ARPHA Preprints

PENSOFT

PREPRINT

Author-formatted, not peer-reviewed document posted on 06/02/2025

DOI: https://doi.org/10.3897/arphapreprints.e148900

A novel approach to the herbarium digitisation workflow

Andriy Novikov, Viktor Nachychko

ARPHA Preprints

Author-formatted, not peer-reviewed document posted on 06/02/2025. DOI:
https://doi.org/10.3897/arphapreprints.e148900

A novel approach to the herbarium digitisation workflow

Andriy Novikov‡, Viktor Nachychko§

‡ State Museum of Natural History of the NAS of Ukraine, Lviv, Ukraine
§ Ivan Franko National University of Lviv, Lviv, Ukraine

Corresponding author: Andriy Novikov (novikoffav@gmail.com)

Abstract

The digitisation workflow currently applied at the herbarium of the State Museum of Natural History of the NAS of Ukraine (LWS) differs from other similar by cascade ('object-to-data-to-image') multilevel organisation. Its application is predicted by the need to preselect specimens by taxon and region, as well as by batched digitisation, which occurs with significant interruptions. Focusing on certain taxonomic groups from specific regions allows to digitise specimens that could be more valuable for early scientific processing. At the same time, the herbarium benefits from such a digitisation model by revising the existing collection classification and keeping the initial ID system. Represented digitisation workflow can be easily reproduced in any herbarium with a limited budget. Provided schemas and notes will help maintain the digitisation, choose appropriate equipment and materials, and pay attention to critical aspects.

Keywords

herbarium, natural history collections, digitisation workflow, imaging, data mobilisation

Introduction

Herbarium digitisation and publishing of mobilised data in open access are crucial modern tasks providing numerous benefits of remote access to collections (Cantrill 2018, Borsch et al. 2020, Nieva de la Hidalga et al. 2020, Powell et al. 2021, Davis 2023). Despite the rapid development of citizen science and specialised biodiversity platforms (e.g., iNaturalist), herbarium collections still remain a valuable source of information for many scientific investigations (Soltis 2017, Funk 2018, Heberling et al. 2019, Heberling 2022, Park et al. 2022, Niedzielski and Markiewicz 2023, Eckert et al. 2024). It has been shown that herbarium specimens better reflect the taxonomic, phylogenetic, and functional diversity of individual regions and allow for much better species distribution modeling, with fewer taxonomic and geospatial errors (Eckert et al. 2024). With the

Author-formatted, not peer-reviewed document posted on 06/02/2025. DOI:
https://doi.org/10.3897/arphapreprints.e148900

advent of remote access to herbarium collections and the development of digital image processing technologies, new approaches to digital data analysis are emerging, including automatic or semi-automatic species identification (Carranza-Rojas et al. 2017, Shirai et al. 2022, Takano et al. 2024), morphological and phenological analyses (Younis et al. 2018, Goëau et al. 2020, Hodač et al. 2024), georeferencing and geotagging (Beaman and Conn 2003, Magdalena et al. 2018, Nowak et al. 2021). Despite impressive progress in the digitisation of the world-leading collections and rapid development and implementation of AI technologies in processing the derivate herbarium data (Pearson et al. 2020, Goëau et al. 2021, Shirai et al. 2022), the herbaria with limited financial and staff facilities are still in search of cost-efficient and simple digitisation solutions (Allan et al. 2019, Takano et al. 2019, Santos et al. 2020, Borsch et al. 2020)

The herbarium of the State Museum of Natural History of the NAS of Ukraine (LWS), located in Lviv city, is one of the oldest in Ukraine. It includes ca. 147,000 specimens, most of which are represented by the specimens of vascular plants. Digitisation at the LWS herbarium was launched in 2012 and occurred sporadically, with interruptions caused mainly by a lack of financial support. Hence, at different times, digitisation focused on different taxonomic groups and aimed to solve various tasks, generally predicting the workflow and final product dissemination. In particular, initially, the type specimens were digitised using outsourced HerbScan hosted at Ivan Franko National University of Lviv (Tasenkevich et al. 2024). Data were captured directly from the received images and stored on JSTOR Global Plants (ITHAKA 2025). Later, data regarding some vascular plants (i.e., rare, relict, and endemic) represented in the flora of the Ukrainian Carpathians were mobilised from the LWS herbarium specimens and published online through the Global Biodiversity Information Facility (GBIF 2025). The digital images for those specimens were not captured or captured occasionally, using different photographic and scanning solutions. Since 2021, the digitisation at the LWS herbarium has become an integral and complex process, including data mobilisation accompanied by specimens imaging, data enhancement and quality control, archiving and publishing. As a result of this effort, recently ca. 8,000 specimens, stored at LWS were digitised and published on GBIF (Novikov and Sup-Novikova 2023, Novikov et al. 2024). Here, we want to share our experience in organising the digitisation process, which could be helpful for institutions holding small- and medium-sized herbaria and having limited financial support.

Digitisation workflow

Digitisation is a complex and multi-level process that is organised differently in different institutions, depending on financial capabilities, technical support, qualified personnel, collection size, and time frames. Usually, the complete digitisation cycle includes several key stages: specimen selection (optional), pre-digitisation curation (cleaning, mounting, barcode placing, etc.), data mobilisation, data enhancement (e.g., georeferencing; optional), and imaging (Takano et al. 2019, Nieva de la Hidalga et al. 2020, Thompson and Birch 2023). Archiving and publishing are stand-alone processes, but they

Author-formatted, not peer-reviewed document posted on 06/02/2025. DOI:
https://doi.org/10.3897/arphapreprints.e148900

nevertheless play a crucial role in completing the digitisation workflow (Heberling 2022, De Smedt et al. 2024). Besides this, data cleaning and quality control occur continuously at different stages and/or after some checkpoints (e.g., after each stage or a certain amount of digitised material), improving the resulting data and images, as well as allowing correcting the digitisation processes in case of need (Nieva de la Hidalga et al. 2020, Engledow 2022, Thompson and Birch 2023, De Smedt et al. 2024).

The general digitisation workflow can be subdivided into three logical lineages depending on the operational objects: specimen, image, and data workflows (Haston et al. 2012a). Different digitisation workflows and their stages can be separated both in time and space, and have irregular or linear order, determining two principal digitisation approaches.

The first approach assumes a linear organisation of digitisation when the specimen operation, imaging, and data processing occur successively. Such an approach is sometimes called 'object-to-image-to-data workflow' (Nelson et al. 2015). It is most widely applied and focuses on obtaining the maximum number of digital images within a limited time. In such a case, data are extracted directly from the specimen image and then processed. Due to high automatisation (sometimes with conveyor belt), often this approach implements fast extraction of a minimum amount of data about the specimen (e.g., barcode and/or specimen ID, taxon name, country and/or local region, and occasionally collector name and collection date) with the idea that this data can be supplemented or corrected later (Haston et al. 2012a, Haston et al. 2012b, Drinkwater et al. 2014, Nelson et al. 2015, Thompson and Birch 2023, De Smedt et al. 2024). This concept of minimal data providing fits well the standard for Minimum Information about a Digital Specimen (MIDS levels 1 or 2 - Haston et al. 2022, Haston and Chapman 2024, De Smedt et al. 2024). Minimising required data and limiting it by standard values that can be automatically read by machine (e.g., using OCR technology or barcode scanners) or unqualified personnel (e.g., volunteers or technicians) has a strategic significance and results in an initial lower number of mistakes. The linear approach, therefore, allows the massive involvement of non-botanists in individual stages, accelerating the general digitisation process. Sometimes, such digitisation is realised by third-party companies specialising in imaging and processing data from natural history collections (e.g., Picturae).

The linear digitisation approach is probably the most cost-effective, but it also has its drawbacks. Automatisation of the processes and the involvement of unqualified personnel can lead to some extra mistakes made in data, which can potentially remain uncorrected or incomplete for an indefinite time. The second disadvantage of such an approach is the potential delay in obtaining even minimal data since they are obtained only after the production of digital images, the processing of which can be terminated. In the case of using automatic text recognition technologies (e.g., OCR), such a delay may be insignificant (Granzow-de la Cerda and Beach 2010, Drinkwater et al. 2014). Recent progress in data extraction from the herbarium labels using machine learning technologies is also promising (Takano et al. 2024). However, sometimes automatic text recognition technologies cannot be applied effectively due to insufficient expertise and/or

Author-formatted, not peer-reviewed document posted on 06/02/2025. DOI:
https://doi.org/10.3897/arphapreprints.e148900

limited sources. Then, a delay in obtaining correct data may occur in the case of barely readable labels, labels with a combination of different handwritings and/or diverse languages, and encrypted or abbreviated labels.

