Can we estimate insect identification ease degrees from their identification key paths?

Zakaria Saoud

Can we estimate insect identification ease degrees from their identification key paths?

Zakaria Saoud[a,*]

[a] *Centre d'Ecologie et des Science de la Conservation, UMR 7204 CNRS-MNHN-SU, Muséum national d'Histoire naturelle, 61 rue Buffon, 75005 Paris, France*

Abstract

Identification key represents a powerful tool, which allows users to identify biological species, such as animals, insects, and plants. This key contains several steps leading to the corresponding species name. In this paper, we propose a new method for estimating insects' identification ease from their charecteristics values (CV). To realize that, we have investigated; 1- the variations of the identification ease level (IES) of the SPIPOLL insects, 2- the relation between the CV and the IES. The obtained results showed that the CV can be used to estimate the IES of the SPIPOLL insects.

Keywords:
Identification Key, Identification tools, Ease estimation, Citizen Sciences, SPIPOLL

1. Introduction

Species identification is a formal process in many biological systematics. Several approaches for species identification have been proposed in the litterateur [20] [5] [22] [25]. These approaches can be classified into five main categories: 1- Morphometric-based approaches [4] [7] [11] [6], 2- DNA barcoding-based approaches [17] [13] [2] [9] [10], 3- Crowd sourcing-based approaches [24] [21] [23] [3], 4- Computer vision and machine learning based approaches [1] [18] [15] [26]

*Corresponding author
Email address: zakaria.saoud@mnhn.fr (Zakaria Saoud)

[14] and 5- key-based approaches [8] [19] [12] [16]. In Morphometric-based approaches, the species are identified according to their length, width, height or the geometry of their morphological structures. DNA barcoding based approaches use information of one or a few gene regions (short genetic marker) to identify the species. In Crowdsourcing-based approaches, citizen scientists and biologists identify the species in a collaborative way, based on different computer-based tools .

Computer vision and machine learning based approaches identify the species automatically using the extracted relevant features from the species images. In the first step, the species images are transformed to a tractable domain and stored in the training dataset. Then, an unsupervised learning algorithm is trained using the training dataset and will be used to identify the species. Key-based approaches allow users to identify the species through an identification key that is considered dichotomous if there is only two descriptions in each step, and polytomous in the other case. Identification key takes the form of a decision tree and is designed to guide the users to find the corresponding species name, through successive identification steps. In each step, the user should answer a question about one characteristic for identifying its description, until finding the corresponding species name. In dichotomous identification key, the users are obliged to follow the order of the proposed questions, while in computer assisted identification key (such in Xper3 identification key) the users can choose to answer any question.

Furthermore, Identification key represents a powerful tool for the teaching of biodiversity for the beginners, and for accelerating the species identification process in citizen sciences website like the SPIPOLL [1]. In this website, the users are asked to update pictures of insects pollinators and flowering plants and identifying them directly by choosing the taxon name, or with the aid of a visual interactive identification key, generated by the Xper3 web platform. Finally a group of expert users will validate the identifications. Figure 1 shows

[1] http://www.spipoll.org/

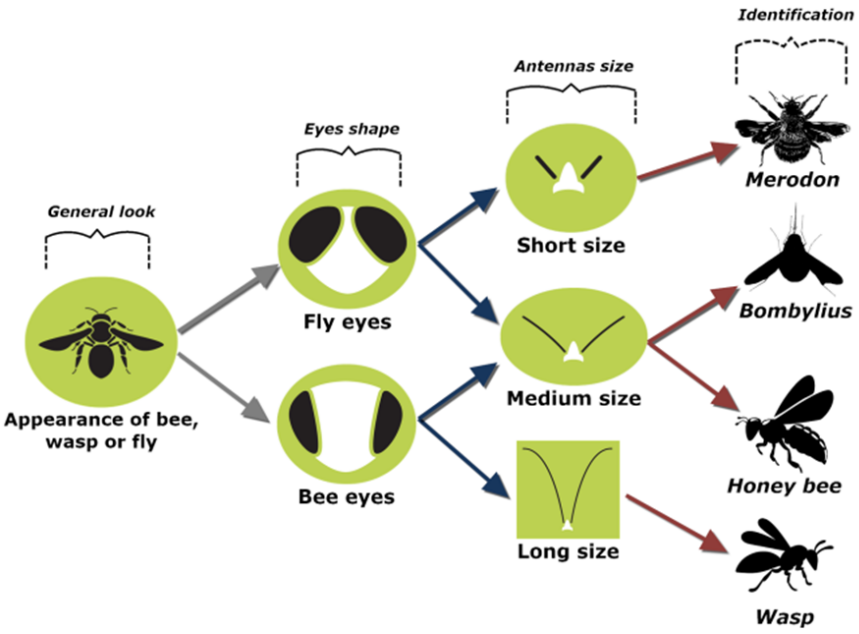

Figure 1: An example of the use of Xper3 identification key.

an example of the use of Xper3 SPIPOLL identification key.

