The Herbarium 2021 Half–Earth Challenge Dataset

Riccardo de Lutio
ETH Zürich
rdelutio@ethz.ch

Damon Little
New York Botanical Garden

Barbara Ambrose
New York Botanical Garden

Serge Belongie
Cornell Tech & Google Research

Abstract

Herbarium sheets present a unique view of the world's botanical history, evolution, and diversity. This makes them an all–important data source for botanical research. With the increased digitisation of herbaria worldwide and the advances in the fine–grained classification domain that can facilitate automatic identification of herbarium specimens, there are a lot of opportunities for supporting research in this field. However, existing datasets are either too small, or not diverse enough, in terms of represented taxa, geographic distribution or host institutions. Furthermore, aggregating multiple datasets is difficult as taxa exist under a multitude of different names and the taxonomy requires alignment to a common reference. We present the Herbarium Half–Earth dataset, the largest and most diverse dataset of herbarium specimens to date for automatic taxon recognition.

1. Introduction

Natural history collections, such as herbarium specimens, contain a plethora of information from phenotype to genotype. Each specimen is a snapshot in time and all together provide a history of plants on Earth since the first herbarium collections were made nearly 500 years ago [18]. Therefore, herbarium specimens are integral for understanding biodiversity and providing data to ameliorate the impacts of habitat loss and climate change [1, 5, 13].

Citizen science initiatives such as iNaturalist [10] and Pl@ntNet [12], have popularised species recognition as a challenging real–world classification task, with large imbalanced fine–grained datasets. Similarly using computer vision methods for the automatic classification of herbarium specimens is a well studied topic, [2, 3, 7, 14, 15, 16, 20, 21, 22, 23, 24, 26]. Many of these works focus on morphological trait recognition [3, 15, 16, 17, 20, 21, 26], while others focus on species recognition from leaves only [21, 23, 24].

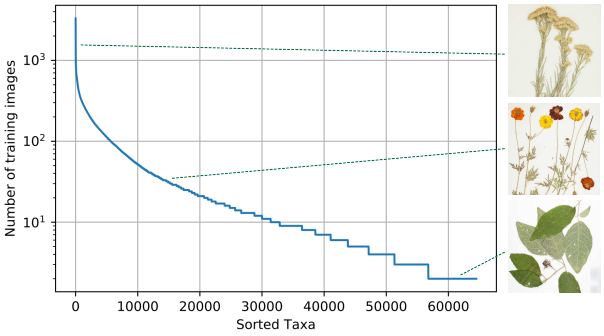

Figure 1. Distribution of training images per taxon. The Herbarium Half–Earth dataset is highly imbalanced. Featured taxa are from top to bottom: *Ericameria nauseosa* (Pall. ex Pursh) G.L. Nesom & G.I. Baird (Asteraceae), *Bidens sulphurea* (Cav.) Sch. Bip. (Asteraceae) and *Solanum rixosum* A.R. Bean (Solanaceae).

However existing datasets designed for computer vision approaches currently present some limitations. They are either small, targeted at specific taxa, only representative of a certain geographic region or coming from a single institution (see Tab. 1). With the Herbarium Half–Earth dataset, we aim to address all these limitations and present the largest and most diverse dataset of herbarium specimens for automatic taxon recognition to date.

Dataset	# Images	# Taxa	# Institutions	Geo. Range
Dillen et al. [7]	1'900	1'580	9	All Continents
Lorieul et al. [15]	163'233	7'782	1	Americas
Herbarium 255 [2]	11'071	255	1	Costa Rica
Herbarium 1K [2]	253'733	1'204	1	France
Herbarium 2019 [19]	46'000	680	1	Americas
Herbarium 2020	1'170'000	32'000	1	Americas
Herbarium 2021	2'500'000	64'500	5	Americas, Oceania and Pacific

Table 1. Summary of existing herbarium sheet datasets. Note that the Herbarium 2019 dataset focuses on the flowering plant family Melastomataceae, while the other datasets present a wider taxonomic diversity.

2. The Herbarium Half–Earth Dataset

The Herbarium Half–Earth dataset [1] includes more than 2.5M images of vascular plant specimens representing nearly 64,500 taxa from the Americas and Oceania.

