LifeCLEF 2020 Teaser: Biodiversity Identification and Prediction Challenges

Alexis Joly, Hervé Goëau, Christophe Botella, Rafael Ruiz de Castaneda, Hervé Glotin, Elijah Cole, Julien Champ, Benjamin Deneu, Maximilien Servajean, Titouan Lorieul, et al.

LifeCLEF 2020 Teaser: Biodiversity Identification and Prediction Challenges

Alexis Joly[1(✉)], Hervé Goëau[2], Stefan Kahl[7], Christophe Botella[1,3], Rafael Ruiz De Castaneda[9], Hervé Glotin[4], Elijah Cole[10], Julien Champ[1], Benjamin Deneu[1], Maximillien Servajean[8], Titouan Lorieul[1], Willem-Pier Vellinga[5], Fabian-Robert Stöter[1], Andrew Durso[9], Pierre Bonnet[2], and Henning Müller[6]

[1] Inria, LIRMM, Montpellier, France
alexis.joly@inria.fr
[2] CIRAD, UMR AMAP, Montpellier, France
[3] INRA, UMR AMAP, Montpellier, France
[4] Aix Marseille Univ, Université de Toulon, CNRS, LIS, DYNI, Marseille, France
[5] Xeno-canto Foundation, Amsterdam, The Netherlands
[6] HES-SO, Sierre, Switzerland
[7] Chemnitz University of Technology, Chemnitz, Germany
[8] LIRMM, Université Paul Valéry, University of Montpellier, CNRS, Montpellier, France
[9] University of Geneva, Geneva, Switzerland
[10] Caltech, Pasadena, USA

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 2020 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, (iii) GeoLifeCLEF: location-based prediction of species based on environmental and occurrence data, and (iv) SnakeCLEF: image-based snake identification.

Keywords: Biodiversity · Machine learning · IA · Species identification · Species prediction · Plant identification · Bird identification · Snake identification · Species distribution model

J. M. Jose et al. (Eds.): ECIR 2020, LNCS 12036, pp. 542–549, 2020.
https://doi.org/10.1007/978-3-030-45442-5_70

1 Introduction

Accurately identifying organisms observed in the wild is an essential step in ecological studies. Unfortunately, observing and identifying living organisms requires 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 [6]. In 2004, Gaston and O'Neill [27] 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 [27], automated species identification has been studied in many contexts [26,29,30,38,41,43,44,48]. This area continues to expand rapidly, particularly due to recent advances in deep learning [25,28,31,42,45–47]. In order to measure progress in a sustainable and repeatable way, the LifeCLEF [15] research platform was created in 2014 as a continuation and extension of the plant identification task [37] that had been run within the ImageCLEF lab [12] since 2011 [32–34]. LifeCLEF expanded the challenge by considering animals in addition to plants, and including audio and video content in addition to images. LifeCLEF 2020 consists of four challenges (PlantCLEF, BirdCLEF, GeoLifeCLEF, and SnakeCLEF), which we will now describe in turn.

2 PlantCLEF 2020 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 [23] data.

Data Collection: The task will rely on a large collection of more than 60,000 herbarium sheets that were collected in French Guyana (The "Herbier IRD de Guyane", CAY [14]) and digitized in the context of the e-ReColNat project [8]. iDigBio [11] hosts millions of images of herbarium specimens. Several tens of thousands of these images, illustrating the French Guyana flora, will be used for the PlantCLEF task this year. A valuable asset of this collection is that several

herbarium sheets are accompanied by a few pictures of the same specimen in the field. For the test set, we will use in-the-field pictures coming different sources including Pl@ntNet [20] and Encyclopedia of Life [9].

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. The metrics used for the evaluation of the task will be the classification accuracy and the mean reciprocal rank.

3 BirdCLEF 2020 Challenge: Bird Species Recognition in Audio Soundscapes

Motivation: Monitoring birds by sound is important for many environmental and scientific purposes. Birds are difficult to photograph and sound offers better possibilities for inventory coverage. A number of participatory science projects have focused on recording a very large number of bird sounds, making it possible to recognize most species by their sound and to train deep learning models to automate this process. It was shown in previous editions of BirdCLEF [35,36] that systems for identifying birds from mono-directional recordings are now performing very well and several mobile applications implementing this are emerging today. However, there is also interest in identifying birds from *omnidirectional* or *binaural* recordings. This would enable more passive monitoring scenarios like networks of static recorders that continuously capture the surrounding sound environment. The advantage of this type of approach is that it introduces less sampling bias than the opportunistic observations of citizen scientists. However, recognizing birds in such content is much more difficult due to the high vocal activity with signal overlap (e.g. during the dawn chorus) and high levels of ambient noise.

