ARPHA Preprints

PENSOFT

PREPRINT

Author-formatted, not peer-reviewed document posted on 29/05/2025

DOI: https://doi.org/10.3897/arphapreprints.e160486

Quantification of plant trait data from herbarium scans in the DiSSCo Research Infrastructure

Rajapreethi Rajendran, Claus Weiland, Jonas Grieb, Soulaine Theocharides, Sam Leeflang, Wouter Addink, Sharif Islam

ARPHA Preprints *Author-formatted, not peer-reviewed document posted on 29/05/2025*. DOI: https://doi.org/10.3897/arphapreprints.e160486

Quantification of plant trait data from herbarium scans in the DiSSCo Research Infrastructure

Rajapreethi Rajendran[‡], Claus Weiland[‡], Jonas Grieb[‡], Soulaine Theocharides[§], Sam Leeflang[§,|], Wouter Addink[§,|], Sharif Islam[§,¶]

‡ Senckenberg – Leibniz Institution for Biodiversity and Earth System Research, Frankfurt am Main, Germany
§ Naturalis Biodiversity Center, Leiden, Netherlands
| Distributed System of Scientific Collections - DiSSCo, Leiden, Netherlands
¶ DiSSCo, Leiden, Netherlands

Corresponding author: Rajapreethi Rajendran (rajapreethi.rajendran@senckenberg.de)

Reviewed v 1

Academic editor: Laurence Livermore

Abstract

The Distributed System for Scientific Collections (DiSSCo) is a research infrastructure to integrate European natural science collections (NSCs) digitally. The aim is to facilitate and enhance the access, management and analysis of collection assets in one unified digital collection. The Machine Annotation Services (MAS) are essential components of DiSSCo's Digital Specimen Architecture (DSArch). These services automate the annotation of digital objects to enable labelling and categorisation of NSC's digital assets.

To further advance this, a Machine Learning as a Service (MLaaS) approach was developed which provides researchers with the access to pre-trained machine-learning models for complex tasks, such as instance segmentation and morphological analysis of datasets. MLaaS enhances the DiSSCo's scalability and flexibility and allows the integration of machine-learning tools in close alignment with the FAIR (Findable, Accessible, Interoperable, Reusable) principles.

This study employs DiSSCO's MLaaS framework for the quantitative analysis of herbarium specimens. Machine-learning models, such as Mask R-CNN and YOLO11, are comparatively applied to detect and generate the pixel-level masks of plant organs in herbarium sheets. Subsequently, these models are used to reconstruct the scale in the herbarium sheet and to calculate the surface area of identified plant organs.

The determination of quantitative characteristics of plant specimens, such as measuring leaf area or the timestamp of the floral transition, opens up herbarium data for reuse in the large prognosis platforms currently developed in the framework of the Common European Data Spaces. In this way, plant trait data mobilised from natural science collections can improve the predictive capability of the vegetation model components of climate-related data spaces.

© Rajendran R et al. This is an open access article distributed under the terms of the Creative Commons Attribution License (CC BY 4.0), which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

Keywords

Digital Specimen Architecture, plant organ detection, quantitative traits, deep learning, DiSSCo, image processing, instance segmentation, Mask R-CNN, YOLO11, Common European Data Spaces

Introduction

The loss of biodiversity in the anthropocene, intensified by climate change, has a significant impact on human societies by reducing the benefits - designated as Ecosystem Services - that humans derive from ecosystems and environment (Pörtner et al. 2021).

To address this critical challenge, natural science collections, including in particular their enriched and annotated digital representations, have a pivotal role for assessment and analysis of the current and future biodiversity loss by providing the fundamental baseline data collection that reflects the actual and past dynamics of marine, freshwater and terrestrial biodiversity. The Distributed System of Scientific Collections (DiSSCo) is a European Research Infrastructure (RI) encompassing over 300 collecting institutions (DiSSCo 2025). A major aim of DiSSCo's infrastructure is thus to open up specimen data and make it widely reusable in compliance with the FAIR Principles (Koureas et al. 2023), preparing data for both discovery by humans and for autonomous processing by machines (i.e. machine-actionability, Jacobsen et al. (2020)).

To achieve this machine-actionability, DiSSCo developed its core data model, the Digital Specimen, in close alignment with the approach of FAIR Digital Objects (FDOs) to represent the physical specimen also in the digital domain (Islam et al. 2020, Hardisty et al. 2022b): A Digital Specimen encompasses or persistently links to all information artefacts (Ceuster and Smith 2015), which are about this physical specimen, such as sequence data, images, chemical measurements or taxonomic determinations. In this way, an object-centred and machine-interpretable representation of a specimen and its connections is realised, which makes independent operations on this digital object by machines possible, for example, machine-learning-based annotation of scanned images and extraction of associated data such as information on labels (Hardisty et al. 2022).

The wider objective of establishing a "Machine Learning as a Service" (MLaaS, Grieb et al. (2021)) framework for DiSSCo is to mobilise traceable geo- and biodiversity data for the large analysis and prognosis infrastructures set up currently in the context of the so-called Common European Data Spaces (CEDS, Scerri et al. (2022)). The CEDS are European Union-led initiatives to enable secure, interoperable data sharing across sectors and borders to enable research, innovation and policy-making. Data Spaces related to the European Green Deal are particularly suited for the integration of biodiversity data, involving the EU-flagship initiative Destination Earth and B-Cubed -

Biodiversity Building Blocks for Policy (Penninga et al. 2021, Groom et al. 2023, Puechmaille et al. 2025). A closer alignment of the natural science collections with the CEDS could thus significantly increase discoverability, availability and subsequent reuse of collection data. Building on that, the objective of the present study is the functional integration of a Machine Annotation Service (MAS) that enables quantitative determinations of morphological parameters into DiSSCo's Digital Specimen infrastructure (DSArch, Leeflang et al. (2022)).

The paper is further organised as follows: In the next section, we outline the components developed for the MAS. Afterwards, we present the results achieved using the framework with Senckenberg – Leibniz Institution for Biodiversity and Earth System Research's herbarium collection. The final section concludes the paper detailing directions for further developments.

