Exploring carnivorous plant habitats based on images from social media

Rafael Blanco, Zujany Salazar, Tobias Isenberg

Exploring Carnivorous Plant Habitats based on Images from Social Media

Rafael Blanco* Télécom SudParis
Zujany Salazar* Télécom SudParis
Tobias Isenberg* Inria

Abstract

We visualize species habitat distribution information based on geo-located images posted on social media. For the example of carnivorous plants, we use published image data to produce interactive maps of the spatial distribution of different species/genuses, histograms of the elevations at which they grow, and plots of the temporal distribution of the photographs. We further discuss the mismatch between our distribution maps and traditionally established maps as well as further possibilities for research with our data.

1 Motivation and Related Work

Many species of plants and animals are endangered today (e. g., see the IUCN Red List 2012; `https://www.iucnredlist.org/`). This status also applies to many carnivorous plants, and past studies have quantified the conservation threats of the different genus and species [4], providing information to prioritize certain areas of conservation. Yet information about the distribution of the different species is often difficult to obtain. Researchers use environment data and current habitat sightings, to apply species distribution modeling (SDM), which uses statistical models to generate an estimate of the plant distribution around the world. In contrast to dedicated campaigns to document species distribution in the wild, we explore the use of geo-located images posted in social media. While the use of geo-located social media images and other posts has been explored in the past for the study of popular places [3] and events [1, 5], we use the social images posted on Panoramio and Flickr not as singular events but as evidence for the existence of a specific plant at a location, providing evidence for the habitat of the depicted species.

2 Data Acquisition

We collected a dataset by searching Panoramio[1] and Flickr for geo-tagged images whose label or description included at least one from a series of search terms. These search terms included the Latin genus and family names, the terms "carnivorous plant(s) or similar, as well as a number of common names for different species in several languages including Chinese, Danish, Dutch, English, French, German, Italian, Japanese, Norwegian, Polish, Portuguese, Russian, Spanish, and Swedish. This resulted in more than 28,700 candidate images. For each of the found images, we looked at both the image itself and its location on the map using an online satellite map, to manually determine whether the images showed a true habitat and whether the location data was believable (e. g., Fig. 24). For example, we removed image locations in the middle of urban areas. As some people posted many images of plants kept at home or used the same location that was not plausibly a habitat location, we removed certain user IDs and locations from consideration to speed-up the inspection. We then recorded resulting locations, resulting in a list of more than 8,800 entries. Out of these, approx. 4,600 are within a radius of approx. 250 m another location with the same species and can thus be considered to be duplicates. For each plant location, we recorded its scientific name,[2] its geographic position on the map, its elevation (based on an elevation look-up for the geographic position), the time the picture was taken, the social media service and the respective image ID, text description of the geographic location such as area name and country, the name and ID of the person who uploaded the images, and the image itself.

Figure 1: Screenshot of our data exploration interface, with markers colored by genus and image thumbnails shown at the bottom.

3 Visualization System

We designed a map-based (using Google's Maps API) data exploration interface that indicates the location of each plant in the dataset with a marker. We provide two visualization modes: a general data point exploration mode and a plant species or genus distribution mode. The *exploration mode* (e. g., Fig. 1) has a filter panel to filter for plant characteristics (genus, species) and/or image data (location, date it was taken, social network). In this mode we provide an image carousel that shows thumbnails of the images of the currently displayed data points. In the *distribution mode* (Fig. 12), we use circles around plant locations to generate distribution maps (e. g., Fig. 3), without applying any statistical methods. We use circles whose size can be changed based on the current map magnification. In addition, as such maps could potentially be published, we add a small random offset to obfuscate the specific locations to prevent potential poaching. Furthermore, we allow users to generate interactive plots of the image dates and elevation histograms using Plotly (e. g., Fig. 2).

Our application relies on Flask for Python, HTML5, CSS3, and JavaScript. We use SQLAlchemy to manage the database that has the information of each image of a plant. Also, we use the Flickr API to allow the users to look up each entry directly on Flickr.

