LifeCLEF 2021 teaser: Biodiversity Identification and Prediction Challenges

Alexis Joly[1 [0000−0002−2161−9940]], Hervé Goëau[2 [0000−0003−3296−3795]], Stefan Kahl[7], Hervé Glotin[4 [0000−0001−7338−8518]], Elijah Cole[10 [0000−0001−6623−0966]], Benjamin Deneu[1 [0000−0003−0640−5706]], Maximillien Servajean[8 [0000−0002−9426−2583]], Titouan Lorieul[1 [0000−0001−5228−9238]], Willem-Pier Vellinga[5], Pierre Bonnet[2 [0000−0002−2828−4389]], Ivan Eggel[6], Henning Müller[6 [0000−0001−6800−9878]]

[1] Inria, LIRMM, Montpellier, France
[2] CIRAD, UMR AMAP, France
[3] INRA, UMR AMAP, France
[4] Aix Marseille Univ, Université de Toulon, CNRS, LIS, DYNI, Marseille, France
[5] Xeno-canto Foundation, The Netherlands
[6] HES-SO, Sierre, Switzerland
[7] Chemnitz University of Technology, Germany
[8] LIRMM, Université Paul Valéry, University of Montpellier, CNRS, France
[9] University of Geneva, Switzerland
[10] Caltech, US
[11] Dept. of Cybernetics, FAV, University of West Bohemia, Czechia

Abstract. Building accurate knowledge of the identity, the geographic distribution and the evolution of species is essential for the sustainable development of humanity, as well as for biodiversity conservation. However, the difficulty of identifying plants and animals in the field is hindering the aggregation of new data and knowledge. Identifying and naming living plants or animals is almost impossible for the general public and is often difficult even for professionals and naturalists. Bridging this gap is a key step towards enabling effective biodiversity monitoring systems. The LifeCLEF campaign, presented in this paper, has been promoting and evaluating advances in this domain since 2011. The 2021 edition proposes four data-oriented challenges related to the identification and prediction of biodiversity: (i) PlantCLEF: cross-domain plant identification based on herbarium sheets, (ii) BirdCLEF: bird species recognition in audio soundscapes, and (iii) GeoLifeCLEF: location-based prediction of species based on environmental and occurrence data.

Keywords: biodiversity · machine learning · IA · species identification · species prediction · plant identification · bird identification · species distribution model.

1 Introduction

Accurately identifying organisms observed in the wild is an essential step in ecological studies. Unfortunately, observing and identifying living organisms re-

quires high levels of expertise. For instance, plants alone account for more than 400,000 different species and the distinctions between them can be quite subtle. Since the Rio Conference of 1992, this *taxonomic gap* has been recognized as one of the major obstacles to the global implementation of the Convention on Biological Diversity [4]. In 2004, Gaston and O'Neill [12] discussed the potential of automated approaches for species identification. They suggested that, if the scientific community were able to (i) produce large training datasets, (ii) precisely evaluate error rates, (iii) scale up automated approaches, and (iv) detect novel species, then it would be possible to develop a generic automated species identification system that would open up new vistas for research in biology and related fields.

Since the publication of [12], automated species identification has been studied in many contexts [10, 25, 24, 20, 14, 22, 21, 29]. This area continues to expand rapidly, particularly due to recent advances in deep learning [13, 15, 23, 9, 28, 27, 26]. In order to measure progress in a sustainable and repeatable way, the LifeCLEF [6] research platform was created in 2014 as a continuation and extension of the plant identification task [19] that had been run within the ImageCLEF lab [5] since 2011 [17, 18, 16]. LifeCLEF expanded the challenge by considering animals in addition to plants, and including audio and video content in addition to images. LifeCLEF 2021 consists of three challenges (PlantCLEF, BirdCLEF, GeoLifeCLEF), which we will now describe in turn.

