Applications in Plant Sciences

An algorithm competition for automatic species identification from herbarium specimens

Damon P. Little[1] ID , Melissa Tulig[1] ID , Kiat Chuan Tan[2], Yulong Liu[2,5], Serge Belongie[2,3] ID , Christine Kaeser-Chen[2], Fabián A. Michelangeli[1] ID , Kiran Panesar[4], R. V. Guha[4], and Barbara A. Ambrose[1,6] ID

PREMISE: Plant biodiversity is threatened, yet many species remain undescribed. It is estimated that >50% of undescribed species have already been collected and are awaiting discovery in herbaria. Robust automatic species identification algorithms using machine learning could accelerate species discovery.

METHODS: To encourage the development of an automatic species identification algorithm, we submitted our Herbarium 2019 data set to the Fine-Grained Visual Categorization sub-competition (FGVC6) hosted on the Kaggle platform. We chose to focus on the flowering plant family Melastomataceae because we have a large collection of imaged herbarium specimens (46,469 specimens representing 683 species) and taxonomic expertise in the family. As is common for herbarium collections, some species in this data set are represented by few specimens and others by many.

RESULTS: In less than three months, the FGVC6 Herbarium 2019 Challenge drew 22 teams who entered 254 models for Melastomataceae species identification. The four best algorithms identified species with >88% accuracy.

DISCUSSION: The FGVC competitions provide a unique opportunity for computer vision and machine learning experts to address difficult species-recognition problems. The Herbarium 2019 Challenge brought together a novel combination of collections resources, taxonomic expertise, and collaboration between botanists and computer scientists.

KEY WORDS   artificial intelligence; computer vision; FGVC; herbarium specimen; Kaggle; machine learning; Melastomataceae.

The major challenge to understanding and cataloging plant diversity is devising novel approaches to speed up the process of discovery. It has taken more than 250 years to name the 400,000 known plant species using a laborious manual process that relies on a shrinking group of experts examining individual specimens in detail. An estimated one million herbarium specimens remain unidentified, and as many as 70,000 flowering plant species alone are yet to be described (Joppa et al., 2011). It is very likely that new species are waiting to be discovered in the unidentified specimen backlog. In fact, an estimated 84% of undescribed plant diversity may be present in herbarium collections; therefore, "herbaria are a major frontier for species discovery" (Bebber et al., 2010).

Herbaria and the massive repository of data they contain provide snapshots of plant diversity through time. The integrity of the plant is maintained in herbaria as a pressed, dried specimen; a specimen collected hundreds of years ago looks much the same as one collected a month ago. Although some specimens may not initially be fully identified, all contain morphological features, and usually

include collection dates and locations, as well as the name of the person who collected the specimen. This information, multiplied by millions of plant collections, provides the framework for understanding plant diversity on a massive scale and learning how it has changed over time.

The use of artificial intelligence (AI) to facilitate the naming of organisms is a new and rapidly developing field (reviewed by Wäldchen et al., 2018). More than 15 years ago, Gaston and O'Neill (2004) proposed the idea of using automated techniques for plant identification. To date, many automated plant species identification studies have focused on leaves alone (Wijesingha and Marikar, 2011; Wilf et al., 2016; Kho et al., 2018). Convolutional neural networks (CNNs) for automated identification were first reported in 2012 and resulted in a dramatic increase in the accuracy of this field (Krizhevsky et al., 2012). CNNs represented a significant improvement on other automated identification algorithms (e.g., Palaniswamy et al., 2010; Unger et al., 2016) because the images require little preprocessing and, therefore, less human input or intervention. Many studies using automated plant species identification have focused on very small ranges of plant taxa, ranging from a few species to two families (Clark et al., 2012; Nasir et al., 2014; Unger et al., 2016; Schuettpelz et al., 2017; Kho et al., 2018; Pryer et al., 2020), whereas other studies have looked at a large number of taxa and shown the power of combining CNNs with herbarium specimens for relatively accurate automatic species identification (Carranza-Rojas et al., 2017; Younis et al., 2018).

