCR to obtain the data from their labels. The images come from specimens from diverse families and collection years. They were photographed from January to September 2023 by the herbarium personnel. This activity constitutes a previous step to the capture of the information of its labels, which is usually carried out by the same staff and volunteers using the IBdata system's capture interface (https://ibdata4.ib.unam.mx, Murguía-Romero, Serrano-Estrada et al., 2023; Murguía-Romero et al., 2024). IBdata is the web system developed by the Instituto de Biología of the Universidad Nacional Autónoma de México through which the database of specimens of the biological collections

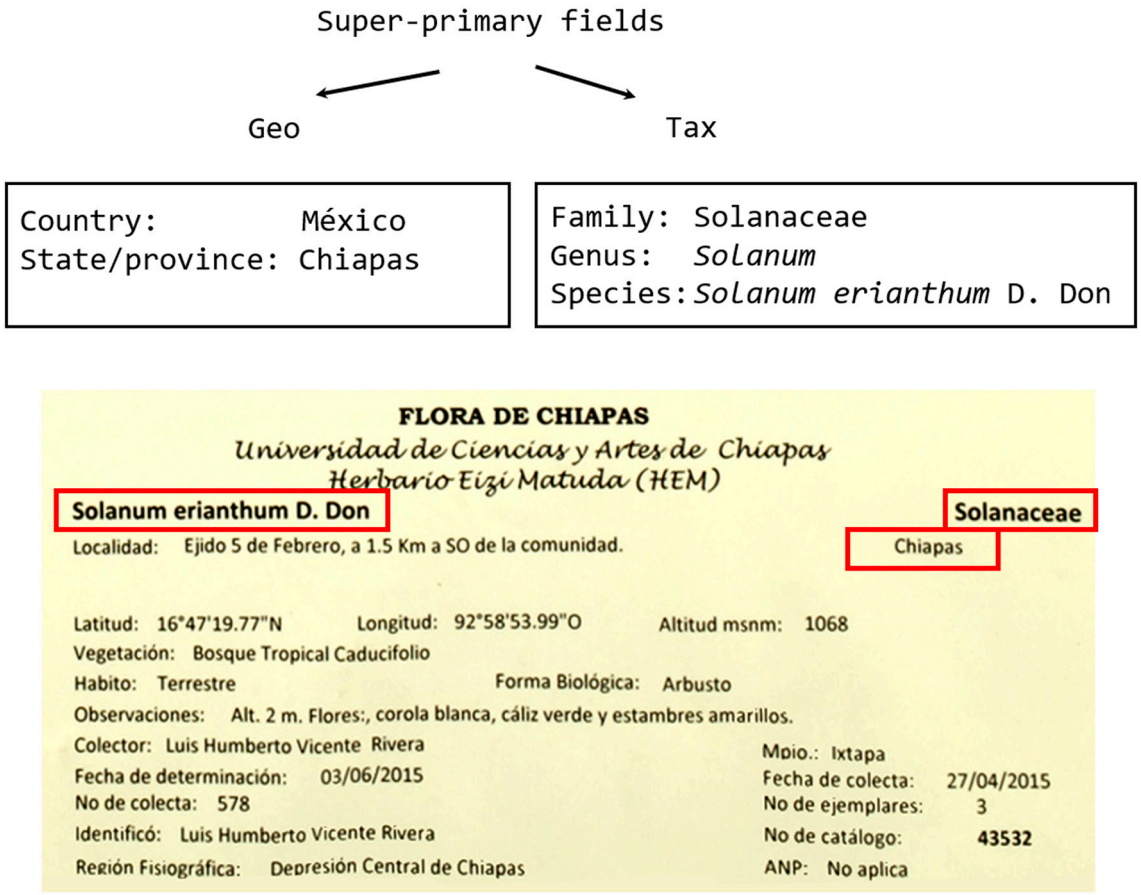

Figure 1. Schematization of the 5 super-primary fields. Upper section: the purpose of the Selective-OCR protocol in this exercise was to recover 5 fields from the specimen labels: 2 geographic and 3 taxonomic. These are defined as GeoTax. Lower section: example of an herbarium specimen label highlighting 3 of the super-primary fields; the genus and country fields are recovered by inferring, in the first case, by looking at the name of the species, and in the second case, by the country-state hierarchy, taking care of verifying the exclusion of homonyms.