Methods

Model selection

Building on previous research on detecting and annotating the plant organs from digitised herbarium scans (Younis et al. 2018, Younis et al. 2020b), we present now an extended approach involving an advanced segmentation technique to facilitate detailed quantitative analysis of morphological traits. The aim of the segmentation method is to subdivide images into objects or regions. Fundamentally, there are two types of segmentation: semantic and instance segmentation (Long et al. 2015, He et al. 2017). Semantic segmentation labels pixels with a class without distinguishing them into individual objects, whereas the instance segmentation labels each pixel and differentiates the individual objects of each class. Instance segmentation, used in the present context, enables the model to detect the plant organs and segment each organ individually. This is particularly beneficial for distinguishing closely-positioned or overlapping specimen organs, such as leaves, stems, fruits, seeds and flowers. Within the framework of this study, Mask R-CNN and YOLO11 were chosen as instance segmentation models. These models are based, as further explained below, on different architecture approaches: Mask R-CNN is a two-stage model, while YOLO11 is single stage. In the initial stage of the Mask R-CCN classification, it identifies regions of interest in an image and, in the second stage, it utilises local features around these proposed regions to determine contained objects. On the other hand, single-stage detectors, like YOLO, employ a fully convolutional approach to gridded regions of an image for the simultaneous prediction of bounding boxes and segmentation in a single pass (Redmon et al. 2016).

Plant organ segmentation dataset preparation

As in the aforementioned previous studies (Younis et al. 2020b), the dataset used for training plant organ segmentation consists of 652 images of herbarium scans from the

Muséum national d'Histoire naturelle's (MNHN) vascular plant collection (Le Bras et al. 2017, Younis et al. 2020a). These images are openly accessible through the Global Biodiversity Information Facility (GBIF) portal (MNHN and Chagnoux 2020).

Out of 652 images, 497 images were used for training and 155 images were used for testing. As shown in Table 1, the images were annotated into six different categories: leaf, stem, flower, fruit, root and seed. These images were previously annotated for object detection in plant organs. The training subset has 15486 annotations and the testing subset has 4137 annotations.

For the instance segmentation task, the annotations from the previous study were further refined to generate detailed pixel-level masks for each plant organ using the Segment Anything Model (SAM, Kirillov et al. (2023)). The files containing the annotations were parsed to extract the bounding box information, which was then passed to the SAM model along with the corresponding original images.

The SAM model iteratively segmented the plant organs for each bounding box and the resulting segmentations were combined for each image. Next, the masks were used to prepare instance segmentation annotations by uniquely identifying and labelling each organ type across all images (Fig. 1). This approach facilitates the reuse of the existing dataset and annotations were produced suitable for MASK R-CNN training. Additionally, the dataset was reformatted to comply with YOLO11's input specifications and allows precise segmentation of individual features in the scans and supports more detailed morphological analysis.

Scale training dataset preparation

The model was further independently trained to detect the scale in the digitised herbarium sheets and to calculate the surface area of plant organs in the digital herbarium sheet. The training dataset consists of 163 annotated images. Amongst these, 32 images are sourced from the Senckenberg herbarium dataset (Younis et al. 2020a), 24 images from the vascular plants collection at the Herbarium of the Muséum national d'Histoire naturelle, Paris (MNHN and Chagnoux 2020) and an additional 107 images from the GBIF to increase variability in the dataset for the training for scale detection. Of the total, 124 images are used for training and 39 for testing. A dataset containing the subset of images included from the Senckenberg Herbarium is available (Rajendran et al. 2025).

Model training and testing

The Mask R-CNN model (He et al. 2018) is trained using PyTorch (Paszke et al. 2019b) to perform instance segmentation on digitised herbarium scans. Out of 652 images, 497 annotated images were used for training and 155 images were used for testing. Data transformations, such as random horizontal and vertical flips, colour jitter, scaling and rotation, were applied during training to introduce variability to improve the model's generalisation.

The Mask R-CNN model used a ResNet-50 backbone to balance feature extraction quality and computational efficiency for the segmentation of high-resolution herbarium images. Training strategies, such as early stopping, were implemented by comparing mean Average Precision (mAP) of the current epoch and previous epoch, involving a patience threshold to avoid overfitting (Ying 2019). Additionally, the model was customised with a modified Region Proposal Network (RPN) anchor generator and adjusted Region Of Interest (ROI) heads to improve performance on the dataset (He et al. 2017).

The model performance was evaluated with metrics, such as precision and recall, for object detection and segmentation. To test the model generalisation, inference was conducted on the images from the Frankfurt Senckenberg Herbarium dataset (Younis et al. 2020a).

The YOLO 11x-seg (Jocher et al. 2023) segmentation model was trained with the same dataset used for MASK R-CNN training (Jocher et al. 2023). The default input image size for the YOLO model training is 640 × 640 pixels, but in this study, a resolution of 1064 × 1064 pixels was used, since it was the largest resolution that could fit into the GPU memory during training. The training strategies, such as early stopping of training, were applied in the same way as in the previous case.

Detection of scale and organ surface area calculation

The measurement of surface area of plant organs is an essential component of morphometric analysis in biodiversity studies (Hodač et al. 2024). To calculate the surface area of a plant organ in the herbarium sheet, the scales need to be detected and the surface area in pixels needs to be calculated and converted into absolute values, which are units of measurement such as square millimetres or square centimetres. This is accomplished through a combination of deep learning techniques for scale detection and Optical Character Recognition (OCR) for extraction of numerical values.

Scale detection

The Mask R-CNN and YOLO11 models were again employed for the detection of scales in the herbarium images. The surface area of plant organs was subsequently calculated. We used 167 images, of them 124 images for training and 43 for testing.

Once the scale was detected by the model, Tesseract OCR was employed to extract the numerical values from the scale present in the images (Smith 2007).