The Xper3 represents a collaborative knowledge base platform that allows creating interactive identification keys. In the SPIPOLL, the set of required steps can represent an identification path (IP) leading to the species name. The length of IP varies from species to another, according to the characteristics values (CV) of the species. Figure 2 shows three IPs of three different flies from the SPIPOLL database. From this figure, we can see that the two last flies that share the most characteristics have longer IPs than the first fly that has specific characteristics (presence of metallic reflections). The two last flies might considered also harder for identification than the first fly, due the resemblance between them. Hence, we can assume that the presence of certain CV on the IP can affect the ease identification degree of the insect. In this paper, we propose a new method for estimating insects' identification ease from their CV.

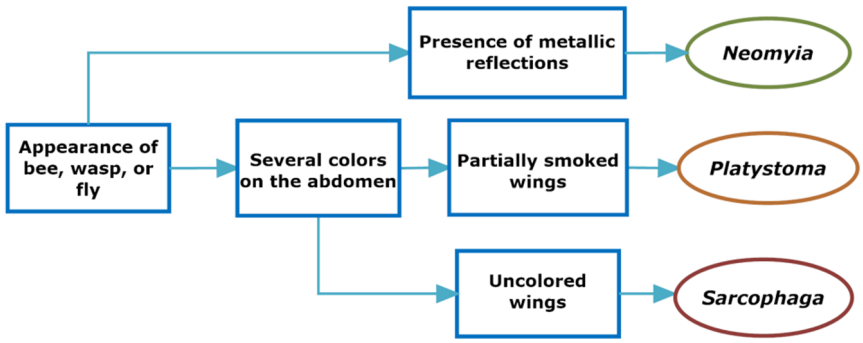

Figure 2: Three IPs of three different flies from the SPIPOLL database.

2. Do the SPIPOLL insects have the same level of identification ease?

In the SPIPOLL, each insect picture is first identified by the observer and then validated by an expert user. This validation represents the final and the true identity of the insect. Therefore, we calculated for each insect, the rate of correct identification. For each picture, we compare its initial identification with the corresponding validation (the final identification). Then, we obtained a score which represents the identification ease score (IES) of the insect. Figure 3 shows the obtained ground truth identification ease scores of the SPIPOLL insects.

3. Is there any relation between the CV and insect IES?

As we said in the introduction, the presence or the absence of certain CV can facilitate the identification process or make it harder. To verify this hypothesis, we have applied a clustering algorithm on a set of 134 insects according to their CV. These insects have diffirent IES and belong to different insect families. First we have constructed a binary matrix wich represents the insects and their CV. In total we used 193 CV. Then we applied k-means algorithm on the obtained matrix and we obtained 6 insects' clusters. This number of clusters represents the ideal number of clusters obtained by the Elbow method as shown in figure 4 .

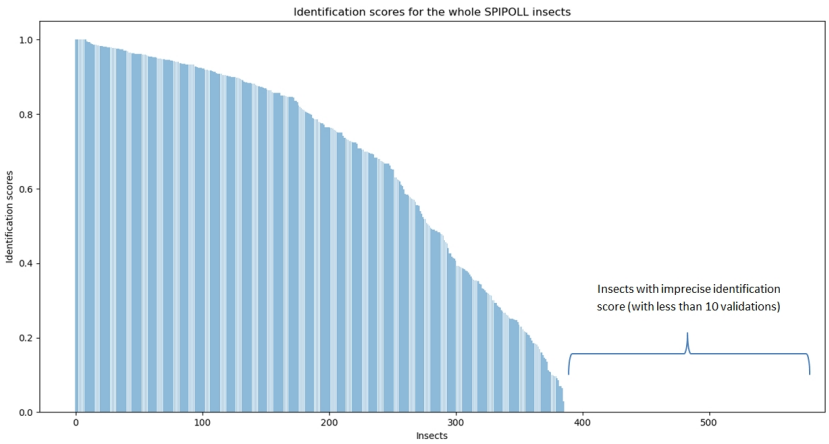

Figure 3: The ground truth identification eases scores of the SPIPOLL insects. The IES of Insects with fewer than ten validations are considered imprecise.