The most exact labels are, in many cases, intraspecific (subspecies, varieties, forms, etc.) or nothospecies (hybrids) neither of which can be characterized as "species", thus we use the terms "taxon" and "taxa" as generic descriptors of taxonomic labels. In addition to labels for species–level and below, we also include labels at higher levels in the taxonomic hierarchy: family and order. This allows for experimentation with methods that address label hierarchy and label similarity. These labels may also be supplemented by more fine–grained estimates of difference among taxa available from other sources [11].

The images are provided by the New York Botanical Garden (NY), Bishop Museum (BPBM), Naturalis Biodiversity Center (NL), Queensland Herbarium (BRI), and Auckland War Memorial Museum (AK).

This dataset has a long tail; there are a minimum of three images per taxon (Fig. 1). However, some taxa can be represented by more than 100 images. This dataset only includes images of vascular plant—the group of plants that includes lycophytes, ferns, gymnosperms, and flowering plants (Fig. 2). The extinct forms of lycophytes are the major component of coal deposits, ferns are indicators of ecosystem health, gymnosperms provide major habitats for animals, and flowering plants provide almost all of our crops, vegetables, and fruits.

Figure 2. Example of images in the Herbarium Half–Earth dataset.

2.1. Dataset Challenges

The Herbarium 2021 Half–Earth dataset is challenging due to multiple reasons. First, of course, due to its large imbalance (Fig. 1), the imbalance factor for the dataset is 1,654.5. Second, the variation within species (Figs. 3) is high. Herbarium specimens can capture plants at different growth–stages (e.g., juvenile versus adult) or with different sets of plant parts (e.g., leaves and flowers versus leaves and fruit; see Fig. 3). In addition, the techniques used to press, dry, and mount specimens vary among collectors and collecting expeditions—these differences can change the appearance of specimens dramatically (e.g., collecting in alcohol often causes leaves to turn black). Arbitrary aesthetic decisions made while processing specimens can result in specimens that differ dramatically in appearance even though they are simply different parts of the same individual plant (Fig. 4). In a herbarium collection, every attempt to conserve dried specimens is made, but in practice older specimens become more fragile and suffer damage as they age leading to some specimens being less complete and more damaged than others. Third, the visual similarity among species can be high (Fig. 5). Finally, the diagnostic morphological features that botanists use to identify species are often very small and thus require a model that is able to handle high–resolution images and can focus on specific details [4, 22].

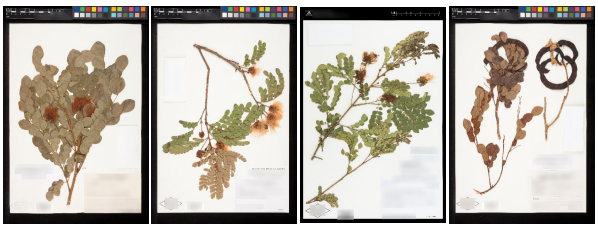

Figure 3. Example of visually different images corresponding to the same species: *Abarema brachystachya* (DC.) Barneby & J.W.Grimes (Fabaceae). The observed differences are primarily due to different reproductive stages: early flowering, late flowering, and fruit.

Figure 4. Different specimens of *Arbutus xalapensis* Kunth (Ericaceae) made from the same individual plant at the same time by the same collector using the same pressing, drying, and mounting protocol.

[1] https://github.com/visipedia/herbarium_comp

Figure 5. Example of visually similar images from different *Alyssum* species (Brassicaceae): *A. alyssoides* (L.) L., *A. desertorum* Stapf, *A. simplex* Rudolphi, *A. szovitsianum* Fisch. & C.A. Mey.

2.2. Data Preprocessing

In this section we give an overview of how we preprocessed the dataset. Figure 6 presents some example of herbarium sheets before and after the preprocessing steps.

2.2.1 Label Alignment

Herbarium specimens of the same taxon may have been labeled in various ways due to differences in the interpretation of taxon circumscriptions, nomenclatural changes, and/or errors. For example, over time *Pilosella piloselloides* (Vill.) Soják (Asteraceae) has been known by at least 526 different names [8]. To ameliorate this situation as much as possible, we have standardized the image labels to the Leipzig Catalogue of Vascular Plants (LCVP v1.0.2) [8]. Labels in our dataset have an LCVP status of either "accepted" or "unresolved".

The data exported from the institutional databases were first processed to find labels that exactly matched LCVP. For labels that did not precisely match, we then searched for long unambiguous partial matches to LCVP: the label was shortened by removing the rightmost word and then we searched for a match that produced only one LCVP output taxon; if no match was found, this was repeated until the label contained only two words. Labels that still did not unambiguously match LCVP, were matched using tre-agrep [25] allowing an increasing amount of mismatch (10–30% of label length; all weights were set to 1). Matches returned by tre-agrep were manually reviewed (8,430 labels passed manual review). Images with labels that could not be coerced into matching LCVP were excluded from the dataset (*c*. 73 thousand images).