Species identification algorithms have been developed for a large number of organisms using an online social network of naturalists, citizen scientists, and biologists, built on the concept of mapping and sharing observations of biodiversity around the globe. One example of an app that uses an online network of users, computer vision, and machine learning is iNaturalist (Van Horn et al., 2017; Van Horn et al., 2018a), an app that helps users identify animal and plant species from pictures they take of an organism. iNaturalist competitions run on the online platform Kaggle (https://www.kaggle.com, described below) demonstrated the feasibility and potential of using deep learning for species recognition, and have resulted in several influential publications (Cui et al., 2018; Sulc and Matas, 2018; Van Horn et al., 2018b), which in turn has helped iNaturalist build better AI models (Van Horn et a., 2017; Van Horn et al., 2018a).

Kaggle is an online platform that hosts competitions to develop algorithms and other analytical tools. Kaggle has more than one million users worldwide, ranging from amateurs to more experienced and professional data scientists. These competitions are based on open data sets ranging from images to text to structured data. Each Kaggle competition crowdsources predictive modeling solutions

for a particular data set, resulting in tens to hundreds of models from just as many competitors. Kaggle provides a platform for anyone wanting to learn new data science skills, and has ongoing competitions for all skill levels (https://www.kaggle.com/docs/competitions). The Kaggle platform provides a real-time leaderboard of competitor performance, a forum for questions, automated model scoring, and a competitor score report. In addition, Kaggle competitions take approximately 3–4 months from data release to the identification of the best-scoring algorithm. Some competitions, although not all, offer cash prizes to the winning algorithm.

To develop an automatic species identification algorithm for herbarium specimens, we used the Kaggle competition platform to crowdsource an algorithm. We chose a well-curated data set for which we have a large number of digitized specimens and a taxonomic expert in the plant family (Fig. 1). In just two months, 254 models were developed to automatically identify the herbarium specimens in our data set. The top four models were all able to identify our specimens to species with >88% accuracy, and these winning models were presented at the sixth Fine-Grained Visual Classification (FGVC6) workshop at the 2019 Computer Vision and Pattern Recognition (CVPR
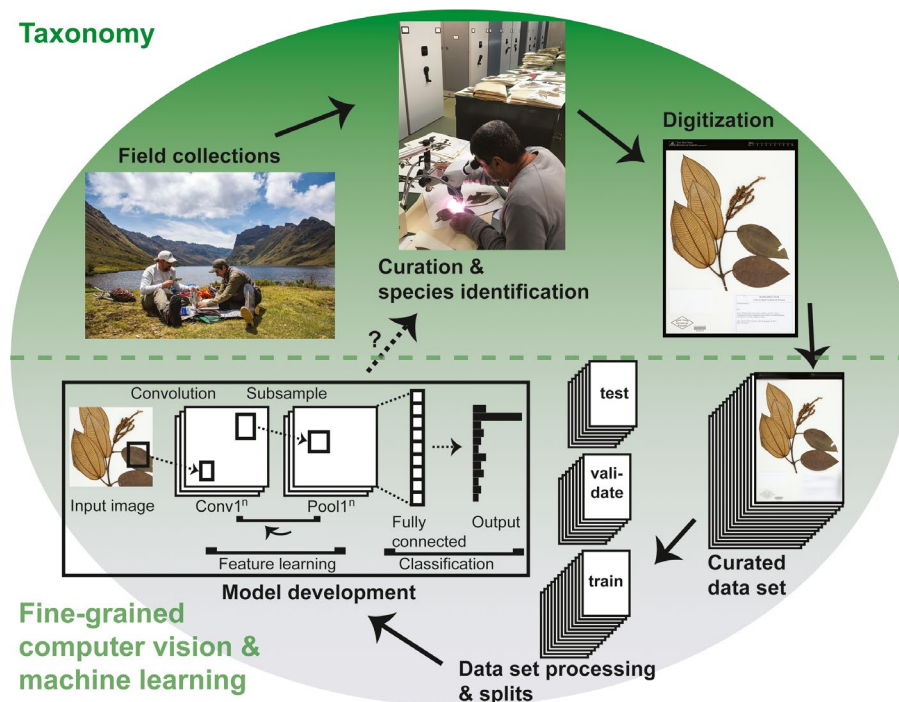