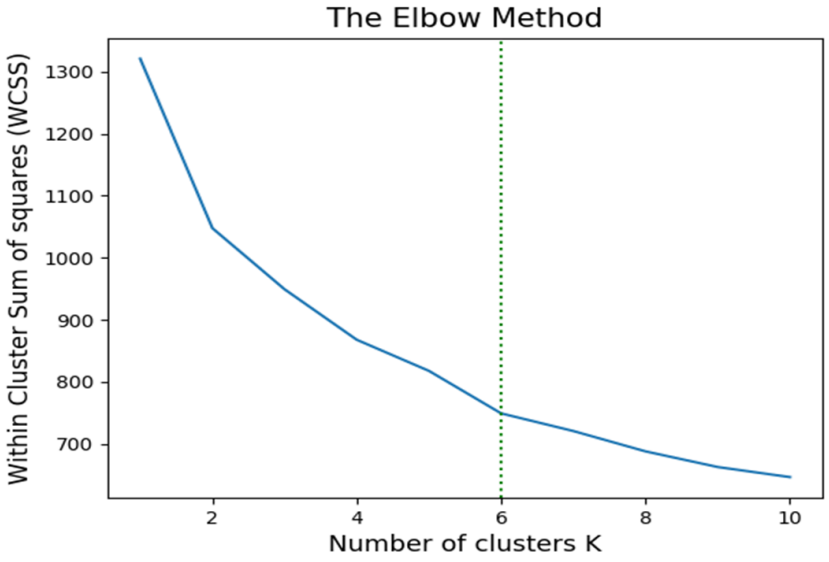

Figure 4: The ideal number of clusters estimated by the Elbow method.

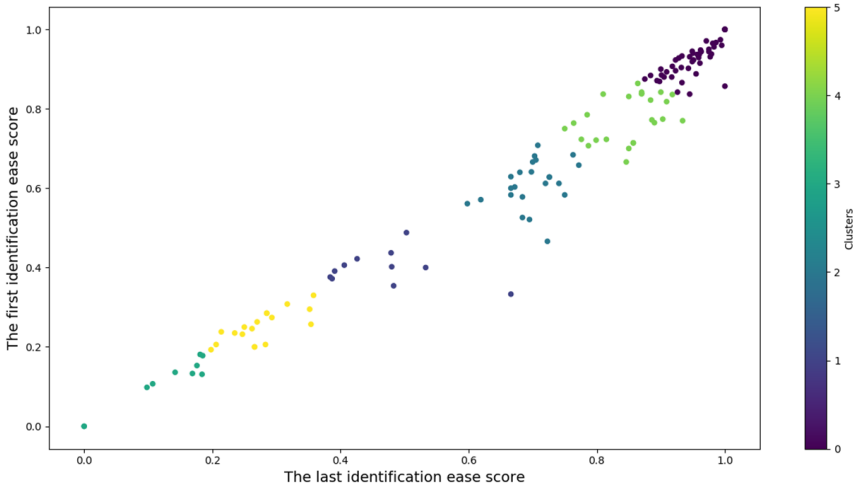

Figure 5: The relation between the IES and the insects clusters. The insects with the same color belong to the same cluster.

Each cluster contains a set of insects with the similar CV. To test if there is any relation between the obtained clusters and the IES, we have designed a graph which shows the variation between the first IES and the last IES for each insect. The first IES is calculated using the first given correct identification. The last IES is calculated using the last given correct identification. Then we have colored the insects of the same clusters with the same color. Figure 5 show the relation between the IES and the insects clusters. From this figure, we can see that for the most insects, the two IES is almost the same. We can see also that the insects of the same cluster have close IES, which prove the relation between the IES and the CV.

4. Can the CV estimate the insect IES?

In our case, each characteristic value CV_i can be used to describe one or more insects from the SPIPOLL. Figure 6 show the relation between the CV and the SPIPOLL insects. We note INS_i the set of insects who have the characteristic value CV_i. Hence, we can calculate for each an IES as follows:

$$IES\left(CV_i\right) = \frac{\sum_{IN_j \in INS_i} IES\left(IN_j\right)}{[INS_i|}$$

Where: IN_j : is an insect belonging the insects set INS_i . $|IN_j|$: is the number of insects that have the characteristic value CV_i .