Numerical extraction using Tesseract OCR

Tesseract (Smith 2007) is a widely used optical character recognition tool to identify text in the images. Image preprocessing techniques, such as rotation of scales to correct orientation, greyscale conversion and bilateral filtering, are applied to enhance the text visibility. Following the preprocessing, the Tesseract OCR was utilised on the scales to

extract the numerical values from the scale. The digits with a confidence level above 75% and in the range of numbers from 0 to 10 are filtered and sorted to identify the consecutive sequences. The pixel distances between adjacent digits are calculated and averaged with the detected numbers to determine the number of pixels corresponding to 1 cm. In cases where the consecutive digits are not detected, the closest pair of digits is used to estimate the pixel distance. The character recognition accuracy is computed by comparing the detected digits with a ground truth reference. Once the pixel distances between consecutive digits on the scale have been calculated, the next step is to establish the conversion factor that relates the relative pixel measurements to absolute measurement in centimetres. The pixel-to-centimetre conversion factor is computed as follows:

$$\text{one cm in pixels} = \frac{\text{Pixel distance between digits A and B}}{\text{Difference between digit B and digit A}}$$

where

- **Pixel Distance Between Digit A and Digit B** is the pixel distance between the detected digits;
- **Difference Between Digit A and Digit B** is the difference between the actual values of the digits.

Surface area calculation of detected plant organs

With the pixel-to-centimetre conversion factor established, the next step is to calculate the surface area of the detected plant organs. The plant organ surface area is calculated as follows:

$$\text{Plant organ surface area } (\text{cm}^2) = \frac{\text{The sum of pixels in the segmented organs}}{(\text{Calculated pixels present in 1cm})^2}$$

where

- **The sum of pixels in the segmented organs** refers to the total number of pixels in the segmented organ regions;
- **Calculated pixels present in 1 cm** is the number of pixels that represent 1 cm on the image, estimated using the detected scale bar.

This additional functionality of calculating surface area supports detailed morphometric analysis of herbarium specimens.

Machine Learning as a Service (MLaaS)

Both models for plant organ segmentation and surface area calculation are then hosted as a single Machine Annotation Service (MAS) to streamline the process of extracting and analysing the herbarium data within the DiSSCo platform. Fig. 2 further illustrates the process flow and communication between the DiSSCo Digital Specimen architecture and the plant organ segmentation MAS.

In DSArch, a service request for a MAS is requested through the user front-end DiSSCover (https://sandbox.dissco.tech) on the digital media. This request adds a message in DiSSCover's Message Broker, which triggers scheduling of a MAS. The corresponding MLaaS APIs are hosted via Uvicorn (Trylesinski and Christie 2019), a high-performance asynchronous python web server designed for efficient handling of API requests on a remote virtual machine (VM) at Senckenberg. Within the scope of our specific use case, the API receives image URLs as the input from the Message Broker through the HTTP POST requests. The API then adds the message to the processing queue, which forwards the message to a web socket service, hosted on a server for performing plant organ segmentation.

Subsequently, the model performs plant organ segmentation and scale detection on the image and sends the output information comprising bounding box coordinates, class labels, confidence scores and area in pixels. Both the pixel-to-centimetre conversion ratio and an area calculation in cm² are returned to the Uvicorn API, which relays the results to the DiSSCo infrastructure. These processed data are then structured into an annotation event complying to DiSSCo's open Digital Specimen (openDS, Addink and Hardisty (2020)) specification. The message is published back to DiSSCo's Core architecture by placing it on the Message Broker, making it in this way available for downstream applications.

Fig. 3 provides a detailed visualisation of the annotations which allows users to explore the extracted information by hovering over specific segments. This includes details, such as the type of organ, confidence score, segmented polygon coordinates, area in pixels, pixel-to-centimetre conversion ratios and calculated areas in square centimetres for each segmented region. Fig. 3 demonstrates the integration of MAS within the DiSSCover platform highlighting its ability to generate enriched and standardised annotations in compliance with openDS .

Results

MASK R-CNN

Mask R-CNN was employed on 203 herbarium images from the Senckenberg collection (Younis et al. 2020a). Fig. 4 shows an example of object detection inference performed by Mask RCNN on a herbarium sheet.

When performing the inference on the Senckenberg images, the most recognised organs were leaves, stems and flowers. The identified regions of interest were then passed on to the segmentation process. In the segmentation process, the model generates segmentation masks for each identified organ and provides pixel-level masks as shown in Fig. 4.

Table 2: Statistics of the MASK R-CNN inference on 203 images. The original annotated herbarium sheet has the total plant organ count of 6693 where only 4022 organs were

detected with a confidence level of 50%. The model did not infer any root or seed due to the strong bias in the training data.

Fig. 5 represents the prediction recall curve of all plant organ detections and segmentations at an intersection of unit (IoU) thresholds of 0.5. The precision for the detection of leaves and stems is better than for flowers and fruits. The model effectively fails to predict roots and seeds.

Overall, the results suggest that the Mask R-CNN model moderately performs in identifying and segmenting the plant organs, achieving an average mean Average Precision (mAP) score of 0.211 for detection and 0.215 for segmentation. However, further addition of new images in the dataset could improve its performance of organ detection and segmentation.

YOLO11

The YOLO11 model was used with the same dataset of 203 herbarium images which was used for inference with Mask R-CNN. The model identified the distinct plant organs, such as leaves, stems, flowers, roots and seeds. In accordance with a similar study (Sapkota et al. 2024), we found that the YOLO11 model performs better than Mask R-CNN; YOLO11 was able to identify all the organs as shown in Fig. 6.

Table 3 below shows the statistics of YOLO11 inference on 203 images. The model was able to detect and segment all plant organs and, compared to Mask R-CNN, YOLO11 prediction count is higher.

Fig. 7 represents the prediction recall curve of all detected and segmented plant organs at IoU thresholds of 0.5. The YOLO11 model demonstrates improved performance compared to Mask R-CNN, achieving an average mean Average Precision (mAP) score of 0.373 for detection and 0.353 for segmentation.