At the LWS herbarium, we applied the second digitisation approach. This so-called 'object-to-data-to-image workflow' (Nelson et al. 2015) has a more complex organisation of digitisation processes and involves data mobilisation before specimen imaging. It means that labels are first quickly photographed and used to gather the data, which are further verified, modified, and supplemented. Only after databasing are the final images of the specimens produced. The cascade approach can be helpful when the preselection of the specimens is required and cannot be implemented automatically (e.g., selection of vouchers of some taxa from a specific region). It is also useful in cases where the herbarium labels are barely readable, partly destroyed, or covered by the specimens. Such an approach allows critical processing of the mobilised data and placing additional notes (e.g., notae criticae) on the specimens before imaging. Specimens preselection and preliminary data processing by specialists result in several advantages, particularly obtaining more complete data of better quality (Tann and Flemons 2008, Granzow-de la Cerda and Beach 2010), albeit for a smaller number of specimens. These enhanced data of MIDS Levels 2 and 3 (Haston and Chapman 2024) can be involved in investigations and published without waiting for the production of digital images. The separation and independence of stages allow optimisation of the digitisation process (Tann and Flemons 2008, Granzow-de la Cerda and Beach 2010). Such an approach also justifies itself in the case of batch digitisation, when digitisation is conducted unsystematically, with lengthy interruptions, and when the staff and available equipment may change significantly. Then, each batch is entirely independent and does not influence the next batch of digitisation. In the cascade approach, the specimen is operated twice: once during the label imaging and once during the final imaging. This results in an extra step of quality control that allows additional specimen selection and solving issues if such issues are discovered.

The main disadvantage of the cascade approach is the general complexity of its logic and automatisation. Requiring qualified personnel makes digitisation more labor-intensive, expensive, and time-consuming. As it was shown, volunteers and inexperienced personnel can be used chiefly for capturing the labels and basic data extraction, while the data interpretation and processing should be implemented instead by experts (Tann and Flemons 2008, Mononen et al. 2014, Brenskelle et al. 2020). In particular, our experience showed that specimens georeferencing by unqualified staff led to misidentified locations that are hard to detect without checking the initial locality data. Among other critical mistakes were incorrectly indicated IDs and arbitrary and incorrectly interpreted specific data (e.g., abbreviations, collector names, and dates), which also required checking the original material. The minor mistakes included overlooked specimens, misspellings of names of taxa, and incomplete or partly missing data (Novikov 2024). This approach supposes a data-intensive strategy (Granzow-de la Cerda and Beach 2010). Hence, the data quality should be prioritised in the case of its application and specialists should be involved as much as possible.

Author-formatted, not peer-reviewed document posted on 06/02/2025. DOI:
https://doi.org/10.3897/arphapreprints.e148900

Current digitisation at the LWS herbarium is focused on the priority group comprising specimens of endemic, rare, and relict taxa of the Ukrainian Carpathians (Novikov et al. 2024b). Hence, due to strict taxonomic and geographic limitations, there is a need to preselect the specimens before their digitisation. In some cases, georeferencing is crucial to verify whether the specimens fit the target group. On the administrative level, it was also decided to keep original specimen IDs. Preselection with preliminary analysis of the label data allows us to prepare an initial list of specimens and place the barcodes with specimen IDs on them before their digitisation. This approach also allowed us to update the existing taxonomy in our collection, merge the specimens placed under different synonymic names into a single folder, and place notae criticae with recent identifications on the specimens before their imaging. Therefore, despite the mentioned disadvantages, it was decided to follow the cascade digitisation approach using the list of tested equipment and services (Suppl. material 1) in the future.

Stage 1. Imaging the labels

Volunteers or technical staff with moderate experience in the taxonomy and geography of the region can be involved. The labels and IDs can be photographed using any digital camera or smartphone. All pictures should have the same orientation (landscape) and size - this will save time during their processing as there will be no need to scale and/or rotate them. If capturing all labels and IDs on a concrete sheet is impossible in the described projection, then a picture of the entire sheet should be taken. The same herbarium can store specimens of the same taxon under different names, as accepted names and taxonomic vision have changed with time. Therefore, it is essential to prepare and follow the checklist of predefined taxa and check both the accepted names and all their possible synonyms (Fig. 1).

The image files should preferably be sorted into folders named correspondingly after the processed taxa (species or infraspecies). However, in the case of intensive digitisation with many different taxa processed daily, placing the files in folders named by working date could be more convenient, with the prospect sorting the files by taxa later. During such bulk photographing, it can be helpful to make contrasting pictures (e.g., of an empty table or blank sheet) at the end of each taxon processing. Such contrasting pictures help to quickly navigate through the massive of pictures and preliminary estimate the number of pictures for each taxon during further processing. Each folder with the multiple pictures obtained in the concrete day should be additionally supported by a text readme file containing the list of processed taxa and any comments on discovered issues.

Stage 2. Data mobilisation and processing

The LWS herbarium contains ca. 147,000 specimens, many of which were collected in the second half of the 1800s - in the 1900s by a few dozen principal collectors. These specimens mostly have handwritten labels, some of which are partly damaged. They were primarily written in Polish, Ukrainian, and Russian. However, many labels here were still written in Latin, Slovakian, Romanian, French, German, and Hungarian. Few

Author-formatted, not peer-reviewed document posted on 06/02/2025. DOI:
https://doi.org/10.3897/arphapreprints.e148900

recent specimens were annotated in English. Some labels combine two languages, e.g., Polish and Latin, Ukrainian and Latin, Russian and Latin, Polish and German. This makes processing the specimens quite labor and requires at least basic knowledge of the mentioned languages and expertise in reading old-manner handwritings.

Therefore, the data from the labels at the LWS herbarium are extracted only by qualified persons directly from the herbarium labels using the two-level protocol (Fig. 2). The involvement of volunteers or unqualified staff is undesirable as it can result in mistakes that are hard to find and fix. The person working with primary data should be familiar with the taxonomy of the selected plant groups and the region's geography. Further data cleaning and cross-validation should preferably be realised by another person experienced in general biodiversity data processing. Otherwise, engaging the volunteers requires the development of an advanced crowdsourcing platform with a rating system, as was proposed by Santos et al. (2020). In such a case, different permission and trust levels should be acquired depending on the transcribers' performance.

The data are input in a Microsoft Excel table preformatted following the DarwinCore standards (TDWG 2025) with the idea to conform the GBIF Occurence class of the datasets (GBIF 2025b). Preliminary data cleaning is performed in the native environment of Microsoft Excel after each portion of work (e.g., after finishing processing all taxa from one family or at the end of the year). However, such data cleaning is not enough for further data processing. Therefore, advanced data cleaning is realised using OpenRefine software (OpenRefine Community 2025). It is worth noting that taxonomic data should be validated in consistency with the GBIF backbone taxonomy (GBIF Secretariat 2023). Mistakes in georeferencing can be checked in different ways, e.g., using QGIS (QGIS Association 2025) or similar software after completing the dataset.

We did not use OCR technologies to parse the data from the herbarium labels due to our lack of expertise and the significant variation of handwritings and languages presented on labels. Unfortunately, the test application of Transkribus (READ-COOP SCE 2024) showed insufficient results, and we failed to train it. At the same time, the relatively low number of specimens can be processed by hand. Therefore, we decided to use the traditional data extraction method until we get an expert in this field.

However, we found it helpful for data processing to create an additional list of the collectors (at the moment, it comprises 374 records), which contains standardised collectors' names in English and shortenings following the IPNI database (IPNI 2025), alternative names (e.g., in original language and other languages), years of life, years of research activity, research interests, geographic coverage. There are also provided links to respective biographic data on IPNI (IPNI 2025), Wikidata (Wikimedia Foundation 2025), HUH Index of Botanists (Harvard University Herbaria & Libraries 2013), Bionomia (Shorthouse 2025), VIAF (OCLC 2025), and other online sources if available. The list is supported by photos of respective handwritings serving as comparison benchmarks. Applying standardised names of collectors allows better data navigation, filtering, and sorting. Moreover, there is no need each time to type the concrete name or verify its correct spelling, especially if it applies diacritic marks. This saves the time a lot.

Author-formatted, not peer-reviewed document posted on 06/02/2025. DOI:
https://doi.org/10.3897/arphapreprints.e148900

Stage 3. Pre-imaging preparation of the specimens

Based on the dataset developed during the label transcription, required specimens are selected from the collection and placed in a separate working table for further pre-imaging preparation. Pre-imaging preparation of the specimens is a multi-level process that includes different tasks, most of which can be done by a technician staff, but specialists should still be involved in quality control. Our experience revealed that even after qualified transcription of the labels, ca. 5-7% of specimens require additional data modifications and clarifications at the stage of pre-imaging preparation.