The obtained IES of CV will be used to calculate an estimated IES for each insect IN_i as follows:

$$IES_{estimated}\left(IN_i\right) = \frac{\sum_{CV_j \in CV_{IN_i}} IES\left(CV_j\right)}{|CV_{IN_i}|}$$

Where: $CV_{\mathrm{IN_i}}$: is the set of characteristics values of the insect IN_i .

$|CV_{\mathrm{IN_i}}|$: is the number of characteristics values of the insect IN_i.

To verify if the CV can estimate the insect IES, we compare the ease identification estimated ranking with the ground truth ease identification rankings for a set of 134 insects. For each insect, we generate its ground truth IES, by comparing its identifications with the corresponding validations. Then, we calculate the ease identification estimated score for each insect using its CV ease identification scores. The comparison between every two scores is based on two evaluation metrics: The Spearman's rho and Kendall's tau, which measure the correlation between the ideal ranking (the ground truth ranking) and the estimated ranking. We calculate Kendall's Tau and Spearman's rho for the 50, 80, 100 and all top ranked insects. Figure Figure 7 illustrates the statistical results regarding Spearman's rho and Kendall's tau distance between the ideal and the estimated rankings.

From this figure, we can see the moderate correlation between the ideal and the most estimated ranking. We can see also that the Spearman's rho correlations reduce when the number of ranked insects increases, unlike the Kendall's tau correlations that increases when the number of ranked insects increases. The two measures give the same correlations for the top 100 ranked insects. This result proves that the CV of insects has a moderate correlation with the insect IES. In the other hand, the Spearman's Rho gives a high correlation for the top 80 ranked insects. Thus, we can see that the CV can be used to

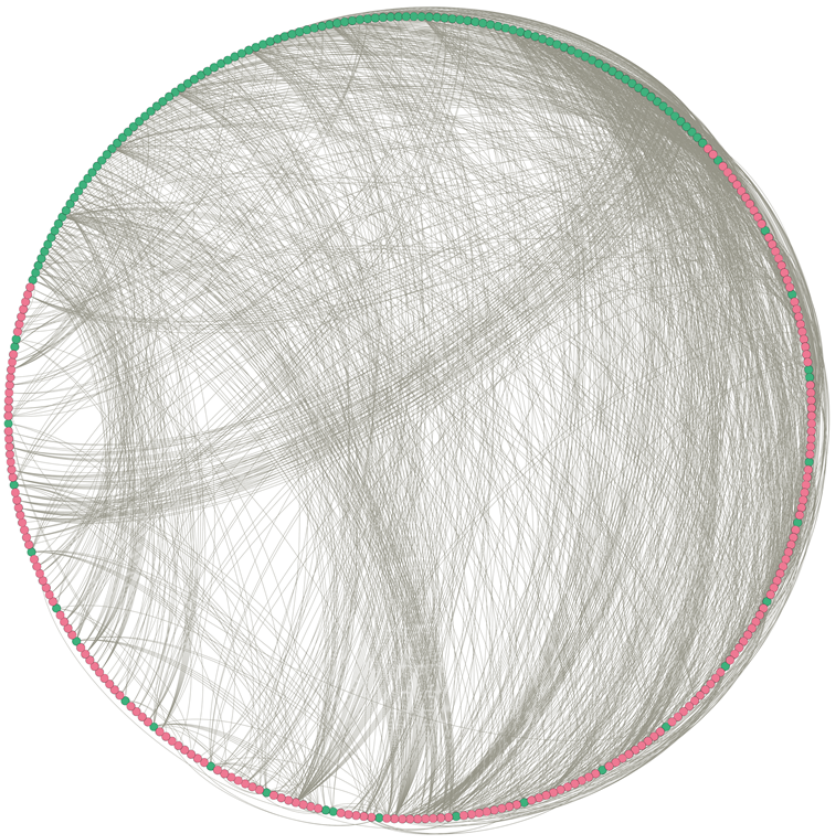

Figure 6: The relation between the CV and the SPIPOLL insects. CV are with red color and insects are with green color.