(zoological, botanical and mycological) can be consulted, which stores the information following the Darwin Core standard (https://dwc.tdwg.org/terms/); in addition to the visualization of the specimen records, it integrates a capture and editing interface with validation rules.

Because the photography process includes controls that ensure that work is not duplicated, this also ensures that there are no duplicate specimens in the image sample used to apply the OCR protocol. Two geographic fields (country and state) and 3 taxonomic fields (family, genus, and species) were defined as super-primary fields, which have also been called GeoTax fields (Murguía-Romero et al., 2024; Fig. 1). To quantify the percentage of error of assignment of values to each of the 5 fields, 30% of the processed records were reviewed along with the image of their labels. The purpose of applying the protocol in this exercise was to assign 5 super-primary fields of the label of each specimen and to make them available for their public consultations in the IBdata institutional web system.

The only information about the specimens that can be determined from their images prior to processing is the catalog number, as this is part of the file name. Therefore, stratified sampling, for example by taxonomic family, collection date, or state, is not possible. To verify that the sample was unbiased, a Wilcoxon test was performed post-processing, that is, after OCR processing, to compare the percentages by state in the sample and across all specimens in the MEXU database. The image sample included specimens from more than 50 herbaria, ensuring a high level of representativeness in the structure of the processed labels.

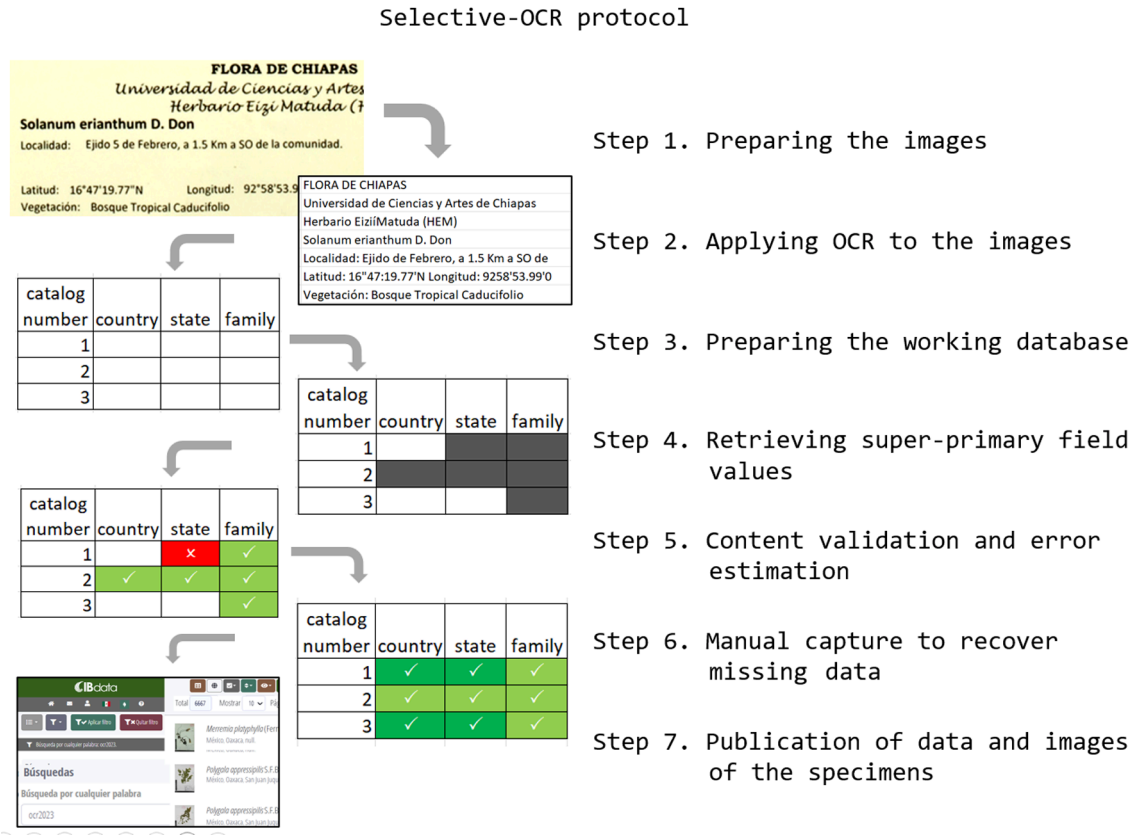

Figure 2. Summary of the Selective-OCR protocol (see Appendix 1 for details).