Data Collection: The training set used for the challenge will be a version of the 2019 training set [40] enriched by new contributions from the Xeno-canto network and a geographic extension. It will contain approximately 80K recordings covering between 1500 and 2000 species from North, Central and South America, as well as Europe. This will be the largest bioacoustic dataset used in the literature. For the test set, three sources of soundscapes will be used: (i) 100+ h of manually annotated soundscapes recorded using 30 field recorders between January and June of 2017 in Ithaca, NY, USA by the Cornell Lab of Ornithology [7], (ii) 10+ h of fully annotated dawn chorus soundscapes recorded using solar-powered field recorders between January and June 2018 near Frankfurt (a.M.), Germany by OekoFor [19], (iii) 2 h acquired at high sampling rate (250 kHz) by binaural antenna in Côte d'Azur, France.

Task Description: In 2020, two scenarios will be evaluated: (i) the recognition of all specimens singing in a long sequence (up to one hour) of raw soundscapes

that can contain tens of birds singing simultaneously, and (ii) chorus source separation in complex soundscapes that were recorded in stereo at very high sampling rate (250 kHz SR). For the first scenario, participants will be asked to provide time intervals of recognized singing birds. Participants will be allowed to use any of the provided metadata complementary to the audio content (`.wav` format, 44.1 kHz, 48 kHz, or 96 kHz sampling rate). The task is focused on developing real-world applicable solutions and therefore requires participants to submit single models trained on none other than the mono-species recordings provided as training data. For the second task on stereophonic recordings, the goal will be to determine the species singing in chorus simultaneously during a time interval. In contrast to task one, the challengers are invited to run automatic source separation before or jointly to the bird species classification, taking advantage of the multi-channel recordings. Participants will be allowed to use any other data than the provided recordings, but will have to provide the scripts to check that their solution is fully automatic. For both tasks, the evaluation measure will be the classification mean average precision (c-mAP, [39]).

4 GeoLifeCLEF 2020 Challenge: Location-Based Prediction of Species Based on Environmental and Occurrence Data

Motivation: Automatic prediction of the list of species most likely to be observed at a given location is useful for many scenarios related to biodiversity management and conservation. First, it could improve species identification tools (whether automatic, semi-automatic or based on traditional field guides) by reducing the list of candidate species observable at a given site. More generally, this could facilitate biodiversity inventories through the development of location-based recommendation services (*e.g.* on mobile phones), encourage the involvement of citizen scientist observers, and accelerate the annotation and validation of species observations to produce large, high-quality data sets. Last but not least, this could be used for educational purposes through biodiversity discovery applications with features such as contextualized educational pathways.

Data Collection: The challenge will rely on a collection of millions of occurrences of plants and animals in the US and France (primarily from GBIF [10], iNaturalist [13], Pl@ntNet [20] and a few expert collections). In addition to geo-coordinates and species name, each occurrence will be matched with a set of geographic images characterizing the local landscape and environment around the occurrence. In more detail, this will include: (i) high resolution (about $1\,m^2$/pixel) remotely sensed imagery (from NAIP [18] for the US and from IGN [2] for France), (ii) bio-climatic rasters ($1\,km^2$/pixel, from WorldClim [24]) and (iii), land cover rasters ($30\,m^2$/pixel from NLCD [17] for the US, $10\,m^2$/pixel from CESBIO [3] for France).

Task Description: The occurrence dataset will be split in a training set with known species name labels and a test set used for the evaluation. For each

occurrence (with geographic images) in the test set, the goal of the task will be to return a candidate set of species with associated confidence scores. The evaluation metrics will be the top-K accuracy (for different values of K) and a set-based prediction metric which will be specified later.