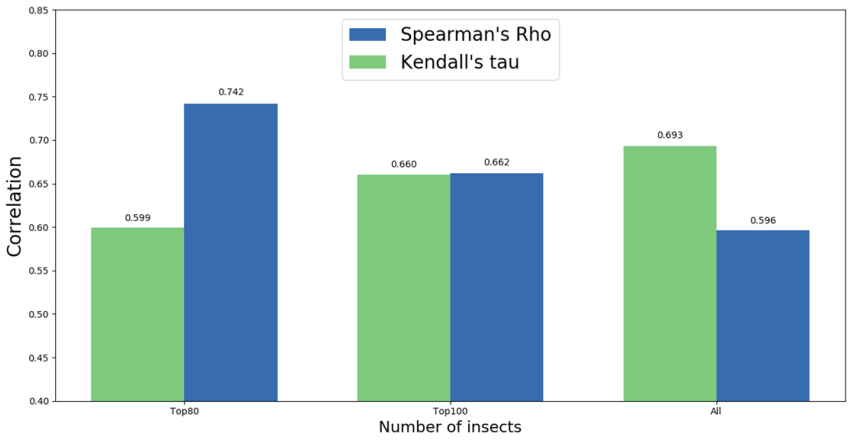

Figure 7: Spearman's rho and and Kendall's tau distance between the ideal and the estimated rankings..

estimate the identification ease degree of the SPIPOLL insects.

5. Conclusion

In this paper, we have studied the relationship between the insect identification ease degree and the identification key. 193 characteristics values (CV) paths have been extracted for a group of 134 insects from the SPIPOLL database. For each CV we have calculated an ease identification score (IES) according to the average of IES of the insects that have this CV. Then for each insect, we have used the obtained IES of the CV to calculate an estimated IES to construct an estimated ranked list of insects. Finally, we have compared the obtained estimated ranking list with the ground truth ranked list of insects. The obtained results showed a high correlation between the estimated IES and the ground truth IES, which prove that the CV can be used to estimate the IES of the SPIPOLL insects.

6. Acknowledgments

This study was supported by l'Agence Nationale de la Recherche (ANR), the National Museum of Natural History (MNHN) and the Office for Insects and their Environment (Opie). We thank Mathieu De Flores (Opie) for his precious help with all the entomological work and Grégoire LOÏS for providing us the SPIPOLL data. Further, we are grateful to all active SPIPOLL users who helped with data collection: Barbara Mai, Gilles Jardinier, Sagittaire06, MichelMarly, Ascalaf07, Janmar, Prisca and Leonlebourdoun.

References

[1] Barré, P., Stöver, B. C., Müller, K. F., and Steinhage, V. (2017). Leafnet: A computer vision system for automatic plant species identification. *Ecological Informatics*, 40:50–56.

[2] Carvalho, D. C., Palhares, R. M., Drummond, M. G., and Frigo, T. B. (2015). Dna barcoding identification of commercialized seafood in south brazil: a governmental regulatory forensic program. *Food Control*, 50:784–788.

[3] Carvalko, J. R. and Morris, C. (2015). Crowdsourcing biological specimen identification: Consumer technology applied to health-care access. *IEEE Consumer Electronics Magazine*, 4(1):90–93.

[4] Cope, J. S., Corney, D., Clark, J. Y., Remagnino, P., and Wilkin, P. (2012). Plant species identification using digital morphometrics: A review. *Expert Systems with Applications*, 39(8):7562–7573.

[5] Dahms, H.-U., Fornshell, J. A., and Fornshell, B. J. (2006). Key for the identification of crustacean nauplii. *Organisms Diversity & Evolution*, 6(1):47–56.

[6] De Luna, E. and Gómez-Velasco, G. (2008). Morphometrics and the identification of braunia andrieuxii and b. secunda (hedwigiaceae, bryopsida). *Systematic Botany*, 33(2):219–228.

[7] Francoy, T. M., Wittmann, D., Drauschke, M., Müller, S., Steinhage, V., Bezerra-Laure, M. A., De Jong, D., and Gonçalves, L. S. (2008). Identification of africanized honey bees through wing morphometrics: two fast and efficient procedures. *Apidologie*, 39(5):488–494.

[8] Gaubert, P., Chalubert, A., and Dubus, G. (2008). An interactive identification key for genets and oyans (carnivora, viverridae, genettinae, genetta spp. and poiana spp.) using xper^2. *Zootaxa*, 1717:39–50.