The specimen's images were subjected to the Selective-OCR protocol (Fig. 2), which can be considered a specialization of the object-to-image-to-data workflow (Nelson et al., 2015). The result of applying OCR to each image was stored in a text file. The average number of lines, after excluding lines with 1 to 2 characters, was 16.5 (SD = 5.11) with a range of 1 to 57. The working database was integrated with 4 tables. The center table "dat_line" was integrated by exporting the lines of text obtained with OCR with 2 additional fields: the row number (sequential) and the name of the image file. The catalog of species, genera, and families "cat_tax" (according to the names accepted by the Kew's World Check List of Vascular Plants, https://powo.science.kew.org; Murguía-Romero, Ortiz et al., 2023), and the catalog from countries and states "cat_geo" (states or provinces) formed 2 other tables. A fourth table "dat_specimen" was created with the list of image names and by adding the 5 super-primary fields (country, state/province, family, genus, and species) and the catalog number to store the identified value in the text lines of the OCR.

Through queries to the database, and by using regular expressions (Friedl, 2006), the names of the species and states were identified in the OCR text table by using the respective catalog tables. The values from the super-primary fields were stored in the particular fields of each record of each label. These searches assume a standard information architecture of a specimen (Morville, 2006), meaning, a pattern of the order and place of each type of information on the surface of the specimen and the label can be proposed. The information architecture proposed includes a label from the specimen, the determination correction labels, the catalog number, and the institutional seals, among other elements (Appendix 2).

Results

The 5 super-primary fields were recovered in 70.6% of the 8,451 OCR analyzed images, in 23.7% at least 1 of the 5 were recovered, and only in 5.7% of cases no fields were recovered (Fig. 3). A manual capture was performed of records corresponding to labels whose primary fields

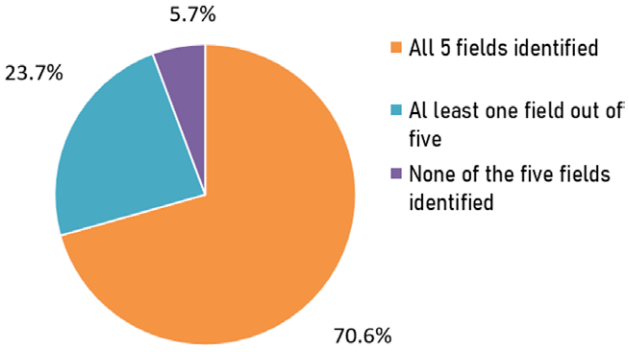

Figure 3. Percentage of correct identification of the values included in the super-primary fields (family, genus, species, country, and state/province) after applying the Selective-OCR protocol.

could not be recovered by processing the OCR lines database. In total, the values of the super-primary fields of 2,483 specimen label records were manually recovered: 2,002 in which the OCR did not identify a super-primary field and 481 in which none were identified.

After processing, it was determined that the specimens in the sample came from 24 different countries, and in the case of Mexico, all 32 states were represented, and more than 200 botanical families were included, which ensures a broad representation of the type of labels on the specimens and reflects the variability of labels in herbaria around the world. No significant differences were found when comparing the percentages of specimens in the 32 Mexican states, which indicates that the sample is not biased in this respect.

One of the main aspects that must be addressed to move towards the evaluation and improvement of data quality is to estimate the error of the proposed protocol (Chapman et al., 2020). To do this, the values of the records that involved manual capture were compared before and after being reviewed and corrected, so the error was estimated based on a 30% sample. The estimated error rate varied from 0.8% to 3.3% depending on the field (Table 1).

Among the main causes identified as to why the values from the super-primary fields could not be recovered with OCR include the badly oriented images, images with degraded letters, text in manuscript, and dot matrix printed labels, among others.

When the values of the 5 super-primary fields for the 8,451 records were registered in the "dat_specimen" table, the catalog identifiers from the IBdata database were assigned and imported into the web system for public consultation. Of this set, 1,145 had already been captured by the herbarium staff and some photographs corresponded to the same specimen, therefore, only 6,667 new records were imported into IBdata. In the IBdata interface (https://ibdata4.ib.unam.mx) the list of these records can be obtained using the Simple Search with the word ocr2023.