5 SnakeCLEF 2020 Challenge: Image-Based Snake Identification

Motivation: Developing a robust system for identifying species of snakes from photographs is an important goal in biodiversity and global health. With over half a million victims of death and disability from venomous snakebite annually, understanding the global distribution of the $>3,700$ species of snakes and differentiating species from images (particularly images of low quality) will significantly improve epidemiology data and treatment outcomes. The goals and usage of image-based snake identification are complementary with those of other challenges: classifying snake species in images, predicting the list of species that are the most likely to be observed at a given location, and eventually developing automated tools that can facilitate integration of changing taxonomies and new discoveries.

Data Collection: Images of about 100 snake species from all around the world (between 300 and 150,000 images per species) will be aggregated from different data sources (including iNaturalist [13]). This will extend the dataset used in a previous challenge [22] hosted on the AICrowd platform. The distribution of the number of images between the classes is highly imbalanced.

Task Description: Given the set of images and corresponding geographic location information, the goal of the task will be to return for each image a ranked list of species sorted according to the likelihood that they are in the image and might have been observed at that location.

6 Timeline and Registration Instructions

All information about the timeline and participation in the challenges is provided on the LifeCLEF 2020 web pages [16]. The system used to run the challenges (registration, submission, leaderboard, etc.) is the AIcrowd platform [1].

7 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 2020 edition of the lab will support this vision and will include the following innovations:

- SnakeCLEF, a challenging new task related to the identification of snake species, will be introduced. It will be organized by the University of Geneva, Switzerland.
- The PlantCLEF task will focus on a brave new challenge: the identification of plant pictures based on a training set of digitized herbarium sheets (cross-domain image classification).
- The BirdCLEF challenge will be enriched with new annotated soundscape data, and with challenging new high sampling rate stereophonic recordings (from SEAMED, SMILES ANR, and SABIOD [21] projects).
- The GeoLifeCLEF challenge will be enriched with new plant and animal occurrences from two continents and high resolution remotely sensed imagery.

The results of this challenge will be published in the proceedings of the CLEF 2020 conference [5] and in the CEUR-WS workshop proceedings [4].

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).

References

1. AIcrowd. https://www.aicrowd.com/
2. BD ORTHO. http://professionnels.ign.fr/bdortho
3. Carte d'ccupation des sols 2016. http://osr-cesbio.ups-tlse.fr/~oso/posts/2017-03-30-carte-s2-2016/
4. CEUR-WS. http://ceur-ws.org/
5. CLEF 2020. https://clef2020.clef-initiative.eu/
6. Convention on Biodiversity. https://www.cbd.int/
7. Cornell Lab of Ornithology. https://www.birds.cornell.edu/home/
8. e-ReColNat. https://www.recolnat.org/
9. Encyclopedia of Life. https://eol.org/
10. Global Biodiversity Information Facility. https://www.gbif.org/
11. iDigBio. https://www.idigbio.org/
12. ImageCLEF. http://www.imageclef.org/
13. iNaturalist. https://www.inaturalist.org/
14. L'Herbier IRD de Guyane. http://publish.plantnet-project.org/project/caypub
15. LifeCLEF. http://www.lifeclef.org/
16. LifeCLEF 2020. https://www.imageclef.org/lifeclef2020
17. Multi-Resolution Land Characteristics Consortium. https://www.mrlc.gov/
18. National Agricultural Imagery Program. https://www.fsa.usda.gov/programs-and-services/aerial-photography/imagery-programs/naip-imagery/
19. OekoFor. http://oekofor.de
20. Pl@ntNet. https://plantnet.org/
21. Scaled Acoustic BIODiversity (SABIOD) Platform. http://sabiod.telemeta.org/
22. Snake Species Identification Challenge. https://www.aicrowd.com/challenges/snake-species-identification-challenge
23. The FAIR Data Principles. https://www.force11.org/group/fairgroup/fairprinciples