FIGURE 1. Workflow diagram of the Herbarium 2019 Challenge. The top half of the oval shows a typical taxonomy workflow while the bottom half illustrates the fine-grained visual classification and machine learning workflow. For Herbarium 2019, the workflow spanned field collections all the way through to the output models. The dotted line connecting model development to species identification and curation remains an open research question. The Herbarium 2019 data set was composed of specimens from the family Melastomataceae, collected in the field, curated, and for which the species had been determined. The physical specimens were previously imaged and placed in the C.V. Starr Virtual Herbarium at New York Botanical Garden. For the Herbarium 2019 competition, the curated Melastomataceae data set contained species with at least 20 specimens each. The data set was resized, and the text and barcodes blurred. The data set was split into training, validation, and test data sets. Competitors developed their models with the test data set using convolutional neural network architectures, which included feature extraction and the modification of variable neuron weights to fit the specimen images. The output classification is represented as a histogram that ranks the probability of the species identification.
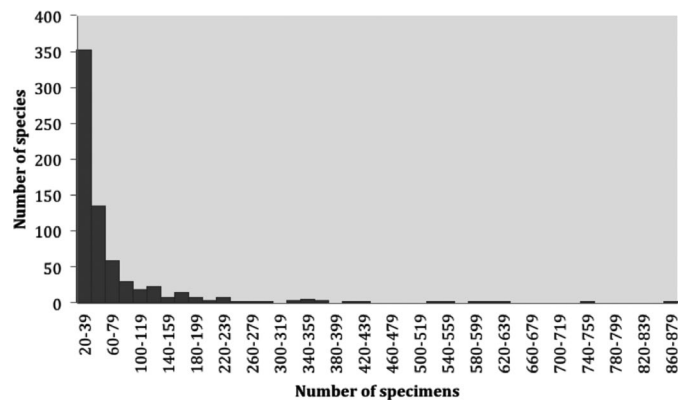
FIGURE 2. Histogram of the long-tailed distribution of the Melastomataceae showing the number of specimens for each species. Only species with a minimum of 20 specimens were included in the data set, and 352 species were represented by 20–39 specimens. One species, *Miconia prasina* (Sw.) DC., was represented by 865 specimens.

2019) meeting of the Institute of Electrical and Electronics Engineers (IEEE) computer society (Fig. 2).

METHODS

Data set

The data set consisted of 46,469 herbarium specimen images representing 683 species from the family Melastomataceae. Details of the data set are provided by Tan et al. (2019). Briefly, the data set was broken down into 75% training, 5% validation, and 20% test specimens that resulted in 34,225, 2679, and 9565 images, respectively (Fig. 1). The data set has a long-tailed distribution, with some species represented by a few specimens and other species represented by many specimens (Fig. 2). This distribution, however, accurately reflects the distribution of specimens per species in herbaria. The data set was limited to species represented by at least 20 specimen images at the lowest end as we thought that this minimum number of images would yield an accurate model while still providing a large data set. In contrast, some species were represented by more than 100 specimen images (Fig. 2). To ensure that specimens of each species were represented in the training, validation, and test data sets at similar proportions, the data set was broken down randomly by species. The images were harvested from Data Commons (http://www.datacommons.org/datasets) or from New York Botanical Garden's Virtual Herbarium (http://sweetgum.nybg.org/science/vh/).