2.2.2 Image Blurring

Herbarium specimens always have a hand–written or printed label on the sheet (usually lower right–hand corner), which includes information about the name of the taxon, the geographic location where it was collected, the date of collection, and the person or team of people who collected it. In addition, annotation labels are often added to the specimen to correct or update information on the original label—these are sequentially added in the empty space above the original label. Specimens often also have institutional labels or stamps indicating the herbarium in which the specimen is archived and a barcode label corresponding to an institutional database entry. Specimens may also include field tags with identification numbers attached directly to the plant. Images usually include color and measurement scales as well as institutional logos. All of these labels can of course, help identify the specimen, thus we blurred this information in the dataset in order to force models to learn about the plants themselves rather than the label text.

To detect these labels we used a pretrained EAST text detection model [27]. This model outputs bounding boxes around the detected text. We merged the bounding boxes that overlapped by a sufficient margin, and filtered out those that were too small. The resulting regions were then heavily blurred. We first applied a mean blur, then a single Gaussian blur with added noise, and finally used a smooth alpha map to blend into the original (Fig. 6). Finally, we excluded images from the dataset where more than 25% of the image was blurred, as we found those to be, in most cases, wrong predictions from our text detection model. We deliberately chose to tune the text detection model to have a high specificity, in order to avoid unnecessarily blurring parts of the plant. Even though, this means that there are images where part of the labels are missed by our blurring algorithm.

2.2.3 Image Resizing

Herbarium sheets are digitized as very high–resolution images to preserve as much of the detail as possible. A common image size is around 6000×4000. This is very large even for networks that are designed to work with higher resolutions. We have resized all images in the dataset to a larger dimension of 1000 (while preserving the aspect ratio), in order to make the overall size of the dataset more accessible.

2.2.4 Dataset Split

Our dataset contains images from 64,500 taxa at the species–level or below with 2,257,759 in the training set and 243,020 in the test set. The data has been split to obtain an approximately even number of images across taxa in the test set. In fact, we capped the maximum number of images per taxon to 10 for the test set. For taxa that have few images we did a 80%/20% split for training/test. Each category has a minimum of three images: at least one in the test set and two in the training set.

Figure 6. Example of images before and after preprocessing.

3. The Herbarium 2021 Half-Earth Challenge

The Herbarium 2021 Half-Earth Challenge is a competition hosted on Kaggle as part of 8^{th} workshop for Fine–Grained Visual Categorization at CVPR 2021. This is the third iteration of the Herbarium Challenge, in this section we give a brief description of the previous challenges.

The Herbarium 2019 Challenge [19, 14] focuses on the flowering plant family Melastomataceae. It contains 46,469 digitally imaged herbarium specimens representing 683 species. The Melastomataceae is a large family with 166 recognized genera and 5,892 species [8]. The overlap with the iNaturalist 2018 challenge dataset [10] is only 2 out of the 683 species in the Herbarium 2019 dataset.

The Herbarium 2020 Challenge dataset contains over 1M images representing over 32,000 plant species. This challenge focuses on vascular land plants of the Americas.

4. Baseline and Evaluation Metric

As a simple baseline we have trained a ResNet-50 [9] for 10 epochs. We split the training set in a stratified manner to create a hold-out validation set, thus the baseline was trained on 80% of the full training set. We used a balanced sampling strategy, so as to mitigate the impact of the imbalance on the classifier. We resized the images to 256×256 and used some standard data augmentations (small rotations, horizontal flips, color-jitter and finally center-crop to 224×224). We initialised the model with weights pretrained on ImageNet [6]. Finally we trained the model using the standard cross-entropy loss, a batch size of 32, a stochastic gradient descent with a learning rate of $1 \cdot 10^{-3}$ which is further reduced when a plateau is reached and a momentum factor of 0.9.

The evaluation metric for the Herbarium 2021 Half-Earth Challenge is the F_1 score:

$$F_1 = 2\frac{\mathrm{Pre} \cdot \mathrm{Rec}}{\mathrm{Pre} + \mathrm{Rec}}\,, \tag{1}$$

where Pre denotes the precision and Rec the recall. Our baseline achieves an F_1 score of 0.46 on the private test set of the competition. For comparison, the first place solution of the competition achieved an F_1 score of 0.76 on the private test set [2].