The initial preparation of the specimen before imaging includes mounting unbound parts of the specimen and labels to the herbarium sheet, packing the small plant parts into the envelope attached to the sheet, restoring labels or preparing new ones, and changing damaged herbarium sheets and/or covers. The second part of the specimen preparation includes placing the stamp of the herbarium, checking and stamping (if needed) the specimen ID, attaching the barcode label, attaching the nota critica with re-identification (if needed), and stamping the sign 'Digitised'. After that, specimens are sorted by accepted taxa (species or infraspecies) and, within each taxon, by ID and then ordered for imaging (Fig. 3).

Nieva de la Hidalga et al. (2020) pointed out that each image of the herbarium specimen should contain five principal image elements required for proper image processing and quality control: color reference chart, scale bar, label, barcode, and institution name. Besides this, we have found it helpful to mark already digitised specimens, which should avoid duplicative processing. Special stickers or a stamp 'Digitised' can be applied for this purpose. Preparation and mounting the stickers is very time-consuming, but it benefits from the possibility of removing them in case of need.

Color reference charts (color checkers) are generally required for digitisation, including the herbarium specimens digitisation (van Dormolen 2012, JSTOR 2018, Guiraud et al. 2019, Rieger et al. 2023). They ensure the proper adjustment of the colors and white balance in the images. Among the most popular color reference charts for herbarium digitisation are X-Rite ColorChecker Classic Nano (40×24 mm) and ISA Golden Thread Object-Level Target x1 (235×25 mm). There are many other options (e.g., Kodak/Tiffen Q13 consisted of two pieces, 203×60 mm each), which differ widely by the number of color patches, their organisation (linear or tabular), size, and price. When choosing the color reference chart, it is important to consider two principal aspects: its size and the ability to automatically calibrate it with specialised software (e.g., ImageZebra). If the chart is too big, it will overlap the specimen. The linear charts can be placed beside the specimen, but tabular charts should be placed on the herbarium sheet to avoid the appearance of extensive empty space on the image. If the chart is not allowed for automatic processing, extra time will be required to calibrate the images. Application of outdated or low-quality charts can result in incorrect image calibration. For the same reason, applying unpopular charts requires an additional indication of its producer and type. The best option in such a case is to indicate the reference color values for all

Author-formatted, not peer-reviewed document posted on 06/02/2025. DOI:
https://doi.org/10.3897/arphapreprints.e148900

patches of the applied chart in CIELAB (L*a*b*) format (Carter et al. 2019). Otherwise, they can be confused and images calibrated improperly. Our previous studies (Novikov et al. 2024b) showed that the ISA Golden Thread Object-Level Target is the best choice since it is less affected by color degradation. Novikov et al. 2024b This target is also preferable due to the extended set of grey patches and the presence of an additional resolution test pattern (so called 'convergent line pair gauge', CLP) for evaluating the images' geometric distortion and spatial frequency response.

Scale bar, crucial for estimation of the sizes of the specimen, is present on many color reference charts. However, many herbaria have branded scale bars that can be placed beside the color reference chart. Some herbaria (e.g., BR, Meise Botanic Garden, Belgium) use a branded scale bar printed on transparent film. This is useful when the scale bar is placed over the specimen. However, special film should be used to print such a scale bar to avoid glares. At the LWS herbarium, we use X-Rite ColorChecker Classic Mini and ISA Golden Thread Object-Level Target charts, each containing scale bars for digitising specimens of vascular plants. Therefore, we do not apply an additional branded scale bar since it requires additional space and manipulations. However, for digitising specimens of non-vascular plants, we use a smaller color reference chart, Charttu Nano (without a preprinted scale bar), supplemented by an additional branded scale bar.

Herbarium digitisation can require introducing a new specimen ID system, which can either be used in parallel with the existing one or will replace the latter. The image should contain a machine-readable barcode or QR code for better automatisation. Specimens are usually ordered in the collection by taxonomic principle, so the numbering of the specimens in the holder or case is not consequent. Sometimes, specimen IDs can be mistakenly duplicated or missing in the collection. The preselection of specimens, printing and placing the barcodes (Code 128, ISO 15417) or QR codes (ISO 18004) corresponding to specimens' original IDs, as well as identification of erroneous IDs, are time-consuming. Therefore, a new ID system is often introduced when barcoded or QR-coded IDs are printed and placed consequently or randomly on the specimens. A good practice in such a case is using globally unique identifiers (GUIDs) to ensure effective digital data curation and databasing (James et al. 2018, Nelson et al. 2018, Kirchhoff et al. 2018).

At the LWS herbarium, we use barcodes with six-digit formatted IDs. Such barcodes are prepared using Zint Barcode Studio software (Stuart 2024) based on the initial dataset before the imaging. They are laser-printed in high-quality mode on the stickers (38.1×21.2 mm) and adhered to the specimens. The application of additional QR codes with GUIDs will be considered when the common digitisation strategy for all the collections kept at the State Museum of Natural History of the NAS of Ukraine is agreed upon.

Keeping the same order of the elements on the herbarium sheet is useful. It helps to navigate over the specimen elements on its image. At the herbarium LWS, the specimen label is typically placed in the right bottom corner, while the specimen ID is in the left bottom corner. The barcode is attached just near or slightly over the original ID. The

Author-formatted, not peer-reviewed document posted on 06/02/2025. DOI:
https://doi.org/10.3897/arphapreprints.e148900

stamp of the herbarium is placed in the right upper corner or, if not possible, in the upper center of the herbarium sheet. The stamp 'Digitised' is placed in the upper left corner of the sheet. Notae criticae, if present and possible, are attached near the original label in the right part of the sheet. The color reference chart, if possible, is placed on the right side of the herbarium sheet just before imaging.

Stage 4. Imaging the specimens

After pre-imaging preparation, the imaging can be done by technician staff, volunteers, or committed to the outsourcing company. This stage is entirely related to image production, adjustment, and file organisation (Fig. 4) and almost eliminates the risk of scientific-related mistakes.

The final imaging of the herbarium specimens can be realised using different technical solutions, i.e., scanners (including flatbed and planetary - see JSTOR (2018), Roma-Marzio et al. (2023), Tasenkevich et al. (2024) for examples) and photo cameras (including DSLR and mirrorless - see Thiers et al. (2016), Sweeney et al. (2018), Takano et al. (2019), Davis et al. (2021) for examples). Each solution has its pros and cons. For example, scanners are relatively expensive, slow, and have a shorter lifespan. Applying flatbed scanners requires additional constructions to revert it and place the specimen under the scanning surface. Such construction is usually bulky (e.g., HerbScan), requiring extra space to hold and providing instead less space to manipulate the specimen. Moreover, the scanner should not contact directly to the specimen to avoid damage to the specimen and pollution of the scanning surface. Master files received with scanners are generally saved in TIFF format and can reach 700 Mb or more, which requires superior saving capacities. However, scanners often produce images of better quality (especially regarding the sharpness and color reproduction) and resolution. Photo stations are usually less expensive to construct, modular, and better suited to various working spaces. A distantly mounted photo camera provides much more space to manipulate the specimen. Photo cameras produce images faster and can be set to automatically create the master files (generally in RAW format but sometimes also in TIFF) and distribution files (lossy images in JPEG format). The obtained files are comparatively smaller in size and require fewer saving capacities. However, the quality of obtained images strongly depends on the camera and lens quality as well as on illumination conditions. Received images can have different optical distortions and uneven lightness uniformity.

The quality is crucial for the images of herbarium specimens, as they can be used for investigations at different magnifications. Nieva de la Hidalga et al. (2020) ascertained general requirements for the images of the herbarium sheets (both stored and web-distributed) and pointed out that color reproduction accuracy (ΔE) should not exceed 5 points. Such color reproduction accuracy conforms to the two-star quality level of the current FADGI guide (Rieger et al. 2023). However, our previous studies (Novikov et al. 2023) showed that many herbarium specimens' images provided in the virtual herbaria do not meet this criterion, reaching only a one-star quality level. Novikov et al. 2023

Author-formatted, not peer-reviewed document posted on 06/02/2025. DOI:
https://doi.org/10.3897/arphapreprints.e148900

However, considering that most herbarium specimens have color much different from living ones, the significance of color reproduction accuracy is questionable. Nieva de la Hidalga et al. (2020) also suggested 72 PPI image resolution sufficient for web publishing and a resolution of 600 PPI - as suitable for preservation and research. As the authors pointed out, such estimates are fair for the scanned images but hardly convertible to apply in photography. Based on our experience, we believe that 10 Mp is the minimum resolution of photo-captured images of the herbarium specimens for web publishing, and 20 Mp is the minimum limit for research and long-term storage.