Scale detection

In addition to the plant organ detection, the YOLO11 model integrates a scale detection component that successfully identifies the scales in 98% of the Senckenberg images. Fig. 8a shows an example of the scale detection.

Once the scale is detected, the model further uses the pixel-to-centimetre ratio to calculate the surface area for the plant organs in absolute values (cm^2). However, variations in scale standards across different organisations necessitate further training to improve scale detection.

Fig. 8b shows detected plant organs and scale with bounding boxes, class labels and confidence scores. Additionally, Fig. 8b illustrates the segmentation mask alongside the organ surface area calculations.

Suppl. material 1 provides a qualitative impression of the applicability of the method to other types of digitised specimen, in which scale detection was applied to images from Senckenberg's marine collections. The system successfully detected scale markers, recognised textual characters and identified morphological structures resembling stems and flowers (Suppl. material 1). These findings demonstrate the potential for extending our approach to applications beyond herbarium specimen analysis. However, successful detection and extraction of morphological traits requires retraining on the corresponding taxonomic groups.

Table 4 represents the statistics of scale detection, text recognition and character accuracy with Mask R-CNN and YOLO11 on 203 images. Although the pipeline of text recognition is the same for both models, there is difference in optical character accuracy due to the number of detected digits. The resulting bounding coordinates have small pixel variation due to the internal architecture of the models. The YOLO11 uses the single shot object detection where the images are passed one time. Mask R-CNN, on the other hand, uses two-stage object detection where the first pass proposes the region of interest and the second pass refines this prediction. The architecture difference impacts the resulting bounding box coordinates and results in a small offset pixel difference on pixel calculation between the models. This slight variation in bounding box calculation influences the OCR performance, as even minimal pixel shifts can affect character segmentation and thereby following recognition accuracy.

The output data then sent to DSArch includes the bounding box of detected organs, class, confidence score, area in pixels, pixel length of one centimetre, area in cm^2 and polygon data.

Discussion

The plant organ segmentation MAS in the DiSSCo architecture has enriched the Digital Specimen data by providing detailed morphological information. Compared to the Mask R-CNN model with a ResNet-50 backbone, YOLO11 has shown a reliable performance in identifying plant organs with computational efficiency. While YOLO11 shows more robust overall detection performance compared to Mask R-CNN, it struggles with small object detection. This highlights the need for further enhancements to improve robustness. The performance can possibly be improved using artificially debiased training data (Kim et al. 2021) and implementing a two-tier IIIF-based approach, starting with a rough initial detection and then using high-resolution image tiles to better detect small objects. The addition of scale detection is an advancement to convert the pixel measurements into real-world units. In this way, NSC's data, mobilised from DiSSCo's collections, can effectively contribute to application areas, such as the assessment of plants functional trait diversity and the biogeography of plant life-cycles (Díaz et al. 2015, Perez et al. 2020 , Poppenwimer et al. 2023).

The pipeline integrated with the Uvicorn API and WebSocket services provides a scalable asynchronous framework for managing high-throughput data streams in line

with DiSSCo's FAIR principles for the accessible and reusable data. This flexibility allows it to be extended for other datasets and segmentation tasks which promises broader applications in biodiversity informatics. Some challenges emerged during the implementation particularly on scale detection due to varying types of scale present in the different herbarium sheets. The scale markings are minute which causes difficulties for OCR systems to interpret the values. Additionally, the presence of various scale types further complicates the process as the OCR system is not optimised for all variations in character formats or standards. To address these issues, different scale formats need to be included in scale detection. The pipeline establishes a strong foundation for data processing to extract additional information on herbarium sheets with future directions to further enhance the scale detection and to adapt batch processing of herbarium sheets for faster real-time performance.

As part of the DiSSCo Transition Project (Koureas et al. 2024), we are also developing comprehensive MAS-related documentation, including service policies, usage guidelines, a MAS policy and a template service-level agreement (SLA). These efforts aim to ensure clarity, reliability, quality and sustainability of MAS offerings within the broader DiSSCo infrastructure.

To facilitate reproducibility and further research in plant organ segmentation, we published an annotated sample dataset including digitised herbarium specimens from Senckenberg's collection (Rajendran et al. 2025). This dataset includes organ-level segmentation masks and scale bar detection annotations and enables training, evaluation and comparison of similar models.

In summary, this study aims to demonstrate a potential pathway for the further societal valorisation of collection data through DiSSCo's Digital Specimen Architecture, particularly by mobilising high-quality datasets extracted from collection data for societal decision-making and action in the context of the Common European Data Spaces.

Acknowledgements

We thank Anke Penzlin, Andreas Allspach, Moritz Sonnewald, Alexander Knorrn, André Freiwald, Kristina Hopf, Stefan Dressler (†) and Marco Schmidt for the provision of data from Senckenberg's collections.

Funding program

Funded by the European Union. Views and opinions expressed are, however, those of the author(s) only and do not necessarily reflect those of the European Union or the European Research Executive Agency. Neither the European Union nor the granting authority can be held responsible for them. DiSSCo Transition grant agreement ID: 101130121, https://doi.org/10.3030/101130121.

Conflicts of interest

The authors have declared that no competing interests exist.