Discussion

Using the Selective-OCR protocol, applied by a 2-person team, it was possible to integrate data from 5 fields with more than 8,000 herborized specimens and assign them to their respective Darwin Core fields in a searchable online database. Measuring the method's efficiency is an important parameter for comparing different approaches. In the proposed protocol, the success rate of assignment to the 5 target fields was estimated at 70.6%, while in 5.7% of cases none of the fields could be assigned. The result of the digitization through the Selective-OCR protocol is available in a public database where the data of the super-primary fields of the specimens that were analyzed can be consulted, which provides to those interested in digitization a context of the strengths and weaknesses of this tool.

Other efforts to systematize the use of OCR for digitizing data from herbarium specimen labels are reported in the literature. For example, Guralnick et al. (2024) reported the digitization of labels from 7,995 specimens using OCR and the participation of 177 volunteers; however, the method they propose focuses on the correct identification

Table 1
Percentage of the effectiveness of the Selective-OCR protocol by super-primary field and the estimated associated error.

Super-primary field	% assigned values	% estimated error (95% CI)	Overall success rate (95% CI)
Country	90.30	0.62 (0.20-1.04)	89.74 (88.13-91.35)
State/Province	83.10	0.78 (0.31-1.25)	82.45 (80.43-84.47)
Family	84.10	1.50 (0.85-2.15)	82.88 (80.88-84.88)
Genus	84.10	3.30 (2.35-4.25)	81.31 (79.24-83.38)
Species	79.30	0.80 (0.33-1.27)	78.67 (76.49-80.85)

of the type of label on the specimen; it does not include the classification of information into Darwin Core fields, which, from the perspective of the authors, should be performed by humans. The digitization result cannot be consulted in a public database, which makes it difficult to evaluate the results. Therefore, the success rate in the automatic identification of information from specific fields is not reported either.

The Selective-OCR protocol provides a methodological framework for organizing biological collection digitization projects, as it is structured in sequential steps that allow for clear management of resources and objectives. Another strength of this procedure is that the information architecture of the herborized specimen and the label constitute a tool that addresses the diversity of label formats and can be adapted to different label models. The proposed information architecture is a tool that guides in the allocation of the Darwin Core field. In fact, this architecture is inspired by the knowledge representation called "scripts" (Schank & Abelson, 1975, 1977) in the context of Artificial Intelligence, proposed to represent paradigmatic situations, which, when framed within a general structure, facilitate information processing in specific contexts. The information architecture can be further improved by incorporating more elements considered in the scripts construct. Kirchhoff et al. (2018) have proposed a service-based workflow for automated information extraction from herbarium specimens in which the `text operator' conducts and coordinates all text processing activities after text extraction by OCR, and is assisted by means of a proper user interface; the protocol proposed here addresses the need to reduce the role of the `text operator' so that the identification of fields with Darwin Core is performed automatically through database queries guided by the information architecture. After identifying the type of information, for example, taxonomic names, the corresponding Darwin Core field must be identified, in this case, if it is the identification of the specimen or associated taxa, or in the case of personal names, if it is the person who determined the specimen or the one who collected it.

One of the weaknesses of the protocol we present here is that it did not estimate productivity or costs. This is because the proposed protocol requires human intervention in 6 of the 7 steps. Just as the speed and quality of records of manually captured specimens depend largely on the profile of the person capturing them, the same is true for human intervention in the proposed protocol.

Although the primary objective of the protocol is to digitize specimens whose data are not yet available in databases, the protocol can also be used to validate the integrity and quality of the information captured manually. For example, we have detected that in some database records there are entries without values in the geographic coordinate fields, however, the labels of the corresponding specimens do specify them, indicating that there was an omission in the capture process. In these cases, although the OCR cannot accurately transcribe the coordinates from the labels in the images, it is possible to detect whether they are coordinates. Therefore, the presence or absence of coordinates in the database records can be compared, thus estimating the degree of completeness in the capture of this field.

This work did not investigate or evaluate Large Language Models (LLM; Shanahan, 2024) tools, which are colloquially known as Generative Artificial Intelligences (GAI; Pavlick, 2025). These tools can be used to assign each line of text extracted from the OCR to a specific field; however, in cases where they are applied, the assignment error must also be estimated in order to report the error percentage in queries that access records constructed with this method. The information architecture of the herborized specimen and the label used for the Selective-OCR protocol can be used as a prompt for the GAI tools (Korzynski et al., 2023).