Nieva de la Hidalga et al. (2020) suggested three types of files that should be produced as a result of the herbarium digitisation, i.e., archiving master file (TIFF), hi-res production file (JPEG2000), and lossy distribution image (JPEG). At the LWS herbarium, we apply only two types: archiving master file (RAW) and distribution file (JPEG). The hi-res production files (80 Mp maximum) can be generated later from the master files if needed. Moreover, we provide distribution files in the highest possible resolution (i.e., 16 or 40 Mp), which is enough for most production and research purposes. Acceptance of such a strategy saves time since both file types are automatically generated by the photo camera.

It is worth noting that image sharpening is not allowed for the archiving master files (van Dormolen 2012, Rieger et al. 2023). It is also recommended to avoid sharpening adjustments in the distribution files. The general processing of the files should include only fixing the image orientation (in most cases to portrait), cropping (FADGI recommends leaving some free space around the specimen; Rieger et al. 2023), and color and white balance adjustment (using presets generated for the applied color reference chart following the producer's recommendations).

Stage 5. Final verification and publishing

The publishing stage includes publishing the data, publishing the images, and crosslinking the data and images (Fig. 5). It also involves data transformation to meet the standard applied in the Data Centre Biodiversity of Ukraine (DCBU 2025), as it does not follow DarwinCore. Such data transformation is a sophisticated process mostly manually implemented by the service provider or trained technicians. Therefore, it is not described here. In general, final data verification and publishing can be done by technician staff or volunteers with elementary experience in IT.

Biodiversity data should be published and permanently archived in appropriate, trusted, general, or domain-specific repositories (Egloff et al. 2016) and follow the FAIR principles (Wilkinson et al. 2016, GoFAIR 2025). Among the important requirements for published data is their persistence, the ability to clearly identify them on the web (e.g., using DOI), and the ability to track the changes. In such a sense, GBIF (2025a) seems to be one of the best solutions for data publishing. However, it is not a specialised virtual herbarium like JACQ (JACQ consortium 2025), Open Herbarium (Open Herbarium 2025), or Reflora (Institute of Research Rio de Janeiro Botanical Garden 2025). GBIF serves as a global data aggregator with outperformed functionality. At the same time, virtual herbaria are

Author-formatted, not peer-reviewed document posted on 06/02/2025. DOI:
https://doi.org/10.3897/arphapreprints.e148900

primary data providers with their specific functional peculiarities mainly focused on work with images and extended specimen annotations.

The herbarium LWS does not have its virtual herbarium platform, and Ukraine does not have a joint specialised national platform or portal for publishing data on digitised collections. Only two independent Ukrainian platforms allow the publication of biodiversity data, including those obtained from the digitised collections, i.e., the Ukrainian Biodiversity Information Network (UkrBin (2025) and the Data Centre Biodiversity of Ukraine (DCBU 2025).DCBU (2025) Both platforms publish different kinds of biodiversity data, including living observations, published reports, and data mobilised from natural history collections. However, both platforms are self-running and do not exchange data with GBIF, limiting data integration and visibility. Moreover, they do not yet allow bulk data export, which makes data extraction laborious. As a result, data published through these platforms are weakly integrated into the research compared to GBIF (2025a).

Considering this, the LWS herbarium data are published online on several platforms simultaneously, i.e., GBIF (2025a), Open Herbarium (2025), and DCBU (2025). GBIF is a principal data-providing platform where the data are deposited as Occurrence class datasets through the Integrated Publishing Toolkit (IPT; Robertson et al. 2014, GBIF 2025c). The images are hosted on the NIRD Service Platform (Sigma2 2025), kindly provided by the Norwegian GBIF node, and synchronised with the dataset using Simple Multimedia extension files (Robertson et al. 2014). The Open Herbarium is a free online platform for publishing virtual herbarium collections developed by Arizona State University (ASU) on the basis of the Symbiota software (Gries et al. 2014, Symbiota Support Hub 2025). The images of the digitised LWS specimens are stored on kindly provided ASU servers and synchronised with the Open Herbarium platform. The DCBU serves as the internal platform to publish and host digitised materials, running independently by the State Museum of Natural History of the NAS of Ukraine. Images of the LWS specimens in the highest possible resolution can be accessed through GBIF or Open Herbarium. At the same time, due to technical limitations, the DCBU hosts images of reduced quality (2.25 Mp maximum).

Stage 6. Archiving

For long-term data storage, FADGI (Rieger et al. 2023) recommends using RAID hard drive massifs with cyclic data corruption control. It is recommended to use several different types of physical media simultaneously (e.g., magnetic tapes and external hard drives; Haston et al. 2012b). Although FADGI does not recommend using optical media for long-term data storage (perhaps due to intensive degradation in the case of direct sunlight and heat; Slattery et al. 2004), they can still be considered a good choice due to their relative longevity in a controlled storage environment and cost value (Brown 2008). Magnetic media also have weaknesses, as they are sensitive to electromagnetic radiation. Such magnetic media as tapes (e.g., LTO tapes) provide superb price/volume value and longevity but can be used only for cold storage as they have limited rewriting

Author-formatted, not peer-reviewed document posted on 06/02/2025. DOI:
https://doi.org/10.3897/arphapreprints.e148900

potential (Wan et al. 2015, Lantz et al. 2025). Moreover, the data on the magnetic tapes can be accessed only consequently, which slows the process. For comparison, any data on the HDD discs, another kind of magnetic storage media, can be accessed at any moment. Direct access to data saves time and ensures that other portions of the HDD magnetic surface are not involved in use and, hence, do not lose their working potential. Nevertheless, HDDs have additional electric and mechanical components, complicating their general construction and, as a result, downgrading the general reliability (Henriksen et al. 2013). Precisely because of the presence of mechanical components, the HDD discs can be easily damaged by drop shock. HDDs have a limited lifetime and their application requires regular data migration like in RAID massifs, which are cost- and energy-consuming (Gu et al. 2014, Bhat 2018). Electronic storage media (e.g., flash drives and SSD discs) strongly depend on electronic components, have limited rewrite resources, and can lose data over time due to storage discharge. In light of this, special optical media like M-discs, BluRay discs with the inorganic MABL recording layer, DVDs with a special recording layer (e.g., Verbatim AZO recording dye) as well as a hard protective layer, can be considered an effective WORM ('write once, read many') data storage. Such optical discs assure long-term storage and reproduction of the data for several decades to hundreds of years (Svrcek 2009, Petrov et al. 2011, Lunt 2011, Iraci 2019, Pioneer 2025).

If archiving is performed on the same type of media, using media from different manufacturers is advisable to avoid possible manufacturing defects and potentially low production quality (Rieger et al. 2023). When archiving on physical media, at least three backup copies should be created, one of which should be stored outside the originating institution. The good practice is to share duplicates of archiving media for storage among specialised libraries, archives, and institutes (Harvey and Mahard 2020, Frick and Greeff 2021). Archiving can also be realised through open online facilities like Zenodo (CERN 2025) or specialised services like Archiver (LIBNOVA Consortium 2025), NIRD (Sigma2 2025), and others.

Similarly to the case of data publishing, the stored materials should be well-structured, well-annotated, clearly identifiable, and represented in open and raw formats (Hart et al. 2016, Wilkinson et al. 2016). There are no clear terms of how long the digitised materials should be preserved, but at least a few decades are expected. Appropriate archiving directly designates the data lifespan. Therefore, it should be realised by archive specialists and cannot be delegated to volunteers or inexperienced staff.

De Smedt et al. (2024) recommend cold storage (i.e., archiving without manipulating the files after that) of herbarium images daily. However, at the LWS herbarium, the images are archived only after renaming and processing at the end of the digitisation bench. They are stored together with metadata and dataset files describing these images.