24. WorldClim Bioclimatic Variables. https://www.worldclim.org/bioclim
25. Bonnet, P., et al.: Plant identification: experts vs. machines in the era of deep learning. In: Joly, A., Vrochidis, S., Karatzas, K., Karppinen, A., Bonnet, P. (eds.) Multimedia Tools and Applications for Environmental & Biodiversity Informatics. MSA, pp. 131–149. Springer, Cham (2018). https://doi.org/10.1007/978-3-319-76445-0_8
26. Cai, J., Ee, D., Pham, B., Roe, P., Zhang, J.: Sensor network for the monitoring of ecosystem: bird species recognition. In: 2007 3rd International Conference on Intelligent Sensors, Sensor Networks and Information, ISSNIP 2007 (2007). https://doi.org/10.1109/ISSNIP.2007.4496859
27. Gaston, K.J., O'Neill, M.A.: Automated species identification: why not? Philos. Trans. R. Soc. London B Biol. Sci. **359**(1444), 655–667 (2004)
28. Ghazi, M.M., Yanikoglu, B., Aptoula, E.: Plant identification using deep neural networks via optimization of transfer learning parameters. Neurocomputing **235**, 228–235 (2017)
29. Glotin, H., Clark, C., LeCun, Y., Dugan, P., Halkias, X., Sueur, J.: Proceedings of the 1st workshop on Machine Learning for Bioacoustics - ICML4B. ICML, Atlanta USA (2013)
30. Glotin, H., LeCun, Y., Artiéres, T., Mallat, S., Tchernichovski, O., Halkias, X.: Proceedings of the Neural Information Processing Scaled for Bioacoustics, from Neurons to Big Data. NIPS International Conference, Tahoe, USA (2013). http://sabiod.org/nips4b
31. 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)
32. Goëau, H., et al.: The ImageCLEF 2013 plant identification task. In: CLEF, Valencia, Spain (2013)
33. Goëau, H., et al.: The ImageCLEF 2011 plant images classification task. In: CLEF 2011 (2011)
34. Goëau, H., et al.: ImageCLEF2012 plant images identification task. In: CLEF 2012, Rome (2012)
35. Goëau, H., Glotin, H., Planqué, R., Vellinga, W.P., Stefan, K., Joly, A.: Overview of BirdCLEF 2018: monophone vs. soundscape bird identification. In: CLEF Working Notes 2018 (2018)
36. Goëau, H., Glotin, H., Vellinga, W., Planqué, B., Joly, A.: LifeCLEF bird identification task 2017. In: Working Notes of CLEF 2017 - Conference and Labs of the Evaluation Forum, Dublin, Ireland, 11–14 September 2017 (2017)
37. Goëau, H., et al.: 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)
38. Joly, A., et al.: Interactive plant identification based on social image data. Ecol. Inform. **23**, 22–34 (2014)
39. Joly, A., et al.: Overview of LifeCLEF 2018: a large-scale evaluation of species identification and recommendation algorithms in the era of AI. In: Bellot, P., et al. (eds.) CLEF 2018. LNCS, vol. 11018, pp. 247–266. Springer, Cham (2018). https://doi.org/10.1007/978-3-319-98932-7_24
40. Kahl, S., Stöter, F.R., Glotin, H., Planqué, R., Vellinga, W.P., Joly, A.: Overview of BirdCLEF 2019: large-scale bird recognition in soundscapes. In: CLEF (Working Notes) (2019)

41. 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)
42. Lee, S.H., Chan, C.S., Remagnino, P.: Multi-organ plant classification based on convolutional and recurrent neural networks. IEEE Trans. Image Process. **27**(9), 4287–4301 (2018)
43. Towsey, M., Planitz, B., Nantes, A., Wimmer, J., Roe, P.: A toolbox for animal call recognition. Bioacoustics **21**(2), 107–125 (2012)
44. 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. J. Acoust. Soc. Am. **123**, 2424 (2008)
45. Van Horn, G., et al.: The inaturalist species classification and detection dataset. In: CVPR (2018)
46. Wäldchen, J., Mäder, P.: Machine learning for image based species identification. Methods in Ecol. Evol. **9**(11), 2216–2225 (2018)
47. Wäldchen, J., Rzanny, M., Seeland, M., Mäder, P.: Automated plant species identification–trends and future directions. PLoS Comput. Biol. **14**(4), e1005993 (2018)
48. Yu, X., Wang, J., Kays, R., Jansen, P.A., Wang, T., Huang, T.: Automated identification of animal species in camera trap images. EURASIP J. Image Video Process. **2013**(1), 1–10 (2013). https://doi.org/10.1186/1687-5281-2013-52