The use of OCR technology is necessarily linked to relational database technology, which is the artifact in which the scan results will ultimately be stored. Therefore, one of the weaknesses of the proposed protocol is that its implementation requires personnel specialized in database management with knowledge of biodiversity data. For example, it will be difficult for a person to review and validate OCR outputs using only spreadsheets without resorting to database management. Likewise, it will be difficult to do it for an IT professional without knowledge of biodiversity as a scientific name, accepted name, synonym, taxonomic hierarchy, among many others. It is therefore important to train personnel specialized in biodiversity informatics, as well as the integration of multidisciplinary work teams.

The Selective-OCR protocol deals with the fact that the use of OCR to recover information from herbarium specimen labels is not all-encompassing. With OCR and its subsequent processing in a relational database using regular expressions, and guided by an information architecture of the specimen, recovering the fields that can be corroborated by catalogs, such as taxonomic and geographic catalogs, is massively feasible and more efficient. The effectiveness and the margin of error of the Selective-OCR protocol reported here can be useful for comparison when applied to more specimens and in other collections.

Acknowledgements

The authors thank the entire MEXU herbarium staff who maintain the collection and participate in photographing the specimens. José Luis Villaseñor reviewed a preliminary version of the manuscript, and his suggestions substantially improved the final version. We are deeply grateful for the comments from the Editor in Chief of RMB and from an anonymous reviewer who substantially improved the final version of the manuscript. Fernanda Y. Bernal Bonilla reviewed and edited the English of the preliminary version.

Appendix 1. Steps of the Selective-OCR protocol.

Step 1. Image preparation

The optical character recognition (OCR) process is carried out on the specimen photographic image. It is important to verify the correct orientation (vertical or horizontal), as well as a minimum resolution of 300 dpi (or even 600 dpi if the font size is very small) to allow recognition using the OCR tool. Two aspects must be decided or taken into account; the first is to define whether character recognition will be performed on the entire image area or only on the area where the data label(s) is (are) located. By choosing the area of the image where only the data label is located, information that can cause noise in the identification of the target information, such as stamps and catalog numbers, can be eliminated. On the other hand, the advantage of applying OCR to the entire image area is that additional information can be identified. What was noise in the context of objective fields, now acquires value, such as the ability to recover catalog numbers. The second aspect is the nomenclature of the image files. It is important that from the creation of the files, names that facilitate their management are assigned, for example, the initials of the person who took the picture, the date of capture, or the room or drawer where the specimens come from, in this last case it might be possible to recognize from the beginning the family or taxonomic group of the specimen.

Step 2. OCR application to the images

OCR will be applied to each of the specimen images, storing in text files all recognized lines. The result will be text files for each image. Diverse programs that do OCR exist, the most adequate one can be chosen depending on the specimen characteristics and the images that will be processed.

Step 3. Working database preparation

The text files coming from the OCR (step 2) will be imported into a database table ("dat_line"). It is recommended to include the specimen or photograph identifier, the line number, and the OCR text line, as fields. The line number is useful when an information architecture is considered for the herborized specimen (Appendix 2), because, when more than one scientific name of a species appears on the label, the order in which they appear is useful to identify the last identification of the specimen and, above all, differences from other scientific names that are indicated as associated taxa (associatedTaxa in Darwin Core, Darwin Core Maintenance Group, https://dwc.tdwg.org/terms/). In addition to the table of OCR lines, it is important to include tables of geographic (county, and state/province) and taxonomic (families, genera, and species) catalogs, such as "cat_geo" and "cat_tax." As well as a fourth table "dat_specimen" where the values from the five super-primary fields will be stored, that can include the specimen identifier or image of the digitized specimen as a primary key.

Step 4. Recovery of values from the super-primary fields

In the "dat_specimen" table, the values from the five super-primary fields will be stored and identified by database queries that use regular expressions in the geographic and taxonomic catalogs. The information architecture of the labels is a key element in differentiating homonym values, for example, the name of the country can be associated with the place of the collector's affiliated institution instead of the collection place or locality (e.g. "Herbario de la Universidad de Sonora"). A taxonomic name may refer to a taxon associated with the locality of collection and not to the taxon identity of the specimen, for instance, "Associated with *Pachycereus weberi*." The order and disposition of the information architecture will determine the type of information it will deal with.

Appendix 1. Continued.

Step 5. Content validation and error estimate

Once the "dat_specimen" table has been completed with the values that the OCR identified, a validation of the allocated data will be carried out by quantifying those cases in which they have been erroneously assigned. This evaluation can be done by either comparing previously captured records done by persons or by selecting a subset of 200 to 400 records and revising them manually with the specimen labels. From this evaluation, the percentage of incorrect allocations of values can be estimated and reported to document the process of digitization via OCR in the respective collection.