2 PlantCLEF 2021 Challenge: Identifying plant pictures from herbarium sheets

Motivation: For several centuries, botanists have collected, catalogued and systematically stored plant specimens in herbaria. These physical specimens are used to study the variability of species, their phylogenetic relationship, their evolution, or phenological trends. One of the key step in the workflow of botanists and taxonomists is to find the herbarium sheets that correspond to a new specimen observed in the field. This task requires a high level of expertise and can be very tedious. Developing automated tools to facilitate this work is thus of crucial importance. More generally, this will help to convert these invaluable centuries-old materials into FAIR [8] data.

Data collection: The task will rely on a large collection of more than 320,000 herbarium sheets used during the last PlantCLEF edition. The specimens were mostly collected in the Guiana shield and the Northern Amazon rainforest, focusing on about 1,000 plant species of the French Guiana flora. A valuable asset of this collection is that several herbarium sheets are accompanied by a few pictures of the same specimen in the field. New information such as morphological, ecological, phenological traits at the species level will be aggregated from various sources (EOL TraitBank, TRY Plant Trait Database, specimen annotations from "Herbier de Cayenne"), and will enrich the data collection this year.

Task description: The challenge will be evaluated as a cross-domain classification task. The training set will consist of herbarium sheets whereas the test

set will be composed of field pictures. To enable learning a mapping between the herbarium sheets domain and the field pictures domain, we will provide both herbarium sheets and field pictures for a subset of species. As was already anticipated in some promising methods evaluated in the last edition, morphological, ecological and phenological traits could potentially be directly integrated into the models and significantly improve the performances on this difficult task.

3 BirdCLEF 2021 Challenge: Bird species recognition in audio soundscapes

Motivation: Recognizing bird sounds in complex soundscapes is an important sampling tool that often helps to reduce the limitations of point counts. In the future, archives of recorded soundscapes will become increasingly valuable as the habitats in which they were recorded will be lost in the near future. This is already the case for soundscapes used for this competition and point counts to assess biodiversity from this particular location in South America will only be possible through soundscape analysis. It is imperative to develop new technologies that can cope with the increasing amount of audio data and that can help to accelerate the process of species diversity assessments. In the past few years, deep learning approaches have transformed the field of automated soundscape analysis. Yet, the results still lack reliability and submitted systems often yield very low scores particularly when the vocal density of species is high. The goal of this competition is to establish training and test datasets that can serve as real-world applicable evaluation scenarios and help the scientific community to advance their conservation efforts through automated bird sound recognition.

Data collection: The 2021 dataset will closely resemble the previously used training and test data. However, we will establish a new subset of data to allow participants that are new to the evaluation campaign to quickly train and test their systems on an entry-level dataset. Training data will again be provided by the Xeno-canto community and will feature almost 1,000 bird species from three continents. The test data will contain expert annotated soundscapes with tens of thousands of labels and high overlap of bird vocalizations. The entry-level portion of the data will contain 20-50 species for training and soundscapes from a selected location in Germany with a runtime of only one hour. This approach reflects the feedback that we received from participants of the 2020 edition and we hope to attract more participating groups and a better turnout in terms of submitted runs and scores.

Task description: The evaluation mode will closely resemble the 2020 test mode and we will use the same established metrics of class-wise and sample-wise mean average precision. However, we will alter the assessment of submitted results to better reflect false positives. The test data annotations have a coverage of 100% of all audio files and we will switch the evaluation mode to not only test for segments that have a label but also for segments that do not contain an annotation (e.g., nighttime recordings). Doing so will allow us to keep our current

(well established) evaluation system in place while better reflecting real-world use cases.

4 GeoLifeCLEF 2021 Challenge: Location-based prediction of species based on environmental and occurrence data

Motivation: Automatically predicting the list of species that are the most likely to be observed at a given location is useful for many scenarios in biodiversity informatics. First of all, it could improve species identification tools by reducing the list of candidate species that are observable at a given location (be they automated, semi-automated or based on classical field guides or flora). More generally, it could facilitate biodiversity inventories through the development of location-based recommendation services (typically on mobile phones), favor the involvement of non-expert nature observers, as well as accelerate the annotation or validation of species observed by non-experts to produce high quality datasets. Last but not least, it might serve educational purposes thanks to biodiversity discovery applications providing functionalities such as contextualized educational pathways.