4 Data Exploration Strategies and Case Studies

Our visualization system allows us to investigate the collected data in interesting ways. Primarily it is straight-forward to examine the geographic distribution of different species and to compare this with known distribution maps. For example, we can find that the distribution of *Drosera rotundifolia* in our data matches with the distribution map on Wikipedia (Fig. 18). However, we also see that our data does not support the full distribution on the established maps. This fact is likely due to a strong bias in our data: we only have pictures from places that people are likely to live at or travel to (Fig. 21, 21(c)). We thus have no evidence for the distribution of *Drosera rotundifolia* in Siberia, for example. In another example, we can also see that our data supports the distribution of *Sarracenia purpurea* in North America (Fig. 19)—including the isolated population in the southern United States but excluding the less populated areas

*e-mail: {rblancog25 | zujany}@gmail.com, tobias.isenberg@inria.fr

[1] The Panoramio service has since been retired by Google.

[2] We generally used the Latin names provided in the descriptions. If no species name was provided or if it was wrong, we either classified the plant ourselves if we knew the species well, or we only recorded the plant's genus.

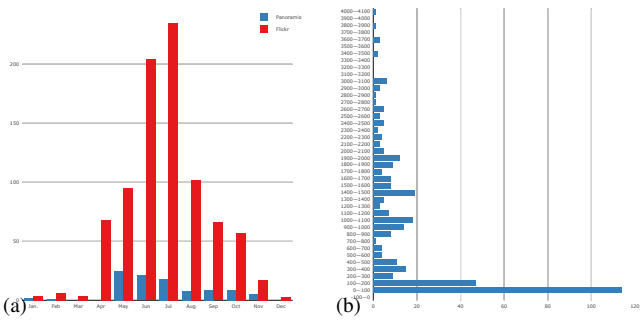


Figure 2: Examples of a generated data plots: (a) sightings of *Sarracenia purpurea* by month and (b) elevation histogram for *Nepenthes*.

of central and northern Canada. Yet we can also see populations of the species in Europe—not only well known places where the plant was introduced in Switzerland but also places in the UK, Scandinavia, and the Netherlands. Similarly, our map of the distribution of *Drosera arcturi* (Fig. 3) shows habitats not only in Australia but also in New Zealand and for *Darlingtonia californica* (Fig. 20) we see also additional locations in western and central California.

In addition, we need to address the inherent unreliability of the posted data—in particular the geographic location could be intentionally or unintentionally incorrect. We thus implemented verification mechanisms that provide information about the reliability of the locations. Specifically, we treat a location as more reliable if other social media users have posted pictures of other plants nearby. Based on a user-selectable search threshold (we experimented, e. g., with 1 km and 4 km; e. g., see Fig. 15 and 16), we search for nearby sites posted by other users and depict sites with a species match in green, sites with a genus match in yellow, sites with only other-genus plants nearby in orange, and the rest in red (other color maps for color-deficient users are possible). The process to compute these trustworthiness values essentially uses a hashed comparison of the sites using 9-neighborhood "buckets" and thus takes about 10 seconds to compute. We allow users to adjust the search radius and thus the trust level they want to work with, which likely depends on a given site and its number of data points. We also explored a verification based on the posts of a single person: We check if the same person posted multiple pictures from exactly the same GPS position—a sign for manipulated coordinates—and color the markers based on how many other pictures were posted a given location (Fig. 17).

In addition to the geographic distributions, we also analyzed abstract data. In particular, we support the analysis of the time the picture was taken and of the elevation of the location (e. g., Fig. 4 shows the number of entries in our dataset by year, and Fig. 5 shows monthly detail since 2003). We also analyzed the dates of the pictures within a year (e. g., Fig. 6 for all entries), and saw that sightings in the Northern Hemisphere with a peak in the middle of the calendar year are prevalent. Looking at specific plants, we can identify the different growth periods between northern and Southern Hemisphere plants (e. g., Fig. 10). The elevation plots (e. g., Fig. 7, 9) also show a quite different behavior of different species.