5. Conclusion

We presented the Herbarium Half-Earth dataset to enable the development of better automatic taxon recognition models. The development of models to automatically identify specimens will reduce the bottleneck of species identification and has the potential to advance biodiversity research at an unprecedented rate.

In the future, we would like to expand the dataset to include specimens collected world-wide. There are more than 35 million digitized specimens in electronic databases representing more than 80% of the known vascular plant diversity.

Acknowledgments We would like to thank our colleagues at the New York Botanical Garden (Damon Little, Kimberly Watson, Barbara Ambrose) and Kiat Chuan Tan from Google for their generous support in making this challenge possible. We are grateful for the participation of our collaborators at Auckland War Memorial Museum (Dhahara Ranatunga, Yumiko Baba), Bishop Museum (Melissa Tulig), Queensland Herbarium at Brisbane Botanic Gardens (Gillian Brown, Gordon Guymer, Andrew Franks), Naturalis Biodiversity Center (Jan Wieringa) and for contributing images to this dataset.

References

[1] Kellen M. Calinger, Simon Queenborough, and Peter S. Curtis. Herbarium specimens reveal the footprint of climate

[2] Results still unverified at the moment of writing.

change on flowering trends across north-central north america. *Ecology Letters*, 16(8):1037–1044, 2013. 1

[2] Jose Carranza-Rojas, Herve Goeau, Pierre Bonnet, Erick Mata-Montero, and Alexis Joly. Going deeper in the automated identification of herbarium specimens. *BMC Evolutionary Biology*, 17(1):181, 2017. 1

[3] Jonathan Y. Clark, David P. A. Corney, and H. Lilian Tang. Automated plant identification using artificial neural networks. In *2012 IEEE Symposium on Computational Intelligence in Bioinformatics and Computational Biology (CIBCB)*, pages 343–348, 2012. 1

[4] James S. Cope, David Corney, Jonathan Y. Clark, Paolo Remagnino, and Paul Wilkin. Plant species identification using digital morphometrics: A review. *Expert Systems with Applications*, 39(8):7562–7573, 2012. 2

[5] Charles C. Davis, Charles G. Willis, Bryan Connolly, Courtland Kelly, and Aaron M. Ellison. Herbarium records are reliable sources of phenological change driven by climate and provide novel insights into species' phenological cueing mechanisms. *American Journal of Botany*, 102(10):1599–1609, 2015. 1

[6] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. ImageNet: A large-scale hierarchical image database. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255, 2009. 4

[7] Mathias Dillen, Quentin Groom, Simon Chagnoux, Anton Güntsch, Alex Hardisty, Elspeth Haston, Laurence Livermore, Veljo Runnel, Leif Schulman, Luc Willemse, Zhengzhe Wu, and Sarah Phillips. A benchmark dataset of herbarium specimen images with label data. *Biodiversity Data Journal*, 7:e31817, 2019. 1

[8] Martin Freiberg, Marten Winter, Alessandro Gentile, Alexander Zizka, Alexandra Nora Muellner-Riehl, Alexandra Weigelt, and Christian Wirth. Lcvp, the leipzig catalogue of vascular plants, a new taxonomic reference list for all known vascular plants. *Scientific Data*, 7(1):416, 2020. 3, 4

[9] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 770–778, 2016. 4

[10] Grant Van Horn, Oisin Mac Aodha, Yang Song, Alexander Shepard, Hartwig Adam, Pietro Perona, and Serge J. Belongie. The inaturalist challenge 2017 dataset. 2018. 1, 4

[11] Yi Jin and Hong Qian. V.phylomaker: an r package that can generate very large phylogenies for vascular plants. *Ecography*, 42(8):1353–1359, 2019. 2

[12] Alexis Joly, Pierre Bonnet, Hervé Goëau, Julien Barbe, Souheil Selmi, Julien Champ, Samuel Dufour-Kowalski, Antoine Affouard, Jennifer Carré, Jean-François Molino, Nozha Boujemaa, and Daniel Barthélémy. A look inside the pl@ntnet experience. *Multimedia Systems*, 22(6):751–766, 2016. 1

[13] Patricia L. M. Lang, Franziska M. Willems, J. F. Scheepens, Hernán A. Burbano, and Oliver Bossdorf. Using herbaria to study global environmental change. *New Phytologist*, 221(1):110–122, 2019. 1