Data collection: The dataset used in 2020 [11] contained about 2 million plant and animal occurrences, each paired with high-resolution covariates (satellite, land cover, altitude) and environmental rasters (bioclimatic variables, soil type, etc.). This dataset of about 840GB took months to build and was delivered quite late to the participants. Training a model on it takes almost two weeks on a machine equipped with several modern GPUs. Last year, only two participants out of the 40 registered managed to submit runs. Therefore, we think it is necessary to keep the same dataset in 2021. However, to facilitate participation and foster consistent progress over last year, we will provide (i) new python tools and intermediate data formats facilitating the training of models, (ii) a validation set allowing participants to compare the performance they obtain with the one of the best method of last year, (iii) an entry-level subset of the whole dataset facilitating debugging before training large-scale models.

Task description: Given the test set of locations (i.e. geo-coordinates) and corresponding high-resolution and environmental covariates, the goal of the task will be to return for each location a ranked list of species sorted according to the likelihood that they might have been observed at that location. The metric used will be the Average-30 accuracy [11].

5 Timeline and registration instructions

All information about the timeline and participation in the challenges is provided on the LifeCLEF 2021 web pages [7]. The system used to run the challenges (registration, submission, leaderboard, etc.) is the AIcrowd platform [1].

6 Discussion and Conclusion

The long-term societal impact of boosting research on biodiversity informatics is difficult to overstate. To fully reach its objective, an evaluation campaign such as LifeCLEF requires a long-term research effort so as to (i) encourage non-incremental contributions, (ii) measure consistent performance gaps, (iii) progressively scale-up the problem and (iv), enable the emergence of a strong community. The 2021 edition of the lab will support this vision and will include the following innovations:

- The PlantCLEF task will be extended with traits information, i.e. structured tags or numerical values related to the morphological, ecological or phenological attributes of species.
- An entry-level dataset (in addition to the official data) will be delivered for the BirdCLEF task in order to allow new participants to quickly get results and progress more iteratively.
- New helper tools and more pre-formatted data will be provided for the GeoLifeCLEF task in order to facilitate participation and build upon the best methods of previous year.

The results of this challenge will be published in the proceedings of the CLEF 2021 conference [3] and in the CEUR-WS workshop proceedings [2].

Acknowledgements: This work is supported in part by the SEAMED PACA project, the SMILES project (ANR-18-CE40-0014), and an NSF Graduate Research Fellowship (DGE-1745301). This work has received funding from the European Union's Horizon 2020 research and innovation program under grant agreement No 863463 (Cos4Cloud project).

References

1. AIcrowd. https://www.aicrowd.com/
2. CEUR-WS. http://ceur-ws.org/
3. CLEF 2021. https://clef2021.clef-initiative.eu/
4. Convention on Biodiversity. https://www.cbd.int/
5. ImageCLEF. http://www.imageclef.org/
6. LifeCLEF. http://www.lifeclef.org/
7. LifeCLEF 2021. https://www.imageclef.org/LifeCLEF2021
8. The FAIR Data Principles. https://www.force11.org/group/fairgroup/fairprinciples
9. Bonnet, P., Goëau, H., Hang, S.T., Lasseck, M., Šulc, M., Malécot, V., Jauzein, P., Melet, J.C., You, C., Joly, A.: Plant identification: experts vs. machines in the era of deep learning. In: Multimedia Tools and Applications for Environmental & Biodiversity Informatics, pp. 131–149. Springer (2018)
10. Cai, J., Ee, D., Pham, B., Roe, P., Zhang, J.: Sensor network for the monitoring of ecosystem: Bird species recognition. In: Intelligent Sensors, Sensor Networks and Information, 2007. ISSNIP 2007. 3rd International Conference on (2007). https://doi.org/10.1109/ISSNIP.2007.4496859