At the LWS herbarium, the data are archived on several types of the physical media: (a) on the internal server using the service of the National Academy of Sciences of Ukraine; (b) on the SD memory cards (SanDisk Extreme Pro); (c) on the BluRay MABL (Verbatim BD-R) discs (Fig. 6). Such a combination of storage media (magnetic, electronic, and

Author-formatted, not peer-reviewed document posted on 06/02/2025. DOI:
https://doi.org/10.3897/arphapreprints.e148900

optical) should provide long-term preservation of the digitised data and images. Besides this, datasets are also archived on Zenodo (CERN 2025). Moreover, since 2024, the data regarding the digitised specimens have been published as printed herbarium catalogs (e.g., Novikov et al. 2024a, Novikov et al. 2024b). These catalogs are supported by DVDs (Verbatim Archival Grade Gold DVD-R) with datasets and images in JPEG format. The catalogs with the attached discs are distributed through the libraries that serve as additional archiving agents. However, such a combination of number archiving media has drawbacks, namely, controlling the preservation conditions and data persistence can be complicated. Moreover, there is still no common archiving and data management strategy at the State Museum of Natural History of the NAS of Ukraine, which is essential to develop shortly. Developing a common archiving strategy would help resolve questions about file formats, priority long-term storage media, cyclic media renewal terms, and the distribution of roles in data archiving.

Conclusions

The digitisation of the herbarium specimens at the LWS herbarium is a multilevel process organised in a cascade manner. This so-called 'object-to-data-to-image' workflow prioritises data extraction and enhancement to meet MIDS levels 2 and 3. Such digitisation is complicated but allows additional quality control and the involvement of volunteers and nonqualified staff at certain stages while the qualified staff is involved in other stages. At the same time, such an approach allows thematic (i.e., focused on specific taxa and certain regions) digitisation focusing on data mobilisation and their publishing through GBIF.

Digitised herbarium materials are a valuable part of a scientific legacy, supporting and accelerating the various research investigations through virtual open access. The digitisation workflow described in this article and illustrated by the detailed schemas is intended to help other herbaria with limited budgets to effectively organise the digitisation and virtual publishing of the data regarding their collections. Clarifications and notes on each digitisation stage aimed to help choose appropriate equipment, materials, archiving media, and other technical solutions.

Acknowledgements

This paper has been prepared as part of the project “Digitisation of natural history collections damaged as a result of hostilities and related factors: development of protocols and implementation on the basis of the State Museum of Natural History of the National Academy of Sciences of Ukraine” (Nr 2022.01/0013), financed by the National Research Foundation of Ukraine in the grant program “Science for the Recovery of Ukraine in the War and Post-War Periods”.

Author-formatted, not peer-reviewed document posted on 06/02/2025. DOI: https://doi.org/10.3897/arphapreprints.e148900

Conflicts of interest

The authors have declared that no competing interests exist.

References

- Allan EL, Dupont S, Hardy H, Livermore L, Price B, Smith V (2019) High-throughput digitisation of natural history specimens. Biodiversity Information Science and Standards 3 https://doi.org/10.3897/biss.3.37337
- Beaman R, Conn B (2003) Automated geoparsing and georeferencing of Malesian collection locality data. Telopea 10 (1): 43-52. https://doi.org/10.7751/telopea20035604
- Bhat WA (2018) Long-term preservation of big data: prospects of current storage technologies in digital libraries. Library Hi Tech 36 (3): 539-555. https://doi.org/10.1108/lht-06-2017-0117
- Borsch T, Stevens A, Häffner E, Güntsch A, Berendsohn W, Appelhans M, Barilaro C, Beszteri B, Blattner F, Bossdorf O, Dalitz H, Dressler S, Duque-Thüs R, Esser H, Franzke A, Goetze D, Grein M, Grünert U, Hellwig F, Hentschel J, Hörandl E, Janßen T, Jürgens N, Kadereit G, Karisch T, Koch M, Müller F, Müller J, Ober D, Porembski S, Poschlod P, Printzen C, Röser M, Sack P, Schlüter P, Schmidt M, Schnittler M, Scholler M, Schultz M, Seeber E, Simmel J, Stiller M, Thiv M, Thüs H, Tkach N, Triebel D, Warnke U, Weibulat T, Wesche K, Yurkov A, Zizka G (2020) A complete digitization of German herbaria is possible, sensible and should be started now. Research Ideas and Outcomes 6 https://doi.org/10.3897/rio.6.e50675
- Brenskelle L, Guralnick R, Denslow M, Stucky B (2020) Maximizing human effort for analyzing scientific images: A case study using digitized herbarium sheets. Applications in Plant Sciences 8 (6). https://doi.org/10.1002/aps3.11370
- Brown A (2008) Selecting storage media for long-term preservation. The National Archives, Kew, Richmond. URL: https://cdn.nationalarchives.gov.uk/documents/information-management/selecting-storage-media.pdf
- Cantrill D (2018) The Australasian Virtual Herbarium: Tracking data usage and benefits for biological collections. Applications in Plant Sciences 6 (2). https://doi.org/10.1002/aps3.1026
- Carranza-Rojas J, Goeau H, Bonnet P, Mata-Montero E, Joly A (2017) Going deeper in the automated identification of Herbarium specimens. BMC Evolutionary Biology 17 (1). https://doi.org/10.1186/s12862-017-1014-z
- Carter EC, Schanda JD, Hirschler R, Jost S, Luo MR, Melgosa M, Ohno Y, Pointer MR, Rich DC, Viénot F, Whitehead L, Wold JH (2019) Colorimetry. 4th Edition. CIE, Vienna, 111 pp. [ISBN 978-3-902842-13-8] https://doi.org/10.25039/TR.015.2018
- CERN (2025) Zenodo. https://zenodo.org/. Accessed on: 2025-1-16.
- Davis C, Kennedy J, Grassa C (2021) Back to the future: A refined single-user photostation for massively scaling herbarium digitization. Taxon 70 (3): 635-643. https://doi.org/10.1002/tax.12459
- Davis C (2023) The herbarium of the future. Trends in Ecology & Evolution 38 (5): 412-423. https://doi.org/10.1016/j.tree.2022.11.015

Author-formatted, not peer-reviewed document posted on 06/02/2025. DOI: https://doi.org/10.3897/arphapreprints.e148900

- DCBU (2025) Data Centre Biodiversity of Ukraine. http://dc.smnh.org/. Accessed on: 2025-1-14.
- De Smedt S, Bogaerts A, De Meeter N, Dillen M, Engledow H, Van Wambeke P, Leliaert F, Groom Q (2024) Ten lessons learned from the mass digitisation of a herbarium collection. PhytoKeys 244: 23-37. https://doi.org/10.3897/phytokeys.244.120112
- Drinkwater R, Cubey R, Haston E (2014) The use of Optical Character Recognition (OCR) in the digitisation of herbarium specimen labels. PhytoKeys 38: 15-30. https://doi.org/10.3897/phytokeys.38.7168
- Eckert I, Bruneau A, Metsger D, Joly S, Dickinson TA, Pollock L (2024) Herbarium collections remain essential in the age of community science. Nature Communications 15 (1). https://doi.org/10.1038/s41467-024-51899-1
- Egloff W, Agosti D, Patterson D, Hoffmann A, Mietchen D, Kishor P, Penev L (2016) Data Policy Recommendations for Biodiversity Data. EU BON Project Report. Research Ideas and Outcomes 2 https://doi.org/10.3897/rio.2.e8458
- Engledow H (2022) Herbarium specimen label interpretation and transcription: First steps used to clean digitized data. Biodiversity Information Science and Standards 6 https://doi.org/10.3897/biss.6.93888
- Frick H, Greeff M (2021) Handbook on natural history collections management – A collaborative Swiss perspective. Swiss Academies Communications 16 (2): 1-185. https://doi.org/10.5281/zenodo.4316838
- Funk V (2018) Collections-based science in the 21st Century. Journal of Systematics and Evolution 56 (3): 175-193. https://doi.org/10.1111/jse.12315
- GBIF (2025a) Global Biodiversity Information Facility. https://www.gbif.org/. Accessed on: 2025-1-03.
- GBIF (2025b) Dataset classes. https://www.gbif.org/dataset-classes. Accessed on: 2025-1-06.
- GBIF (2025c) IPT: The Integrated Publishing Toolkit. https://www.gbif.org/uk/ipt. Accessed on: 2025-1-14.
- GBIF Secretariat (2023) GBIF Backbone Taxonomy. https://doi.org/10.15468/39omei. Accessed on: 2025-1-06.
- Goëau H, Mora-Fallas A, Champ J, Love NR, Mazer S, Mata-Montero E, Joly A, Bonnet P (2020) A new fine-grained method for automated visual analysis of herbarium specimens: A case study for phenological data extraction. Applications in Plant Sciences 8 (6). https://doi.org/10.1002/aps3.11368
- Goëau H, Bonnet P, Joly A (2021) AI-based identification of plant photographs from herbarium specimens. Biodiversity Information Science and Standards 5 https://doi.org/10.3897/biss.5.73751
- GoFAIR (2025) FAIR Principles. https://www.go-fair.org/fair-principles/. Accessed on: 2025-1-16.
- Granzow-de la Cerda Í, Beach J (2010) Semi-automated workflows for acquiring specimen data from label images in herbarium collections. TAXON 59 (6): 1830-1842. https://doi.org/10.1002/tax.596014
- Gries C, Gilbert E, Franz N (2014) Symbiota – A virtual platform for creating voucher-based biodiversity information communities. Biodiversity Data Journal 2 https://doi.org/10.3897/bdj.2.e1114