All herbarium specimens contain text, usually in the bottom right-hand corner, that includes the species name and collection details. In addition, each herbarium sheet contains a barcode. To ensure that the species-identification algorithms are developed based on the species image rather than the text present on the specimen, the text and barcodes were blurred. PhotoOCR (Bissacco et al., 2013) was used to detect the text and a Gaussian blur was used to obscure it, as previously described (Tan et al., 2019). Finally, the digitized images were resized to 1024 pixels in the largest dimension, while maintaining aspect resulting in a 25-fold reduction in the size of the data set (Tan et al., 2019).

Herbarium 2019 Challenge hosted by Kaggle

This competition was termed Herbarium 2019, a sub-competition of FGVC6, and was listed on a Google website for the workshop (https://sites.google.com/view/fgvc6/competitions/herbarium-2019). The competition was hosted on the Kaggle platform (https://www.kaggle.com/c/herbarium-2019-fgvc6). The Kaggle page included the following tabs: Overview, Data, Notebooks, Discussion, Leaderboard, and Rules. The Overview tab included a description of the data set, evaluation details, a timeline of the competition, and information about FGVC6 and CVPR 2019. The Data tab included details of, and links to, the data set (hosted on the Google Cloud platform). Both low- and high-resolution images were available to competitors. The Discussion tab provided an online forum for competitors and hosts to discuss the Herbarium 2019 Challenge. Finally, the Leaderboard tab provided real-time scoring of the teams. Details of the competition were also available on GitHub at https://github.com/visipedia/herbarium_comp. The data set was released online on 5 April 2019 and the competition closed on 7 June 2019. No prize was offered for the winning model.

RESULTS

Competition overview

A summary of the Herbarium 2019 Challenge results was presented by Kiat Chuan Tan at the FGVC6 workshop held during the CVPR 2019 conference. Briefly, 22 teams submitted 254 algorithms to identify Melastomataceae species from herbarium specimens (Table 1). The top four teams developed algorithms that correctly identified the species with >88% accuracy. These teams were (ordered by descending accuracy): MEGVII Research Nanjing (Boyan Zhou, Quan Cui, Borui Zhao, and Yanping Xie from MEGVII Research Nanjing, Beijing, China) with 89.8% accuracy, PEAK (Chunqiao Xu, Shao Zeng, Qiule Sun, Shuyu Ge, and Peihua Li from Dalian University of Technology in Liaoning, China) with 89.1% accuracy, Miroslav Valan (from the Swedish Museum of Natural History in Stockholm, Sweden) with 89.0% accuracy, and Hugo Touvron and Andrea Vedaldi (from Facebook AI Research, London, United Kingdom) with 88.9% accuracy.

Details of available best models

The teams are not obliged to make models open access; however, two teams made their models publicly available and we were able to analyze them here.

TABLE 1. The top four Herbarium 2019 Challenge competitors with the accuracy of their models and the model architectures they used in the competition.

Name of team	Accuracy	Model architecture
MEGVII Research Nanjing (Beijing, China)	89.8%	SeResNeXt-101, ResNet-50
PEAK (Dalian University of Technology, Liaoning, China)	89.1%	Not made publicly available
Miroslav Valan (Swedish Museum of Natural History, Stockholm, Sweden)	89.0%	SENet-154, ResNet-50
Hugo Touvron and Andrea Vedaldi (Facebook AI Research, London, UK)	88.9%	SENet-154, ResNet-50

Two of the top-performing teams, MEGVII and Touvron–Vedaldi, both implemented their solutions in PyTorch and utilized a strategy of training on lower-resolution images and testing on higher-resolution images. The latter feature is credited with the high accuracy obtained (Touvron et al., 2019).