We note that our data relies on a correct classification of the species. While due to our manual process we are certain that we use the correct genus name for all entries, the same cannot be said for the species. Here we relied on the classification provided in the social media post (if any), and only re-classified some specific species which we knew well. A sizable portion of our entries (approx. 12%) thus only contains the genus of the respective plants locations. An expert botanist with a research background on these plants would certainly be able to address these challenges. The abstract data plots as well as the geographic maps are likely to support such an analysis, identifying outliers or non-classified plants in specific regions.

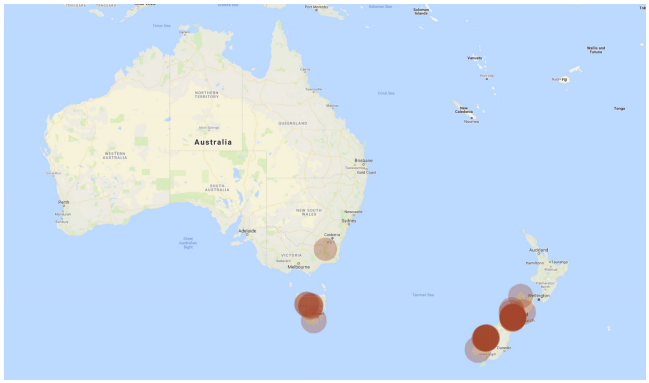


Figure 3: Example of a generated habitat map for *Drosera arcturi*.

5 Conclusion and Future Work

Our work showcases an interesting and, to the best of our knowledge, previously not explored use of social media data for longer-term "events" such as plant grows in a specific habitat/region. We explored both geographic as well as abstract data analysis techniques for this data, and it would be interesting in the future to also explore space-time representations [2]. We are also planning to investigate machine learning approaches to be able to identify flowers on the pictures, which would allow us to plot flowering periods of the plants. We also plan to investigate if ML approaches could support genus or species identification to be able to further check the data or automate the data collection process. The latter should also generally be integrated in the visualization tool, so that we can merge the currently separate processes of data collection, correction, and visualization.

We collected our dataset and built our visual exploration tool due to an interest in carnivorous plants, it can thus be used by enthusiasts to visit habitats. Our tool also has the potential to be used by plant biologists or botanists to study the habitat information we collected as well as by conservationists to become aware of unknown sites as well as the evolution of known sites. For the latter of the two applications, however, we point out that our dataset is limited because for some species we only have very few entries (e. g., currently 110 species with 5 entries or fewer each). So users need to be aware of small-number statistics issues when analyzing the data. In addition, the previously mentioned data biases need to be considered. Based on the protection status of many carnivorous plants, however, we will not make the tool or data publicly available to prevent poaching and will only share it with researchers in the application domain.

References

[1] G. Andrienko, N. Andrienko, P. Bak, S. Kisilevich, and D. Keim. Analysis of community-contributed space- and time-referenced data (example of Flickr and Panoramio photos). In *Proc. VAST (Posters)*, pp. 213–214. IEEE Computer Society, Los Alamitos, 2009. doi: 10.1109/VAST.2009.5333472

[2] B. Bach, P. Dragicevic, D. Archambault, C. Hurter, and S. Carpendale. A descriptive framework for temporal data visualizations based on generalized space-time cubes. *Computer Graphics Forum*, 36(6):36–61, 2017. doi: 10.1111/cgf.12804

[3] D. J. Crandall, L. Backstrom, D. Huttenlocher, and J. Kleinberg. Mapping the world's photos. In *Proc. WWW*, pp. 761–770. ACM, New York, 2009. doi: 10.1145/1526709.1526812

[4] D. E. Jennings and J. R. Rohr. A review of the conservation threats to carnivorous plants. *Biological Conservation*, 144(5):1356–1363, 2011. doi: 10.1016/j.biocon.2011.03.013