11. Cole, E., Deneu, B., Lorieul, T., Servajean, M., Botella, C., Morris, D., Jojic, N., Bonnet, P., Joly, A.: The geolifeclef 2020 dataset. arXiv preprint arXiv:2004.04192 (2020)
12. Gaston, K.J., O'Neill, M.A.: Automated species identification: why not? Philosophical Transactions of the Royal Society of London B: Biological Sciences **359**(1444), 655–667 (2004)
13. Ghazi, M.M., Yanikoglu, B., Aptoula, E.: Plant identification using deep neural networks via optimization of transfer learning parameters. Neurocomputing **235**, 228–235 (2017)
14. Glotin, H., LeCun, Y., Artiéres, T., Mallat, S., Tchernichovski, O., Halkias, X.: Proc. Neural Information Processing Scaled for Bioacoustics, from Neurons to Big Data. NIPS Int. Conf., Tahoe USA (2013), http://sabiod.org/nips4b
15. Goeau, H., Bonnet, P., Joly, A.: Plant identification based on noisy web data: the amazing performance of deep learning (LifeCLEF 2017). In: CLEF 2017-Conference and Labs of the Evaluation Forum. pp. 1–13 (2017)
16. Goëau, H., Bonnet, P., Joly, A., Bakic, V., Barthélémy, D., Boujemaa, N., Molino, J.F.: The ImageCLEF 2013 plant identification task. In: CLEF. Valencia, Spain (2013)
17. Goëau, H., Bonnet, P., Joly, A., Boujemaa, N., Barthélémy, D., Molino, J.F., Birnbaum, P., Mouysset, E., Picard, M.: The ImageCLEF 2011 plant images classification task. In: CLEF 2011 (2011)
18. Goëau, H., Bonnet, P., Joly, A., Yahiaoui, I., Barthélémy, D., Boujemaa, N., Molino, J.F.: ImageCLEF2012 plant images identification task. In: CLEF 2012. Rome (2012)
19. Goëau, H., Joly, A., Bonnet, P., Bakic, V., Barthélémy, D., Boujemaa, N., Molino, J.F.: The imageclef plant identification task 2013. In: Proceedings of the 2nd ACM international workshop on Multimedia analysis for ecological data. pp. 23–28. ACM (2013)
20. ICML int. Conf.: Proc. 1st workshop on Machine Learning for Bioacoustics - ICML4B (2013), http://sabiod.univ-tln.fr
21. Joly, A., Goëau, H., Bonnet, P., Bakić, V., Barbe, J., Selmi, S., Yahiaoui, I., Carré, J., Mouysset, E., Molino, J.F., et al.: Interactive plant identification based on social image data. Ecological Informatics **23**, 22–34 (2014)
22. Lee, D.J., Schoenberger, R.B., Shiozawa, D., Xu, X., Zhan, P.: Contour matching for a fish recognition and migration-monitoring system. In: Optics East. pp. 37–48. International Society for Optics and Photonics (2004)
23. Lee, S.H., Chan, C.S., Remagnino, P.: Multi-organ plant classification based on convolutional and recurrent neural networks. IEEE Transactions on Image Processing **27**(9), 4287–4301 (2018)
24. Towsey, M., Planitz, B., Nantes, A., Wimmer, J., Roe, P.: A toolbox for animal call recognition. Bioacoustics **21**(2), 107–125 (2012)
25. Trifa, V.M., Kirschel, A.N., Taylor, C.E., Vallejo, E.E.: Automated species recognition of antbirds in a Mexican rainforest using hidden Markov models. The Journal of the Acoustical Society of America **123**, 2424 (2008)
26. Van Horn, G., Mac Aodha, O., Song, Y., Cui, Y., Sun, C., Shepard, A., Adam, H., Perona, P., Belongie, S.: The inaturalist species classification and detection dataset. CVPR (2018)
27. Wäldchen, J., Mäder, P.: Machine learning for image based species identification. Methods in Ecology and Evolution **9**(11), 2216–2225 (2018)

28. Wäldchen, J., Rzanny, M., Seeland, M., Mäder, P.: Automated plant species identification—trends and future directions. PLoS computational biology **14**(4), e1005993 (2018)
29. Yu, X., Wang, J., Kays, R., Jansen, P.A., Wang, T., Huang, T.: Automated identification of animal species in camera trap images. EURASIP Journal on Image and Video Processing (2013)