The approaches also differed in several aspects. The final output of MEGVII is a weighted average of five SeResNeXt (Xie et al., 2016; Hu et al., 2017) and ResNet (He et al., 2015) variants pretrained on ImageNet data that was supplemented with iNaturalist 2018 data in some cases (Russakovsky et al., 2015; Van Horn et al, 2018a). In the network training and testing, class weights were applied in proportion to the frequency of the class occurrence in the training data; images were resized; colors were left unaltered; and either cross-entropy or focal loss functions were used. The output from the neural networks was weighted to 0, 1, 2, or 3 in order to maximize ensemble accuracy, with the weights being determined using a brute force algorithm set to optimize accuracy.

In contrast, the final output of Touvron–Vedaldi is an equal-weighted average of two models: a SENet (Hu et al., 2017) variant and a ResNet (He et al., 2015) variant pretrained on ImageNet data. In the network training and testing, class weights were not used; images were resized; color was standardized; and a cross-entropy loss function was used. The output from the neural networks was averaged. Additional details of the Touvron–Vedaldi solution are available from Touvron et al. (2019).

DISCUSSION

To develop automatic plant species identification algorithms, we need to bring together data scientists and botanists. As botanists, we need tools to tackle large-scale questions or those that not only require computing power but the development of new algorithms or software. On the other hand, data scientists, particularly those working in AI, need large, well-curated, open access data sets to develop their models.

The Kaggle platform hosts competitions for data science to solve predictive modeling problems by crowdsourcing algorithms. The Kaggle platform is very accessible and attracts competitors from a range of organizations around the world. The top four best-performing teams for the Herbarium 2019 Challenge included companies (MEGVII and Facebook AI Research), a public university (Dalian University of Technology), and an individual (Miroslav Valan). Interestingly, Miroslav Valan is a veterinarian who joined the competition during his vacation and designed the models on his phone. The neural network training and testing was then done using computer power provided by Kaggle.

Three previous studies used similar approaches to the Herbarium 2019 Challenge by using CNNs to identify plant species from herbarium specimens (Carranza-Rojas et al., 2017; Schuettpelz et al., 2017; Younis et al., 2018). The Schuettpelz et al. (2017) study resembles the Herbarium 2019 Challenge in that it used the entire herbarium specimen; however, their study focused on identifying only plants within one of two plant families, Lycopodiaceae or Selaginellaceae, with another part of the study focused on detecting mercury contamination on herbarium sheets. The studies by Carranza-Rojas et al. (2017) and Younis et al. (2018) relied on larger numbers of specimens representing a relatively broad span of taxonomic diversity, like the Herbarium 2019 Challenge. Younis et al. (2018) selected the 1000 species represented by the largest numbers of specimens

available on the Global Biodiversity Information Facility (GBIF), with some of the species represented by well over 1000 images and the least-represented species by over 500. Although they achieved relatively accurate identification results (82.4%), their data set contained very distantly related (and phenetically different) species, and the image-to-species distribution curve does not reflect the situation of most herbaria. Carranza-Rojas et al. (2017) also included a large set of images, but their data set also included pictures of live plants and individual leaf scans. Moreover, their data set was really an aggregation of five different data sets from a broad geographic and taxonomic scope. While these three studies (Carranza-Rojas et al., 2017; Schuettpelz et al., 2017; Younis et al., 2018) proved the utility of CNNs for specimen identification, we feel that our data set is a more accurate representation of herbarium data sets and the potential of CNNs in species identification.

We concentrated on a single species-rich family, in which some species are highly morphologically variable while others are not, but also our data set more accurately reflects data sets present in herbaria, with some species represented by many specimens and others represented by few.

We chose to focus on one plant family for the Herbarium 2019 Challenge as we have a taxonomist on staff who specializes in this plant family; therefore, we knew we had a large and well-curated data set for the development of automatic species identification algorithms. This turned out to be very important as it allowed for the development of improved training strategies for CNNs (Touvron et al., 2019). However, for the development of future herbarium models, we need to determine the minimum number of images per species needed for the training data set without a significant decrease in accuracy, which would hopefully allow an increase in the number of species represented in our data sets. Finally, flagging specimens that have a low probability of being assigned to any species in the training data set would also be crucial for the discovery of new species "hidden" in the herbarium (Bebber et al., 2010), as contemporary neural network designs are incapable of unambiguously indicating that an image does not match any of the samples it was trained with. Therefore, one of the more interesting avenues for future research is the design of new neural network layers or training techniques that will allow for the identification of unknown species.