References

- Addink W, Hardisty A (2020) 'openDS' – Progress on the New Standard for Digital Specimens. Biodiversity Information Science and Standards 4 https://doi.org/10.3897/biss.4.59338
- Ceuster W, Smith B (2015) Aboutness: towards foundations for the Information Artifact Ontology. In: Ceuster W, Smith B (Eds) International Conference on Biomedical Ontology. URL: https://api.semanticscholar.org/CorpusID:15812859
- Díaz S, Kattge J, Cornelissen JC, Wright I, Lavorel S, Dray S, Reu B, Kleyer M, Wirth C, Colin Prentice I, Garnier E, Bönisch G, Westoby M, Poorter H, Reich P, Moles A, Dickie J, Gillison A, Zanne A, Chave J, Joseph Wright S, Sheremet'ev S, Jactel H, Baraloto C, Cerabolini B, Pierce S, Shipley B, Kirkup D, Casanoves F, Joswig J, Günther A, Falczuk V, Rüger N, Mahecha M, Gorné L (2015) The global spectrum of plant form and function. Nature 529 (7585): 167-171. https://doi.org/10.1038/nature16489
- DiSSCo (2025) Distributed System of Scientific Collections. https://www.dissco.eu/. Accessed on: 2025-8-01.
- Grieb J, Weiland C, Hardisty A, Addink W, Islam S, Younis S, Schmidt M (2021) Machine Learning as a Service for DiSSCo's Digital Specimen Architecture. Biodiversity Information Science and Standards 5 https://doi.org/10.3897/biss.5.75634
- Groom Q, Abraham L, Adriaens T, Breugelmans L, Clarke D, Fernández M, Hendrickx L, Hui C, Kumschick S, Martini M, McGeoch M, Metodiev T, Miller J, Oldoni D, Pereira H, Preda C, Robertson T, Rocchini D, Seebens H, Teixeira H, Trekels M, Wilson JR, Yovcheva N, Zengeya T, Desmet P (2023) B-Cubed: Leveraging Analysis-Ready Biodiversity Datasets and Cloud Computing for Timely and Actionable Biodiversity Monitoring. Biodiversity Information Science and Standards 7 https://doi.org/10.3897/biss.7.110734
- Hardisty A, Brack P, Goble C, Livermore L, Scott B, Groom Q, Owen S, Soiland-Reyes S (2022a) The Specimen Data Refinery: A Canonical Workflow Framework and FAIR Digital Object Approach to Speeding up Digital Mobilisation of Natural History Collections. Data Intelligence 4 (2): 320-341. https://doi.org/10.1162/dint_a_00134
- Hardisty A, Ellwood E, Nelson G, Zimkus B, Buschbom J, Addink W, Rabeler R, Bates J, Bentley A, Jos&eacute AB, Hansen S, Macklin J, Mast A, Miller J, Monfils A, Paul D, Wallis E, Webster M (2022b) Digital Extended Specimens. Enabling an Extensible Network of Biodiversity Data Records as Integrated Digital Objects on the Internet, BioScience 72 (ue 10). https://doi.org/10.1093/biosci/biac060
- He K, Gkioxari G, Dollar P, Girshick R (2017) Mask R-CNN. 2017 IEEE International Conference on Computer Vision (ICCV) https://doi.org/10.1109/iccv.2017.322
- He K, Gkioxari G, Dollar P, Girshick R (2018) Mask R-CNN. IEEE transactions on pattern analysis and machine intelligence 42 (2): 386-397. https://doi.org/10.1109/TPAMI.2018.2844175

- 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
- Islam S, Hardisty A, Addink W, Weiland C, Glöckler F (2020) Incorporating RDA Outputs in the Design of a European Research Infrastructure for Natural Science Collections. Data Science Journal 19 https://doi.org/10.5334/dsj-2020-050
- Jacobsen A, de Miranda Azevedo R, Juty N, Batista D, Coles S, Cornet R, Courtot M, Crosas M, Dumontier M, Evelo C, Goble C, Guizzardi G, Hansen KK, Hasnain A, Hettne K, Heringa J, Hooft RW, Imming M, Jeffery K, Kaliyaperumal R, Kersloot M, Kirkpatrick C, Kuhn T, Labastida I, Magagna B, McQuilton P, Meyers N, Montesanti A, van Reisen M, Rocca-Serra P, Pergl R, Sansone S, da Silva Santos LOB, Schneider J, Strawn G, Thompson M, Waagmeester A, Weigel T, Wilkinson M, Willighagen E, Wittenburg P, Roos M, Mons B, Schultes E (2020) FAIR Principles: Interpretations and Implementation Considerations. Data Intelligence 2: 10-29. https://doi.org/10.1162/dint_r_00024
- Jocher G, Qiu J, Chaurasia A (2023) Ultralytics YOLO. Version 8.0.0. URL: https://github.com/ultralytics/ultralytics
- Kim E, Lee J, Lee J, Lee J, Choo J (2021) Learning Debiased Representation via Disentangled Feature Augmentation. Neural Information Processing Systems https://doi.org/10.48550/arXiv.2107.01372
- Kirillov A, Mintun E, Ravi N, Mao H, Rolland C, Gustafson L, Xiao T, Whitehead S, Berg A, Lo W, Dollár P, Girshick R (2023) Segment Anything. 2023 IEEE/CVF International Conference on Computer Vision (ICCV) https://doi.org/10.1109/iccv51070.2023.00371
- Koureas D, Livermore L, Alonso E, Addink W, Alves MJ, Casino A, Curral L, Enghoff H, Guiraud M, Hardy H, Hoffmann J, Landel S, Paleco C, Petersen M, Scory S, Smith V, Weiland C, Wesche K, Woodburn M (2023) DiSSCo Prepare Project: Increasing the Implementation Readiness Levels of the European Research Infrastructure. Research Ideas and Outcomes 9 https://doi.org/10.3897/rio.9.e113906
- Koureas D, Livermore L, Addink W, Alonso E, Alonso J, Casino A, Dusoulier F, Ferreira V, Grieb J, Groom Q, Islam S, Kõljalg U, Lymer G, Marhold K, Paleco C, Pijls S, Scory S, Scott B, Weiland C, Worley K (2024) DiSSCo Transition Abridged Grant Proposal. Research Ideas and Outcomes 10 https://doi.org/10.3897/rio.10.e118241
- Le Bras G, Pignal M, Jeanson M, Muller S, Aupic C, Carré B, Flament G, Gaudeul M, Gonçalves C, Invernón V, Jabbour F, Lerat E, Lowry P, Offroy B, Pimparé EP, Poncy O, Rouhan G, Haevermans T (2017) The French Muséum national d'histoire naturelle vascular plant herbarium collection dataset. Scientific Data 4 (1). https://doi.org/10.1038/sdata.2017.16
- Leeflang S, Addink W, Theocharides S (2022) Human and Machine Working Together towards High Quality Specimen Data: Annotation and Curation of the Digital Specimen. Biodiversity Information Science and Standards 6 https://doi.org/10.3897/biss.6.90987
- Long J, Shelhamer E, Darrell T (2015) Fully convolutional networks for semantic segmentation. 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR) https://doi.org/10.1109/cvpr.2015.7298965
- MNHN, Chagnoux S (2020) The vascular plants collection (P) at the Herbarium of the Muséum national d'Histoire naturelle (MNHN - Paris). 69.186. GBIF. URL: https://doi.org/10.15468/nc6rxy