[9] Hebert, P. D., Stoeckle, M. Y., Zemlak, T. S., and Francis, C. M. (2004). Identification of birds through dna barcodes. *PLoS biology*, 2(10):e312.

[10] Holmes, B. H., Steinke, D., and Ward, R. D. (2009). Identification of shark and ray fins using dna barcoding. *Fisheries Research*, 95(2-3):280–288.

[11] Jacquiet, P., Cabaret, J., Cheikh, D., and Thiam, E. (1996). Identification of haemonchus species in domestic ruminants based on morphometrics of spicules. *Parasitology research*, 83(1):82–86.

[12] Kirchoff, B. K., Leggett, R., Her, V., Moua, C., Morrison, J., and Poole, C. (2011). Principles of visual key construction—with a visual identification key to the fagaceae of the southeastern united states. *AoB Plants*, 2011.

[13] Kress, W. J., Wurdack, K. J., Zimmer, E. A., Weigt, L. A., and Janzen, D. H. (2005). Use of dna barcodes to identify flowering plants. *Proceedings of the National Academy of Sciences*, 102(23):8369–8374.

[14] Kumar, N., Belhumeur, P. N., Biswas, A., Jacobs, D. W., Kress, W. J., Lopez, I. C., and Soares, J. V. (2012). Leafsnap: A computer vision system for automatic plant species identification. In *European Conference on Computer Vision*, pages 502–516. Springer.

[15] Larios, N., Deng, H., Zhang, W., Sarpola, M., Yuen, J., Paasch, R., Moldenke, A., Lytle, D. A., Correa, S. R., Mortensen, E. N., et al. (2008). Automated insect identification through concatenated histograms of local appear-

ance features: feature vector generation and region detection for deformable objects. *Machine Vision and Applications*, 19(2):105–123.

[16] Martellos, S. (2010). Multi-authored interactive identification keys: The frida (friendly identification) package. *Taxon*, 59(3):922–929.

[17] Meier, R., Shiyang, K., Vaidya, G., and Ng, P. K. (2006). Dna barcoding and taxonomy in diptera: a tale of high intraspecific variability and low identification success. *Systematic biology*, 55(5):715–728.

[18] Nikolaou, N., Sampaziotis, P., Aplikioti, M., Drakos, A., Kirmitzoglou, I., Argyrou, M., Papamarkos, N., and Promponas, V. J. (2010). Vestis: A versatile semi-automatic taxon identification system from digital images. EUT Edizioni Università di Trieste.

[19] Nimis, P., Riccamboni, R., and Martellos, S. (2012). Identification keys on mobile devices: The dryades experience. *Plant Biosystems-An International Journal Dealing with all Aspects of Plant Biology*, 146(4):783–788.

[20] Norton, G. A., Patterson, D. J., and Schneider, M. (2012). Lucid: A multimedia educational tool for identification and diagnostics. *International Journal of Innovation in Science and Mathematics Education (formerly CAL-laborate International)*, 4(1).

[21] Rahman, M., Blackwell, B., Banerjee, N., and Saraswat, D. (2015). Smartphone-based hierarchical crowdsourcing for weed identification. *Computers and Electronics in Agriculture*, 113:14–23.

[22] Ribeiro, R. D., Cardoso, D. B. O. S., and de Lima, H. C. (2015). A new species of hymenaea (leguminosae: Caesalpinioideae) with a revised identification key to the genus in the brazilian atlantic forest. *Systematic Botany*, 40(1):151–156.

[23] Siddharthan, A., Lambin, C., Robinson, A.-M., Sharma, N., Comont, R., O'mahony, E., Mellish, C., and Wal, R. V. D. (2016). Crowdsourcing without

a crowd: Reliable online species identification using bayesian models to minimize crowd size. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 7(4):45.

[24] Silvertown, J., Harvey, M., Greenwood, R., Dodd, M., Rosewell, J., Rebelo, T., Ansine, J., and McConway, K. (2015). Crowdsourcing the identification of organisms: A case-study of ispot. *ZooKeys*, (480):125.

[25] Sweeney, C. A. (2004). A key for the identification of stomata of the native conifers of scandinavia. *Review of Palaeobotany and Palynology*, 128(3-4):281–290.

[26] Wang, J., Lin, C., Ji, L., and Liang, A. (2012). A new automatic identification system of insect images at the order level. *Knowledge-Based Systems*, 33:102–110.