AI and machine learning provide a great opportunity for rapidly identifying herbarium specimens and holds promise for specimen curation. Equally beneficially, a large well-curated data set provides the opportunity to make advancements in machine learning models. Finally, the Herbarium 2019 Challenge and resulting models have established a platform to expand to a larger specimen data set beyond the Melastomataceae and to push the lower limits of species representation in the data set. This would be a grand challenge, but one that may help us to more quickly curate and one day discover the new species already housed in herbaria around the world.

ACKNOWLEDGMENTS

DATA AVAILABILITY

An overview of the Herbarium 2019 Challenge along with the data set can be found at https://www.kaggle.com/c/herbarium-2019-fgvc6/overview. Details of the model from Touvron–Vedaldi can be found at https://github.com/facebookresearch/FixRes.

LITERATURE CITED

Bebber, D., M. Carine, J. Wood, A. Wortley, D. Harris, G. Prance, G. Davidse, et al. 2010. Herbaria are a major frontier for species discovery. *Proceedings of the National Academy of Sciences, USA* 107: 22169–22171.

Bissacco, A., M. Cummins, Y. Netzer, and H. Neven. 2013. PhotoOCR: Reading text in uncontrolled conditions, 785–792. *In* Proceedings of the IEEE International Conference on Computer Vision, Sydney, Australia.

Carranza-Rojas, J., H. Goeau, P. Bonnet, E. Mata-Montero, and A. Joly. 2017. Going deeper in the automated identification of Herbarium specimens. *BMC Evolutionary Biology* 17: 181.

Clark, J., D. Corney, and H. Tang. 2012. Automated plant identification using artificial neural networks. *In* Proceedings of the IEEE Symposium on Computational Intelligence in Bioinformatics and Computational Biology (CIBCB), San Diego, California, USA.

Cui, Y., Y. Song, C. Sun, A. Howard, and S. Belongie. 2018. Large scale fine-grained categorization and domain-specific transfer learning. arXiv 1806.06193 [Preprint]. Published 16 June 2018 [accessed 6 May 2020]. Available at: https://arxiv.org/abs/1806.06193.

Gaston, K. J., and M. O'Neill. 2004. Automated species identification: Why not? *Philosophical Transactions of the Royal Society B Biological Sciences* 359(1444): 655–657.

He, K., X. Zhang, S. Ren, and J. Sun. 2015. Deep residual learning for image recognition. arXiv 1512.03385 [Preprint]. Published 10 December 2016 [accessed 6 May 2020]. Available at: https://arxiv.org/abs/1512.03385

Hu, J., L. Shen, S. Albanie, G. Sun, and E. Wu. 2017. Squeeze-and-excitation networks. arXiv 1709.01507 [Preprint]. Published 5 September 2017 [accessed 6 May 2020]. Available at: https://arxiv.org/abs/1709.01507.

Joppa, L. N., D. L. Roberts, and S. L. Pimm. 2011. How many species of flowering plants are there? *Proceedings of the Royal Society B Biological Sciences* 278: 554–559.

Kho, S., S. Manickam, S. Malek, M. Mosleh, and S. Dhillon. 2018. Automated plant identification using artificial neural networks and support vector machine. *Frontiers in Life Science* 10: 98–107.

Krizhevsky, A., I. Sutskever, and G. Hinton. 2012. ImageNet classification with deep convolutional neural networks. *In* Advances in Neural Information Processing Systems 25 (Neural Information Processing Systems 2012), Lake Tahoe, Nevada, USA.

Nasir, A. F., M. N. A. Rahman, N. Mat, and R. Mamat. 2014. Automatic identification of *Ficus deltoidea* Jack (Moraceae) varieties based on leaf. *Modern Applied Science* 85: 121–131.