- Guiraud M, Groom Q, Bogaerts A, De Smedt S, Dillen M, Saarenmaa H, Wijkamp N, Van der Mije S, Wijers A, Wu Z (2019) Best practice guidelines for imaging of herbarium specimens. ICEDIG, 41 pp. https://doi.org/10.5281/zenodo.3524263
- Gu M, Li X, Cao Y (2014) Optical storage arrays: a perspective for future big data storage. Light: Science & Applications 3 (5). https://doi.org/10.1038/lsa.2014.58
- Hart E, Barmby P, LeBauer D, Michonneau F, Mount S, Mulrooney P, Poisot T, Woo K, Zimmerman N, Hollister J (2016) Ten simple rules for digital data storage. PLOS Computational Biology 12 (10). https://doi.org/10.1371/journal.pcbi.1005097
- Harvard University Herbaria & Libraries (2013) Index of Botanists. https://kiki.huh.harvard.edu/databases/botanist_index.html. Accessed on: 2025-1-07.
- Harvey R, Mahard MR (2020) The preservation management handbook: A 21st-century guide for libraries, archives, and museums. 2nd Edition. Rowman & Littlefield Publishers, Lanham. [ISBN 978-1-5381-0900-7]
- Haston E, Cubey R, Harris D (2012a) Data concepts and their relevance for data capture in large scale digitisation of biological collections. International Journal of Humanities and Arts Computing 6: 111-119. https://doi.org/10.3366/ijhac.2012.0042
- Haston E, Cubey R, Pullan M, Atkins H, Harris D (2012b) Developing integrated workflows for the digitisation of herbarium specimens using a modular and scalable approach. ZooKeys 209: 93-102. https://doi.org/10.3897/zookeys.209.3121
- Haston E, Hardisty A, Chapman C (2022) MIDS definition. 0.16. Release date: 2022-5-22. URL: https://github.com/tdwg/mids/blob/working-draft/old-drafts/MIDS-definition-v0.16-28May2022.md
- Haston E, Chapman C (2024) Minimum Information about a Digital Specimen (MIDS). https://www.tdwg.org/community/cd/mids/. Accessed on: 2024-1-03.
- Heberling JM, Prather LA, Tonsor SJ (2019) The changing uses of herbarium data in an era of global change: An overview using automated content analysis. BioScience 69 (10): 812-822. https://doi.org/10.1093/biosci/biz094
- Heberling JM (2022) Herbaria as big data sources of plant traits. International Journal of Plant Sciences 183 (2): 87-118. https://doi.org/10.1086/717623
- Henriksen SF, Seuskens W, Wijers G (2013) D6.2 Best practices for a digital storage infrastructure for the long-term preservation of digital files. ICT PSP URL: https://pro.europeana.eu/files/Europeana_Professional/Projects/Project_list/Digitising_Contemporary_Art/Deliverables/DCA_D62_Best_practices_for_a_digital_storage_infrastructure_20130506_Version1.pdf
- Hodač L, Karbstein K, Kösters L, Rzanny M, Wittich HC, Boho D, Šubrt D, Mäder P, Wäldchen J (2024) Deep learning to capture leaf shape in plant images: Validation by geometric morphometrics. The Plant Journal 120 (4): 1343-1357. https://doi.org/10.1111/tpj.17053
- Institute of Research Rio de Janeiro Botanical Garden (2025) Reflora. https://reflora.jbrj.gov.br/. Accessed on: 2025-1-16.
- IPNI (2025) International Plant Names Index. https://www.ipni.org/. Accessed on: 2025-1-07.
- Iraci J (2019) Longevity of recordable CDs, DVDs and Blu-rays — Canadian Conservation Institute (CCI) Notes 19/1. https://www.canada.ca/en/conservation-institute/services/conservation-preservation-publications/canadian-conservation-institute-notes/longevity-recordable-cds-dvds.html. Accessed on: 2025-1-16.
- ITHAKA (2025) JSTOR Global Plants. https://plants.jstor.org/. Accessed on: 2025-1-03.

Author-formatted, not peer-reviewed document posted on 06/02/2025. DOI: https://doi.org/10.3897/arphapreprints.e148900

- JACQ consortium (2025) JACQ. https://www.jacq.org/. Accessed on: 2025-1-16.
- James S, Soltis P, Belbin L, Chapman A, Nelson G, Paul D, Collins M (2018) Herbarium data: Global biodiversity and societal botanical needs for novel research. Applications in Plant Sciences 6 (2). https://doi.org/10.1002/aps3.1024
- JSTOR (2018) JSTOR PLANTS Handbook. JSTOR URL: http://www.snsb.info/SNSBInfoOpenWiki/attach/Attachments/JSTOR-Plants-Handbook.pdf
- Kirchhoff A, Bügel U, Santamaria E, Reimeier F, Röpert D, Tebbje A, Güntsch A, Chaves F, Steinke K, Berendsohn W (2018) Toward a service-based workflow for automated information extraction from herbarium specimens. Database 2018 https://doi.org/10.1093/database/bay103
- Lantz M, Furrer S, Petermann M, Rothuizen H, Brach S, Kronig L, Iliadis I, Weiss B, Childers E, Pease D (2025) Magnetic tape storage technology. ACM Transactions on Storage 21 (1): 1-70. https://doi.org/10.1145/3708997
- LIBNOVA Consortium (2025) Archiver. https://archiver-project.eu/. Accessed on: 2025-1-16.
- Lunt B (2011) How long is long-term data storage? In: Society for Imaging Science & Technology (Ed.) Archiving. IS&T Archiving Conference 2011, Salt Lake City, UT, USA, 2011, May 16-19. Society for Imaging Science & Technology, Springfield [ISBN 978-1-63266-640-6].
- Magdalena UR, Silva LAE, Lima RO, Bellon E, Ribeiro R, Oliveira FA, Siqueira MF, Forzza RC (2018) A new methodology for the retrieval and evaluation of geographic coordinates within databases of scientific plant collections. Applied Geography 96: 11-15. https://doi.org/10.1016/j.apgeog.2018.05.002
- Mononen T, Tegelberg R, Sääskilahti M, Huttunen M, Tähtinen M, Saarenmaa H (2014) DigiWeb - a workflow environment for quality assurance of transcription in digitization of natural history collections. Biodiversity Informatics 9 (1). https://doi.org/10.17161/bi.v9i1.4748
- Nelson G, Sweeney P, Wallace L, Rabeler R, Allard D, Brown H, Carter JR, Denslow M, Ellwood E, Germain-Aubrey C, Gilbert E, Gillespie E, Goertzen L, Legler B, Marchant DB, Marsico T, Morris A, Murrell Z, Nazaire M, Neefus C, Oberreiter S, Paul D, Ruhfel B, Sasek T, Shaw J, Soltis P, Watson K, Weeks A, Mast A (2015) Digitization workflows for flat sheets and packets of plants, algae, and fungi. Applications in Plant Sciences 3 (9). https://doi.org/10.3732/apps.1500065
- Nelson G, Sweeney P, Gilbert E (2018) Use of globally unique identifiers (GUIDs) to link herbarium specimen records to physical specimens. Applications in Plant Sciences 6 (2). https://doi.org/10.1002/aps3.1027
- Niedzielski P, Markiewicz J (2023) Digitalization of herbarium collections as a tool for the commercialization of scientific knowledge. Procedia Computer Science 225: 2194-2203. https://doi.org/10.1016/j.procs.2023.10.210
- Nieva de la Hidalga A, Rosin P, Sun X, Bogaerts A, De Meeter N, De Smedt S, Strack van Schijndel M, Van Wambeke P, Groom Q (2020) Designing an herbarium digitisation workflow with built-in image quality management. Biodiversity Data Journal 8 https://doi.org/10.3897/bdj.8.e47051
- Novikov A, Sup-Novikova M (2023) Endemic vascular plants of the Ukrainian Carpathians. Version 1.7. https://doi.org/10.15468/5hrh87. Accessed on: 2025-1-03.