- Paszke A, Gross S, Massa F, Lerer A, Bradbury J, Chanan G, Killeen T, Lin Z, Gimelshein N, Antiga L, Desmaison A, Köpf A, Yang E, DeVito Z, Raison M, Tejani A, Chilamkurthy S, Steiner B, Fang L, Bai J, Chintala S (2019) PyTorch: An Imperative Style, High-Performance Deep Learning Library. arXiv https://doi.org/10.48550/arxiv.1912.01703
- Penninga F, Lutz M, Minghini M, Cetl V, Robbrecht J (2021) INSPIRE, a public sector contribution to the European green deal data space - A vision for the technological evolution of Europe's spatial data infrastructures for 2030. Publications Office of the European Union URL: https://data.europa.eu/doi/10.2760/8563
- Perez T, Rodriguez J, Mason Heberling J (2020) Herbarium-based measurements reliably estimate three functional traits. American Journal of Botany 107 (10): 1457-1464. https://doi.org/10.1002/ajb2.1535
- Poppenwimer T, Mayrose I, DeMalach N (2023) Revising the global biogeography of annual and perennial plants. Nature 624 (7990): 109-114. https://doi.org/10.1038/s41586-023-06644-x
- Pörtner H, Scholes R, Agard J, Archer E, Arneth A, Bai X, Barnes D, Burrows M, Chan L, Cheung WL(, Diamond S, Donatti C, Duarte C, Eisenhauer N, Foden W, Gasalla M, Handa C, Hickler T, Hoegh-Guldberg O, Ichii K, Jacob U, Insarov G, Kiessling W, Leadley P, Leemans R, Levin L, Lim M, Maharaj S, Managi S, Marquet P, McElwee P, Midgley G, Oberdorff T, Obura D, Osman Elasha B, Pandit R, Pascual U, Pires AF, Popp A, Reyes-García V, Sankaran M, Settele J, Shin Y, Sintayehu D, Smith P, Steiner N, Strassburg B, Sukumar R, Trisos C, Val A, Wu J, Aldrian E, Parmesan C, Pichs-Madruga R, Roberts D, Rogers A, Díaz S, Fischer M, Hashimoto S, Lavorel S, Wu N, Ngo H (2021) Scientific outcome of the IPBES-IPCC co-sponsored workshop on biodiversity and climate change. Zenodo https://doi.org/10.5281/zenodo.4659158
- Puechmaille D, Schick M, Saulyak B, Dillmann M, Wolf L (2025) Destination Earth Data Lake unlocking Big Earth Data processing. EGU General Assembly 2024. https://doi.org/10.5194/egusphere-egu24-11115
- Rajendran R, Marco S, Claus W, Jonas G (2025) Plant organ segmentation images and annotations on digitized herbarium scans. Senckenberg Society for Nature Research https://doi.org/10.12761/fj4m-zr97
- Redmon J, Divvala S, Girshick R, Farhadi A (2016) You Only Look Once: Unified, Real-Time Object Detection. 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)779-788. https://doi.org/10.1109/cvpr.2016.91
- Sapkota R, Ahmed D, Karkee M (2024) Comparing YOLOv8 and Mask R-CNN for instance segmentation in complex orchard environments. Artificial Intelligence in Agriculture 13: 84-99. https://doi.org/10.1016/j.aiia.2024.07.001
- Scerri S, Tuikka T, de Vallejo IL, Curry E (2022) Common European Data Spaces: Challenges and Opportunities. Data Spaces337-357. https://doi.org/10.1007/978-3-030-98636-0_16
- Smith R (2007) An Overview of the Tesseract OCR Engine. Ninth International Conference on Document Analysis and Recognition (ICDAR 2007) Vol 2629-633. https://doi.org/10.1109/icdar.2007.4376991
- Trylesinski M, Christie T (2019) Uvicorn. URL: https://github.com/encode/uvicorn
- Ying X (2019) An Overview of Overfitting and its Solutions. Journal of Physics: Conference Series 1168 https://doi.org/10.1088/1742-6596/1168/2/022022

- 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
- Younis S, Schmidt M, Dressler S (2020a) Plant organ detections and annotations on digitized herbarium scans [dataset]. *PANGAEA*. Release date: 2020-8-27. URL: https://doi.org/10.1594/PANGAEA.920895
- Younis S, Schmidt M, Weiland C, Dressler S, Seeger B, Hickler T (2020b) Detection and annotation of plant organs from digitised herbarium scans using deep learning. Biodiversity Data Journal 8 https://doi.org/10.3897/bdj.8.e57090

Figure 1.

Mask generated by the SAM (Segment Anything Model) for a herbarium specimen scan of *Rubus pottianus* H.E. Weber. The figure shows the segmentation mask output produced by the SAM model for a digitised herbarium sheet labelled FR-0030810 (CETAF ID: https://id.senckenberg.de/object/sesam-353465) from the training dataset (Rajendran et al. 2025).

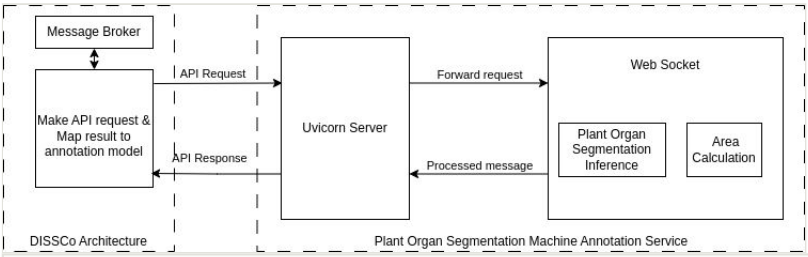

Figure 2.