Palaniswamy, S., N. Thacker, and C. Klingenberg. 2010. Automatic identification of landmarks in digital images. *IET Computer Vision* 5: 247–260.

Pryer, K. M., C. Tomasi, X. Wang, E. K. Meineke, and M. D. Windham. 2020. Using computer vision on herbarium specimen images to discriminate among closely related horsetails (*Equisetum*). *Applications in Plant Sciences* 8(6): e11372.

Russakovsky, O., J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, et al. 2015. ImageNet large scale visual recognition challenge. *International Journal of Computer Vision* 115: 211–252.

Schuettpelz, E., P. Frandsen, R. Dikow, A. Brown, S. Orli, M. Peters, A. Metallo, et al. 2017. Applications of deep convolutional neural networks to digitized natural history collections. *Biodiversity Data Journal* 5: e21139.

Sulc, M., and J. Matas. 2018. Improving CNN classifiers by estimating test-time priors. arXiv 1805.08235 [Preprint]. Submitted 21 May 2018 [accessed 6 May 2020]. Available at: https://arxiv.org/abs/1805.08235.

Tan, K. C., Y. Liu, B. Ambrose, M. Tulig, and S. Belongie. 2019. The Herbarium Challenge 2019 Dataset. arXiv 1906.05372v2 [Preprint]. Published 12 June 2019 [accessed 6 May 2020]. Available at: https://arxiv.org/abs/1906.05372v2.

Touvron, H., A. Vedaldi, M. Douze, and H. Jégou. 2019. Fixing the train-test resolution discrepancy. arXiv 1906.06423v2 [Preprint]. Published 14 June 2019 [accessed 6 May 2020]. Available at: https://arxiv.org/abs/1906.06423v2.

Unger, J., D. Merhof, and S. Renner. 2016. 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: 248.

Van Horn, G., O. Mac Aodha, Y. Song, Y. Cui, C. Sun, A. Shepherd, H. Adam, et al. 2017. The iNaturalist challenge 2017 data set. arXiv 1707.06642 [Preprint]. Published 20 July 2017 [accessed 6 May 2020]. Available at: https://arxiv.org/abs/1707.06642.

Van Horn, G., O. Mac Aodha, Y. Song, Y. Cui, C. Sun, A. Shepherd, H. Adam, et al. 2018a. The iNaturalist species classification and detection dataset. *In* Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 8769–8778. IEEE, Piscataway, New Jersey, USA.

Van Horn, G., S. Branson, S. Loarie, S. Belongie, and P. Perona. 2018b. Lean multiclass crowdsourcing. *In* Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2714–2723. IEEE, Piscataway, New Jersey, USA.

Wäldchen, J., M. Rzanny, M. Seeland, and P. Mäder. 2018. Automated plant species identification—Trends and future directions. *PLoS Computational Biology* 14(4): e1005993.

Wijesingha, D., and F. Marikar. 2011. Automatic detection system for the identification of plants using herbarium specimen images. *Tropical Agricultural Research* 23: 42–50.

Wilf, P., S. Zhang, S. Chikkerur, S. Little, S. Wing, and T. Serre. 2016. Computer Vision cracks the leaf code. *Proceedings of the National Academy of Sciences, USA* 113: 3305–3310.

Younis, S., C. Weiland, R. Hoehndorf, S. Dressler, T. Hickler, B. Seeger, and M. Schmidt. 2018. Taxon and trait recognition from digitized herbarium specimens using deep convolutional neural networks. *Botany Letters* 165: 377–383.

Xie, S., R. Girshick, P. Dollár, Z. Tu, and K. He. 2016. Aggregated residual transformations for deep neural networks. arXiv 1611.05431 [Preprint]. Published 16 November 2016 [accessed 6 May 2020]. Available at: https://arxiv.org/abs/1611.05431.