Author-formatted, not peer-reviewed document posted on 06/02/2025. DOI: https://doi.org/10.3897/arphapreprints.e148900

- Novikov A, Sup-Novikova M, Nachychko V, Kuzyarin O (2023) Testing budget photosystems to reach an optimal solution for the herbarium digitization purposes. Plant Introduction 99/100: 36-50. https://doi.org/10.46341/pi2023010
- Novikov A (2024) On the importance of systematicity and reliability when mobilizing herbarium data. In: Novikov A (Ed.) Digitization of natural history collections: challenges and achievements. Digitization of natural history collections: challenges and achievements, Lviv, 11 October 2024. State Museum of Natural History of the NAS of Ukraine, Lviv, 27-28 pp. [In Ukrainian].
- Novikov A, Kuzyarin O, Nachychko V, Susulovska S (2024a) LWS herbarium catalogue. Vascular plants. Volume 1. Rare, relict and range-limited taxa in the Ukrainian Carpathians and adjacent territories. State Museum of Natural History of the NAS of Ukraine, Lviv, 288 pp. https://doi.org/10.5281/zenodo.13993748
- Novikov A, Kuzyarin O, Nachychko V, Susulovska S (2024b) LWS herbarium catalogue. Vascular plants. Volume 2. Endemic and subendemic taxa in the Ukrainian Carpathians and adjacent territories. State Museum of Natural History of the NAS of Ukraine, Lviv, 97 pp. https://doi.org/10.5281/zenodo.13993782
- Novikov A, Nachychko V, Kuzyarin O, Susulovska S (2024c) LWS herbarium. Vascular plants. Version 1.14. https://doi.org/10.15468/58zxna. Accessed on: 2025-1-03.
- Novikov A, Sup-Novikova M, Nachychko V, Kuzyarin O (2024d) Testing color reference charts for the herbarium digitization purposes. Plant Introduction 101/102: 34-43. https://doi.org/10.46341/pi2024004
- Novikov A, Savytska A, Kuzyarin O, Nachychko V, Susulovska S, Rizun V, Susulovsky A, Hushtan H, Hushtan K, Leleka D (2024e) Data mobilisation in the LWS Herbarium: success and prospects. Biodiversity Data Journal 12 https://doi.org/10.3897/bdj.12.e117292
- Nowak M, Słupecka K, Jackowiak B (2021) Geotagging of natural history collections for reuse in environmental research. Ecological Indicators 131 https://doi.org/10.1016/j.ecolind.2021.108131
- OCLC (2025) VIAF: The Virtual International Authority File. https://viaf.org/. Accessed on: 2025-1-07.
- Open Herbarium (2025) Open Herbarium Portal. https://openherbarium.org/. Accessed on: 2025-1-14.
- OpenRefine Community (2025) OpenRefine. URL: https://openrefine.org/
- Park D, Lyra G, Ellison A, Maruyama RKB, Torquato DdR, Asprino R, Cook B, Davis C (2022) Herbarium records provide reliable phenology estimates in the understudied tropics. bioRxiv https://doi.org/10.1101/2022.08.19.504574
- Pearson KD, Nelson G, Aronson MFJ, Bonnet P, Brenskelle L, Davis CC, Denny EG, Ellwood ER, Goëau H, Heberling JM, Joly A, Lorieul T, Mazer SJ, Meineke EK, Stucky BJ, Sweeney P, White AE, Soltis PS (2020) Machine learning using digitized herbarium specimens to advance phenological research. BioScience 70 (7): 610-620. https://doi.org/10.1093/biosci/biaa044
- Petrov V, Kryuchyn A, Gorbov I (2011) High-density optical disks for long-term information storage. SPIE Proceedings 8011 https://doi.org/10.1117/12.900745
- Pioneer (2025) Advantages of BD-R for long-term storage and results of lifetime estimate testing from third party organizations (ADTC). https://pioneer-blurayodd.eu/features/01/. Accessed on: 2025-1-16.

Author-formatted, not peer-reviewed document posted on 06/02/2025. DOI: https://doi.org/10.3897/arphapreprints.e148900

- Powell C, Krakowiak A, Fuller R, Rylander E, Gillespie E, Krosnick S, Ruhfel B, Morris A, Shaw J (2021) Estimating herbarium specimen digitization rates: Accounting for human experience. Applications in Plant Sciences 9 (4). https://doi.org/10.1002/aps3.11415
- QGIS Association (2025) QGIS Geographic Information System. URL: https://www.qgis.org/
- READ-COOP SCE (2024) Transkribus. https://www.transkribus.org/. Accessed on: 2024-6-20.
- Rieger T, Phelps KA, Beckerle H, Brown T, Frederick R, Mitrani S, Breen P, Breitbart M, Williams D, Triplett R, Horsley M (Eds) (2023) Technical guidelines for digitizing cultural heritage materials. 3rd. Federal Agencies Digitization Guidelines Initiative, USA, 129 pp. URL: https://www.digitizationguidelines.gov/guidelines/digitize-technical.html
- Robertson T, Döring M, Guralnick R, Bloom D, Wieczorek J, Braak K, Otegui J, Russell L, Desmet P (2014) The GBIF integrated publishing toolkit: Facilitating the efficient publishing of biodiversity data on the internet. PLOS One 9 (8). https://doi.org/10.1371/journal.pone.0102623
- Roma-Marzio F, Maccioni S, Dolci D, Astuti G, Magrini N, Pierotti F, Vangelisti R, Amadei L, Peruzzi L (2023) Digitization of the historical Herbarium of Michele Guadagno at Pisa (PI-GUAD). PhytoKeys 234: 107-125. https://doi.org/10.3897/phytokeys.234.109464
- Santos J, Rupino da Cunha P, Sales F (2020) A strategy to digitise natural history collections with limited resources. Biodiversity Data Journal 8 https://doi.org/10.3897/bdj.8.e55959
- Shirai M, Takano A, Kurosawa T, Inoue M, Tagane S, Tanimoto T, Koganeyama T, Sato H, Terasawa T, Horie T, Mandai I, Akihiro T (2022) Development of a system for the automated identification of herbarium specimens with high accuracy. Scientific Reports 12 (1). https://doi.org/10.1038/s41598-022-11450-y
- Shorthouse D (2025) Bionomia: Linking natural history specimens to the world's collectors. https://bionomia.net/. Accessed on: 2025-1-07.
- Sigma2 (2025) NIRD Service Platform. https://www.sigma2.no/service/nird-service-platform. Accessed on: 2025-1-14.
- Slattery O, Lu RC, Zheng J, Byers F, Tang X (2004) Stability comparison of recordable optical discs - A study of error rates in harsh conditions. Journal of Research of the National Institute of Standards and Technology 109 (5). https://doi.org/10.6028/jres.109.038
- Soltis P (2017) Digitization of herbaria enables novel research. American Journal of Botany 104 (9): 1281-1284. https://doi.org/10.3732/ajb.1700281
- Stuart R (2024) Zint. 2.4. URL: https://zint.org.uk/
- Svrcek I (2009) Accelerated life cycle comparison of Millenniata Archival DVD. China Lake Department of Defence, China Lake, 75 pp. URL: https://archive.org/details/millenniata-archival-dvd-m-disk-china-lake-full-report-november-10-2009/mode/2up
- Sweeney P, Starly B, Morris P, Xu Y, Jones A, Radhakrishnan S, Grassa C, Davis C (2018) Large-scale digitization of herbarium specimens: Development and usage of an automated, high–throughput conveyor system. Taxon 67 (1): 165-178. https://doi.org/10.12705/671.10
- Symbiota Support Hub (2025) Symbiota. Open-source biodiversity data management software. https://symbiota.org/. Accessed on: 2025-1-14.