Schematic overview illustrating the information flow between DiSSCo core architecture and the MAS workflow deployed at Senckenberg. (i) Message Broker, which handles asynchronous communication; (ii) MLaaS (Machine Learning as a Service), API hosted via Uvicorn, serving as the inference interface; and (iii) Machine Annotation Service (MAS) modules responsible for task orchestration. The architecture diagram highlights how herbarium image data and metadata are processed, annotated and returned to the DiSSCo system in a scalable and modular fashion.

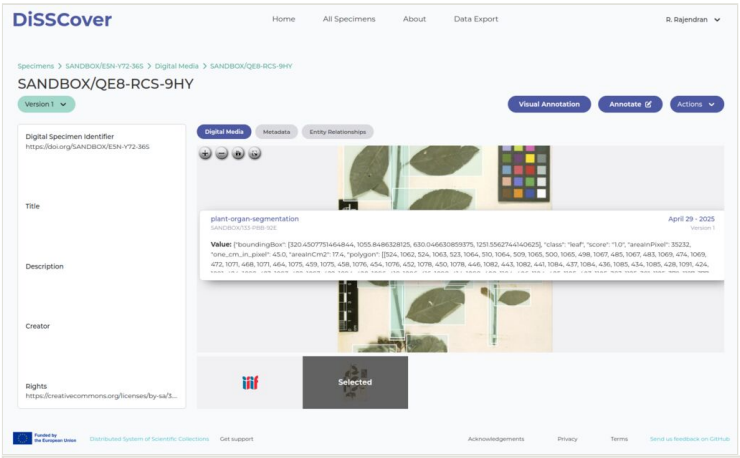

Figure 3.

Annotated herbarium specimen sheet processed through the DiSSCover platform. The figure presents an annotated herbarium image processed using the DiSSCover pipeline, which includes detection of plant organs and segmentation. The visual overlays include bounding boxes, class labels, confidence scores from the prediction model, area in pixels, pixel-to-centimetre conversions and polygon coordinates for each detected organ.

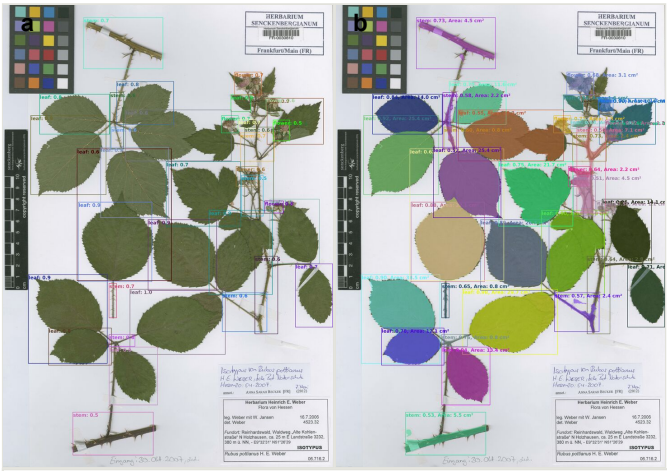

Figure 4.

Plant organ detection and segmentation using the Mask R-CNN model:

a: Plant organ detection: Detection output showing bounding boxes, confidence score and class labels for distinct plant organs on the specimen from Fig. 1;
b: Plant organ segmentation: Segmentation results for the same specimen illustrating pixel-wise masks, confidence score, class labels for each detected plant organ.

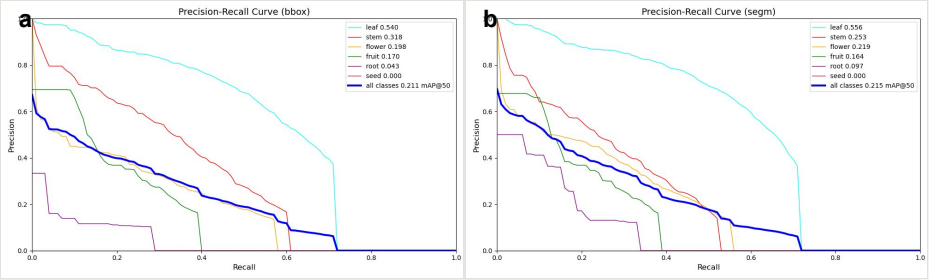

Figure 5.

Precision–Recall (PR) curves evaluating Mask R-CNN performance at 50% confidence threshold:

a: Object detection: PR curves for plant organ detection, shown both overall and per class, with precision and recall calculated at 50% confidence;
b: Segmentation: PR curves for plant organ segmentation presented overall and for each class at 50% confidence threshold.

ARPHA Preprints *Author-formatted, not peer-reviewed document posted on 29/05/2025*. DOI: https://doi.org/10.3897/arphapreprints.e160486

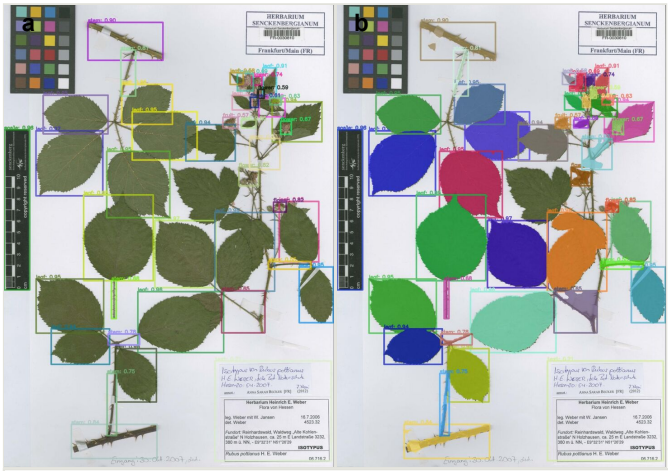

Figure 6.