Step 6. Manual capture to recover missing data

Optionally, the values from the fields that could not be identified by the OCR will be completed. This is the purpose of identifying the possible causes of error; so that the whole set of processed fields has the values from the five super-primary fields. That way, the users can filter out and select the subsets of interest according to geographic and taxonomic criteria.

Step 7. Publication of data from specimen records

Finally, from the "dat_specimen" table, the analyzed specimen data will be published in repositories that allow for public consultation.

Appendix 2. Information architecture of the herborized specimen and the label used for the Selective-OCR protocol.

The information architecture of a specimen (Fig. A2.1) describes the topology of the elements in a paradigmatic herbarium sample that is used as a guide to identify the correct values of the objective fields of the OCR. Nevertheless, each element could not be in the place described by the information architecture, it serves as a general guide that works for most records.

As part of the information architecture of a specimen, the information architecture of a specimen label is also included (Fig. A2.2), which describes the topology of the information within the main label of the specimen.

In the case that the information is preceded by a category (for example, "Det", "Collection date", etc.), it can be identified easily by the analysis of the texts obtained from the OCR in the database. But, if it is the opposite case, the information architecture helps in this process since the elements corresponding to each field, by the order or apparition on the label, can be excluded or selected. For example, the architecture indicates that the associated taxa are positioned after the taxonomic identification, that way, the first taxonomic name that appears on the texts can be assigned as the identity of the specimen (scientificName), while the subsequent ones, to the "associated taxa" field.

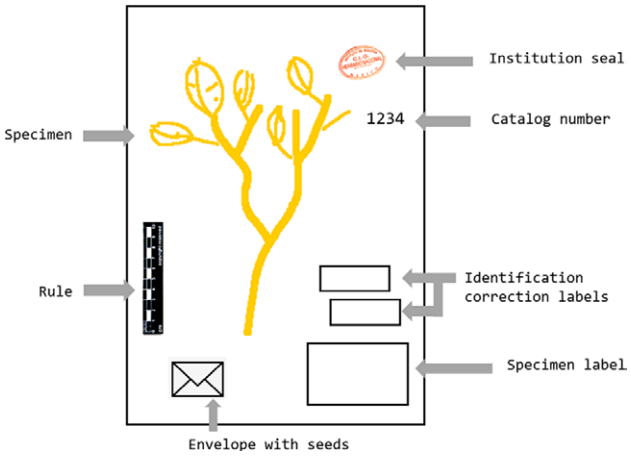

Figure A2.1. Information architecture of a herbarium specimen.

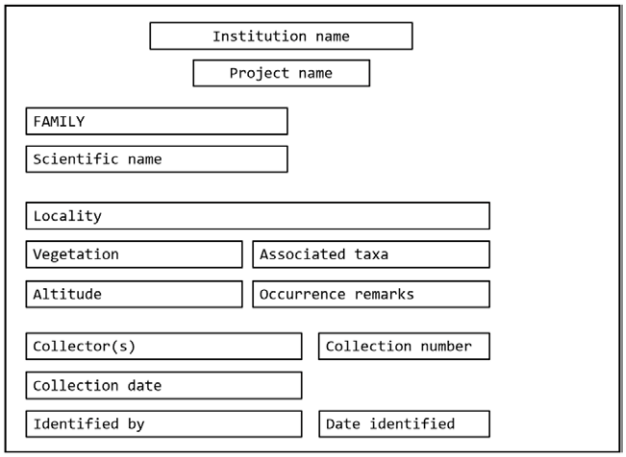

Figure A2.2. Information architecture of a herbarium specimen label.

References

Blagoderov, V., Kitching, I. J., Livermore, L., Simonsen, T. J., & Smith, V. S. (2012). No specimen left behind: industrial scale digitization of natural history collections. *Zookeys*, *209*, 133–146. https://doi:10.3897/zookeys.209.3178

Chapman, A. D., Belbin, L., Zermoglio, P. F., Wieczorek, J., Morris, P. J., Nicholls, M. et al. (2020). Developing standards for improved data quality and for selecting fit for use biodiversity data. *Biodiversity Information Science and Standards*, *4*, e50889. https://doi.org/10.3897/biss.4.50889

Conabio (Comisión Nacional para el Conocimiento y Uso de la Biodiversidad). (2019). *Datos primarios de ejemplares del Sistema Nacional sobre Biodiversidad (SNIB) – características y reglas –*. Ciudad de México. Last access June 3, 2025. www.snib.mx/ejemplares/docs/CONABIO-SNIB-ProtocoloCalidadI.pdf

Corlett, R. T. (2022). We need to accelerate the digitization of existing botanical information and complete the global plant inventory. *Integrative Conservation*, *1*, 6–7. https://doi.org/10.1002/inc3.12

Friedl, J. E. (2006). *Mastering regular expressions*. Sebastopol, CA: O'Reilly Media, Inc.