[14] Damon P. Little, Melissa Tulig, Kiat Chuan Tan, Yulong Liu, Serge Belongie, Christine Kaeser-Chen, Fabián A Michelangeli, Kiran Panesar, R.V. Guha, and Barbara A Ambrose. An algorithm competition for automatic species identification from herbarium specimens. *Applications in plant sciences*, 8(6):e11365–e11365, 07 2020. 1, 4

[15] Titouan Lorieul, Katelin D. Pearson, Elizabeth R. Ellwood, Hervé Goëau, Jean-Francois Molino, Patrick W. Sweeney, Jennifer M. Yost, Joel Sachs, Erick Mata-Montero, Gil Nelson, Pamela S. Soltis, Pierre Bonnet, and Alexis Joly. Toward a large-scale and deep phenological stage annotation of herbarium specimens: Case studies from temperate, tropical, and equatorial floras. *Applications in Plant Sciences*, 7(3):e01233, 2019. 1

[16] Katelin D Pearson, Gil Nelson, Myla F J Aronson, Pierre Bonnet, Laura Brenskelle, Charles C Davis, Ellen G Denny, Elizabeth R Ellwood, Herve Goeau, J Mason Heberling, Alexis Joly, Titouan Lorieul, Susan J Mazer, Emily K Meineke, Brian J Stucky, Patrick Sweeney, Alexander E White, and Pamela S Soltis. Machine Learning Using Digitized Herbarium Specimens to Advance Phenological Research. *BioScience*, 70(7):610–620, 05 2020. 1

[17] Kathleen M. Pryer, Carlo Tomasi, Xiaohan Wang, Emily K. Meineke, and Michael D. Windham. Using computer vision on herbarium specimen images to discriminate among closely related horsetails (*Equisetum*). *Applications in Plant Sciences*, 8(6):e11372, 2020. 1

[18] Anastasia Stefanaki, Henk Porck, Ilaria Maria Grimaldi, Nikolaus Thurn, Valentina Pugliano, Adriaan Kardinaal, Jochem Salemink, Gerard Thijsse, Claudine Chavannes-Mazel, Erik Kwakkel, and Tinde van Andel. Breaking the silence of the 500-year-old smiling garden of everlasting flowers: The en tibi book herbarium. *PloS one*, 14(6):e0217779–e0217779, 06 2019. 1

[19] Kiat Chuan Tan, Yulong Liu, Barbara Ambrose, Melissa Tulig, and Serge Belongie. The herbarium challenge 2019 dataset. *CVPRW*, 2019. 6th Fine-Grained Visual Categorization Workshop (FGVC6). 1, 4

[20] Jordan R. Ubbens and Ian Stavness. Deep plant phenomics: A deep learning platform for complex plant phenotyping tasks. *Frontiers in Plant Science*, 8:1190, 2017. 1

[21] Jakob Unger, Dorit Merhof, and Susanne Renner. Computer vision applied to herbarium specimens of german trees: testing the future utility of the millions of herbarium specimen images for automated identification. *BMC Evolutionary Biology*, 16(1):248, 2016. 1

[22] Jana Wäldchen and Patrick Mäder. Plant species identification using computer vision techniques: A systematic literature review. *Archives of Computational Methods in Engineering*, 25(2):507–543, 2018. 1, 2

[23] D Wijesingha and Faiz Marikar. Automatic detection system for the identification of plants using herbarium specimen images. *Tropical Agricultural Research*, 23, 09 2012. 1

[24] Peter Wilf, Shengping Zhang, Sharat Chikkerur, Stefan A. Little, Scott L. Wing, and Thomas Serre. Computer vision

cracks the leaf code. *Proceedings of the National Academy of Sciences*, 113(12):3305–3310, 2016. 1

[25] Sun Wu and Udi Manber. Fast text searching: allowing errors. *Communications of the ACM*, 35:83–91, 1992. 3

[26] Sohaib Younis, Claus Weiland, Robert Hoehndorf, Stefan Dressler, Thomas Hickler, Bernhard Seeger, and Marco Schmidt. Taxon and trait recognition from digitized herbarium specimens using deep convolutional neural networks. *Botany Letters*, 165(3-4):377–383, 2018. 1

[27] Xinyu Zhou, Cong Yao, He Wen, Yuzhi Wang, Shuchang Zhou, Weiran He, and Jiajun Liang. East: An efficient and accurate scene text detector. 2017. 3