Plant organ detection and segmentation using the YOLO11 model:

a: Plant organ detection: Detection output showing bounding boxes, confidence score and class labels for distinct plant organs on the specimen from Fig. 1;
b: Plant organ segmentation: Segmentation results for the same specimen illustrating pixel-wise masks, confidence score, class labels for each detected plant organ.

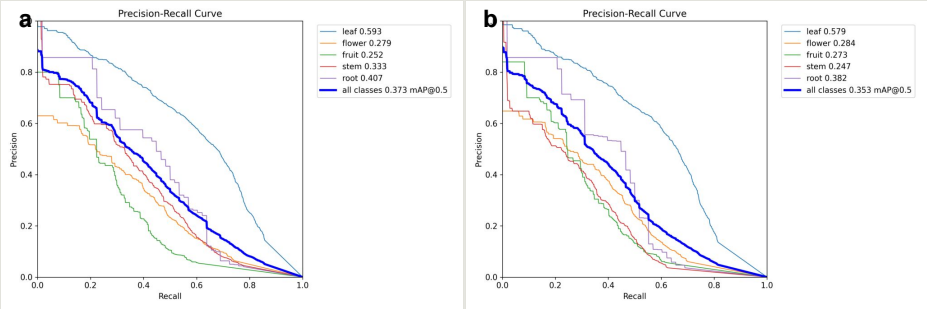

Figure 7.

Precision–Recall (PR) curves evaluating YOLO11 performance at 50% confidence threshold:

a: Object detection: PR curves for plant organ detection, showing both overall and per class, with precision and recall calculated at 50% confidence threshold;

b: Segmentation: PR curves for plant organ segmentation presented overall and for each class at 50% confidence threshold.

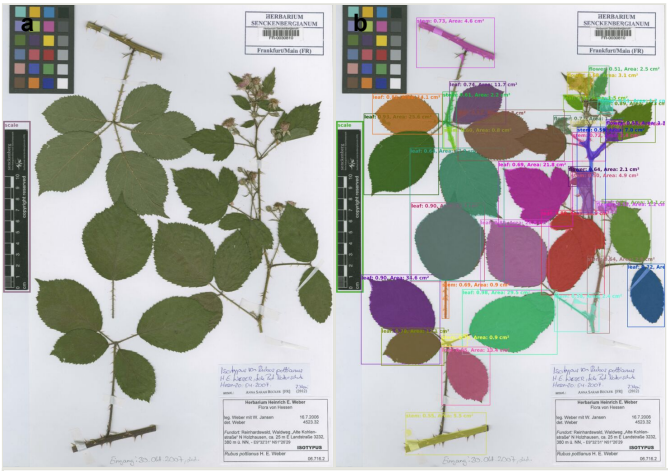

Figure 8.

Scale detection and area calculation in herbarium scans:

a: Scale detection: Detection of the scale bar on the specimen from Fig. 1, used to convert pixel measurements to absolute units (centimetres);
b: Area Calculation: Area calculation of segmented plant organs using the detected scale enabling measurement of organ size in absolute units.

Table 1.

The number of annotated bounding boxes and segmentation masks for each plant organ category is presented for both the training and testing subsets.

Category	**Training subset** (497 images)	**Testing subset** (155 images)	**Complete dataset** (652 images)
Leaf	7865	2051	9916
Stem	3315	961	4276
Flower	3179	763	3942
Fruit	1045	296	1341
Root	78	60	138
Seed	4	6	10
Total	**15486**	**4137**	**19623**

Table 2.

Plant organ counts in the inferred images compared to detected counts by the Mask R-CNN model at a 50% confidence threshold.

Category	Organ Count	Detected Count
Leaf	3362	3072
Stem	1063	782
Flower	1921	1407
Fruit	183	144
Root	117	77
Seed	47	5
Total	6693	5487

Table 3.

Plant organ counts of the inferred images compared to detected counts by the YOLO11 model at a 50% confidence threshold.

Category	Organ Count	Detected Count
Leaf	3362	3072
Stem	1063	782
Flower	1921	1407
Fruit	183	144
Root	117	77
Seed	47	5
Total	6693	5487

Table 4.

Performance metrics for scale detection and scale text counting and in addition overall average OCR character accuracy for both Mask R-CNN and YOLOv11.

Metrics	MASK R-CNN	YOLO11
Scale Detection Counter	203	202
Scale Text Counter	196	195
Overall Average OCR Character Accuracy	23.76%	57.16%

ARPHA Preprints *Author-formatted, not peer-reviewed document posted on 29/05/2025*. DOI: https://doi.org/10.3897/arphapreprints.e160486

Supplementary material

Suppl. material 1: Scale and Structure Recognition Beyond Herbarium Specimens

Authors: Rajapreethi Rajendran, Jonas Grieb, Claus Weiland, Anke Penzlin, Andreas Allspach, Moritz Sonnewald, Alexander Knorrn, André Freiwald, Kristina Hopf

Data type: Image

Brief description: The top figure represents *Galeoides decadactylus* (Bloch, 1795) with Senckenberg catalogue number SMF-39828 (this specimen is currently in process of ingestion into Senckenberg's Collection Management System), the scale was detected and all digits within the scales were successfully recognised. However, no plant organ like structure was detected in the image and, consequently, no area was calculated.

The second figure represents *Syngnathus acus* (Linnaeus, 1758) associated with CETAF ID: https://id.senckenberg.de/object/sesam-1710769. In this case, the specimen was incorrectly identified as the stem due to its structural similarity and the scale was detected and digits '6' and '8' within the scale regions were identified through optical character recognition (OCR). Pixel-wise segmentation is performed and area measurement was successfully computed.

In the third figure, the specimen shows *Pisa tetraodon* (Pennant, 1777), associated with CETAF ID: https://id.senckenberg.de/object/sesam-1710770. In this case, the specimen was incorrectly identified as flower due to its structural similarity and the scale was detected. Although scale was detected, no numerical digits were recognised by OCR and, as a result, area calculation was not conducted.

Download file (11.01 MB)