Author-formatted, not peer-reviewed document posted on 06/02/2025. DOI: https://doi.org/10.3897/arphapreprints.e148900

- Takano A, Horiuchi Y, Fujimoto Y, Aoki K, Mitsuhashi H, Takahashi A (2019) Simple but long-lasting: A specimen imaging method applicable for small- and medium-sized herbaria. PhytoKeys 118: 1-14. https://doi.org/10.3897/phytokeys.118.29434
- Takano A, Cole TH, Konagai H (2024) A novel automated label data extraction and data base generation system from herbarium specimen images using OCR and NER. Scientific Reports 14 (1). https://doi.org/10.1038/s41598-023-50179-0
- Tann J, Flemons PKJ (2008) Data capture of specimen labels using volunteers. Australian Museum.
- Tasenkevich L, Skrypets K, Seniv M, Khmil T (2024) Digitization of the oldest botanical collection in Ukraine (LW Herbarium): a case study. Biodiversity: Research and Conservation 75: 1-8. https://doi.org/10.14746/biorc.2024.75.2
- TDWG (2025) Darwin Core. https://dwc.tdwg.org/. Accessed on: 2024-1-06.
- Thiers B, Tulig M, Watson K (2016) Digitization of The New York Botanical Garden Herbarium. Brittonia 68 (3): 324-333. https://doi.org/10.1007/s12228-016-9423-7
- Thompson K, Birch J (2023) Mapping the digitisation workflow in a university herbarium. Research Ideas and Outcomes 9 https://doi.org/10.3897/rio.9.e106883
- UkrBin (2025) Ukrainian Biodiversity Information Network . https://ukrbin.com/. Accessed on: 2025-1-15.
- van Dormolen H (2012) Metamorfoze preservation imaging guidelines. Koninklijke Bibliotheek, 44 pp.
- Wan S, Cao Q, Xie C (2015) Optical storage: an emerging option in long-term digital preservation. Frontiers of Optoelectronics 7 (4): 486-492. https://doi.org/10.1007/s12200-014-0442-2
- Wikimedia Foundation (2025) Wikidata. https://www.wikidata.org/. Accessed on: 2025-1-07.
- Wilkinson M, Dumontier M, Aalbersberg IJ, Appleton G, Axton M, Baak A, Blomberg N, Boiten J, da Silva Santos LB, Bourne P, Bouwman J, Brookes A, Clark T, Crosas M, Dillo I, Dumon O, Edmunds S, Evelo C, Finkers R, Gonzalez-Beltran A, Gray AG, Groth P, Goble C, Grethe J, Heringa J, 't Hoen PC, Hooft R, Kuhn T, Kok R, Kok J, Lusher S, Martone M, Mons A, Packer A, Persson B, Rocca-Serra P, Roos M, van Schaik R, Sansone S, Schultes E, Sengstag T, Slater T, Strawn G, Swertz M, Thompson M, van der Lei J, van Mulligen E, Velterop J, Waagmeester A, Wittenburg P, Wolstencroft K, Zhao J, Mons B (2016) The FAIR Guiding Principles for scientific data management and stewardship. Scientific Data 3 (1). https://doi.org/10.1038/sdata.2016.18
- Younis S, Weiland C, Hoehndorf R, Dressler S, Hickler T, Seeger B, Schmidt M (2018) Taxon and trait recognition from digitized herbarium specimens using deep convolutional neural networks. Botany Letters 165: 377-383. https://doi.org/10.1080/23818107.2018.1446357

ARPHA Preprints
Author-formatted, not peer-reviewed document posted on 06/02/2025. DOI:
https://doi.org/10.3897/arphapreprints.e148900

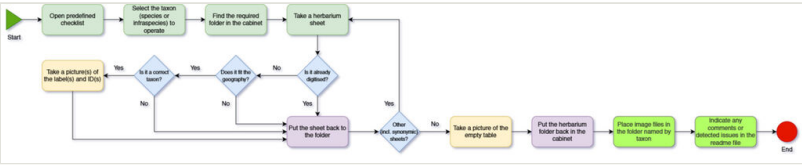

Figure 1.

The flow chart of imaging the labels of specimens of predefined taxa from a particular region.

ARPHA Preprints
Author-formatted, not peer-reviewed document posted on 06/02/2025. DOI:
https://doi.org/10.3897/arphapreprints.e148900

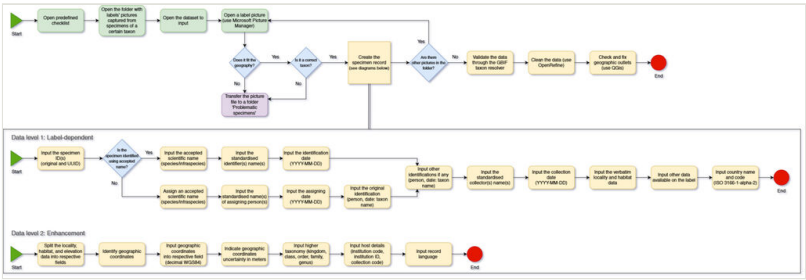

Figure 2.

The flow chart of data mobilisation from the herbarium labels and data enhancement.

ARPHA Preprints
Author-formatted, not peer-reviewed document posted on 06/02/2025. DOI:
https://doi.org/10.3897/arphapreprints.e148900

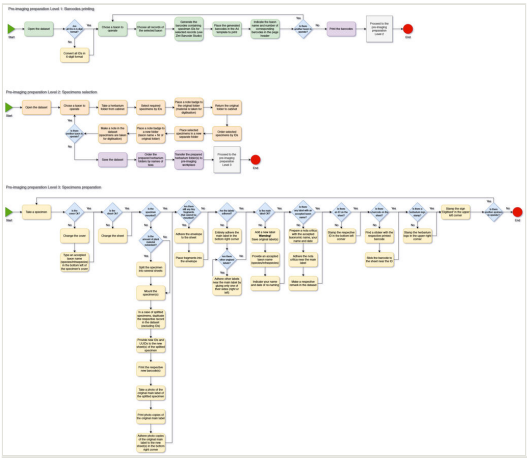

Figure 3.

The flow chart of pre-imaging preparation of the specimens.

ARPHA Preprints
Author-formatted, not peer-reviewed document posted on 06/02/2025. DOI:
https://doi.org/10.3897/arphapreprints.e148900

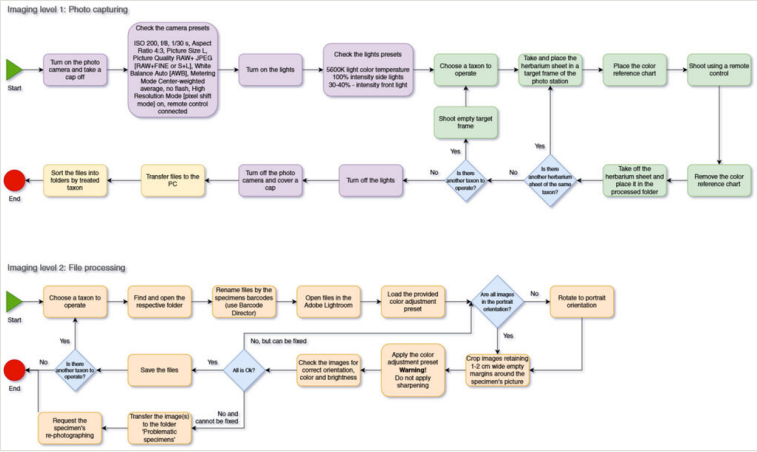

Figure 4.

The flow chart of imaging and image processing.

Author-formatted, not peer-reviewed document posted on 06/02/2025. DOI:
https://doi.org/10.3897/arphapreprints.e148900

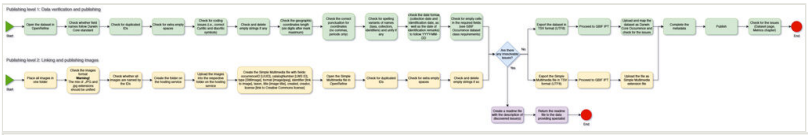

Figure 5.

The flow chart of final data verification, cross-linking with images, and publishing on GBIF.

Author-formatted, not peer-reviewed document posted on 06/02/2025. DOI:
https://doi.org/10.3897/arphapreprints.e148900

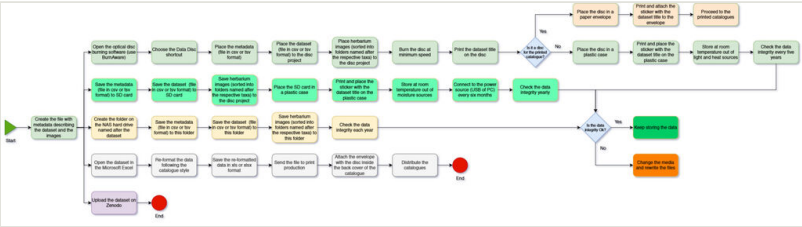

Figure 6.

The flow chart of the data and images archiving.

ARPHA Preprints
Author-formatted, not peer-reviewed document posted on 06/02/2025. DOI:
https://doi.org/10.3897/arphapreprints.e148900

Supplementary material

Suppl. material 1: The equipment, materials, software and other sources applied during the digitisation of the specimens at the LWS herbarium

Authors: Andriy Novikov & Viktor Nachychko
Data type: Textual description
Brief description: The supplement contains the list of the equipment, materials, software, and online resources applied during the digitisation of the specimens at the LWS herbarium.
Download file (23.10 kb)