Guralnick, R., LaFrance, R., Denslow, M., Blickhan, S., Bouslog, S. M., Yost, J. et al. (2024). Humans in the loop: community science and machine learning synergies for overcoming herbarium digitization bottlenecks. *Applications in Plant Sciences*, *12*, e11560. https://doi.org/10.1002/aps3.11560

Kirchhoff, A., Bügel, U., & Santamari, E. (2018). Toward a service-based workflow for automated information extraction from herbarium specimens. *Database*, *2018*, bay103. https://doi.org/10.1093/database/bay103

Korzynski, P., Mazurek, G., Krzypkowska, P., & Kurasinski, A. (2023). Artificial intelligence prompt engineering as a new digital competence: analysis of generative AI technologies such as ChatGPT. *Entrepreneurial Business and Economics Review*, *11*, 25–37. https://doi.org/10.15678/EBER.2023.110302

Morville, P., & Rosenfeld, L. (2006). *Information architecture for the World Wide Web: designing large-scale web sites*. Sebastopol, CA: O'Reilly Media, Inc.

Murguía-Romero, M., Ortiz, E., Serrano-Estrada, B., & Villaseñor, J. L. (2023). The Kew's" World Checklist of Vascular Plants" and its relevance to the knowledge of the flora of Mexico. *Botanical Sciences*, *101*, 632–653. https://doi.org/10.17129/botsci.3223

Murguía-Romero, M., Serrano-Estrada, B., Salazar, G., Gernandt, D. S., Melo-Samper-Palacios, U., Sánchez-González, G. E. et al. (2023). *IBdata v3 "Helia Bravo Hollis". Manual de uso*. Ciudad de México: Instituto de Biología, Universidad Nacional Autónoma de México.

Murguía-Romero, M., Serrano-Estrada, B., Salazar, G. A., Sánchez-González, G. E., Melo-Samper-Palacios, U., Gernandt, D. S. et al. (2024). The IBdata web system for biological collections: design focused on usability. *Biodiversity Informatics*, *18*, 1–12. https://doi.org/10.17161/bi.v18i.20516

Nelson, G., Sweeney, P., Wallace, L. E., Rabeler, R. K., Allard, D., Brown, H. et al. (2015). Digitization workflows for flat sheets and packets of plants, algae, and fungi. *Applications in Plant Sciences*, *3*, 1500065. https://doi.org/10.3732/apps.1500065

Pavlick, G. (2025). *What is Generative AI (GenAI)? How does it work?* Oracle. Accessed June 6, 2025: https://www.oracle.com/uk/artificial-intelligence/generative-ai/what-is-generative-ai/

Schank, R. C., & Abelson, R. P. (1975). Scripts, plans, and knowledge. In *IJCAI'75: Proceedings of the 4th International Joint Conference on Artificial Intelligence, Volume 1* (pp. 151–157). San Francisco, CA: Morgan Kaufmann Publishers Inc. https://dl.acm.org/doi/10.5555/1624626.1624649

Schank, R. C., & Abelson, R. P. (1977). *Scripts, plans, goals, and understanding: an inquiry into human knowledge structures*. Hillsdale, New Jersey: Lawrence Eribaum Associates Inc. Publishers.

Shanahan, M. (2024). Talking about large language models. *Communications of the ACM*, *67*, 68–79. https://doi.org/10.1145/3624724

Soltis, P. S., Nelson, G., & James, S. A. (2018). Green digitization: online botanical collections data answering real-world questions. *Applications in Plant Sciences*, *6*, e1028. https://doi.org/10.1002/aps3.1028

Takano, A., Cole, T. C., & Konagai, H. (2024). A novel automated label data extraction and database generation system from herbarium specimen images using OCR and NER. *Scientific Reports*, *14*, 112. https://doi.org/10.1038/s41598-023-50179-0