[5] S. Kisilevich, M. Krstajic, D. Keim, N. Andrienko, and G. Andrienko. Event-based analysis of people's activities and behavior using Flickr and Panoramio geotagged photo collections. In *Proc. Information Visualisation*, pp. 289–296. IEEE Computer Society, Los Alamitos, 2010. doi: 10.1109/IV.2010.94

Exploring Carnivorous Plant Habitats based on Images from Social Media

Additional material

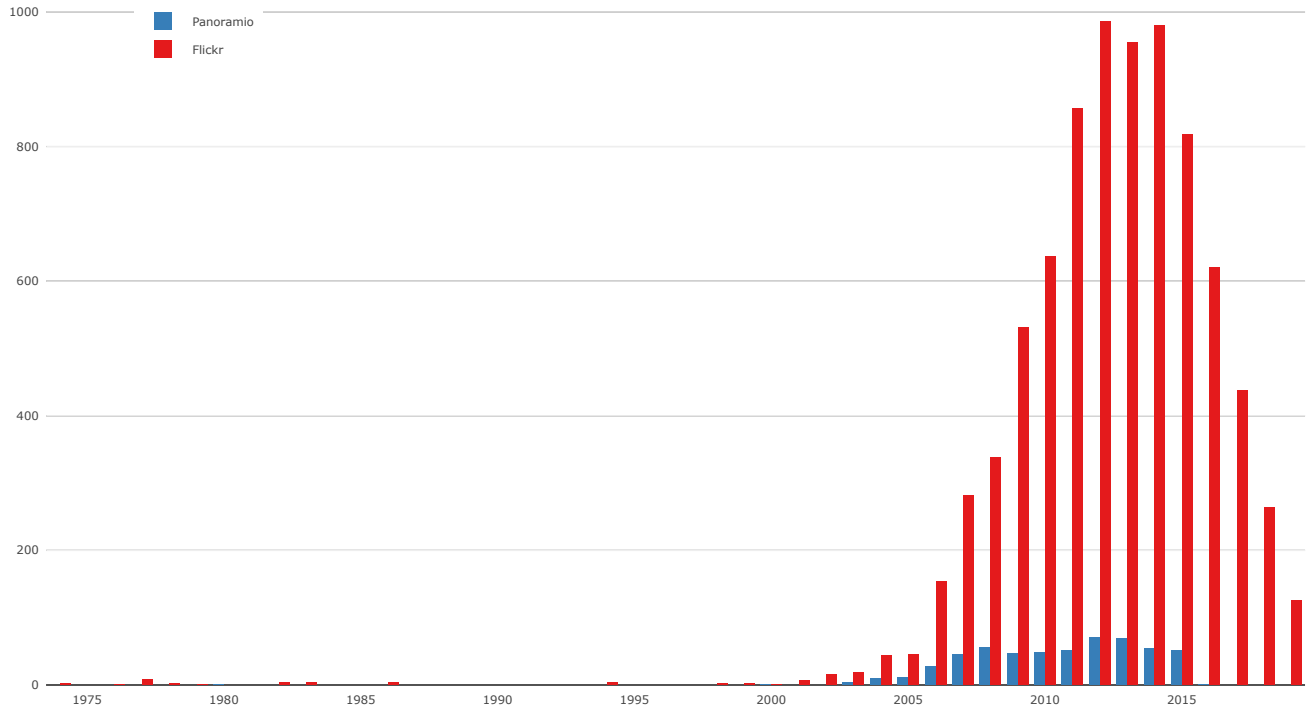


Figure 4: Date histogram by year and service, complete dataset. Interestingly, the number of posts per year declines after 2014, possibly to due the fact that people post their pictures only after some time and not directly when taking them and/or due to Flickr's changed upload policies which apparently have let quite a number people to delete their images from the service (see Fig. 23 for a graph that supports this hypothesis).

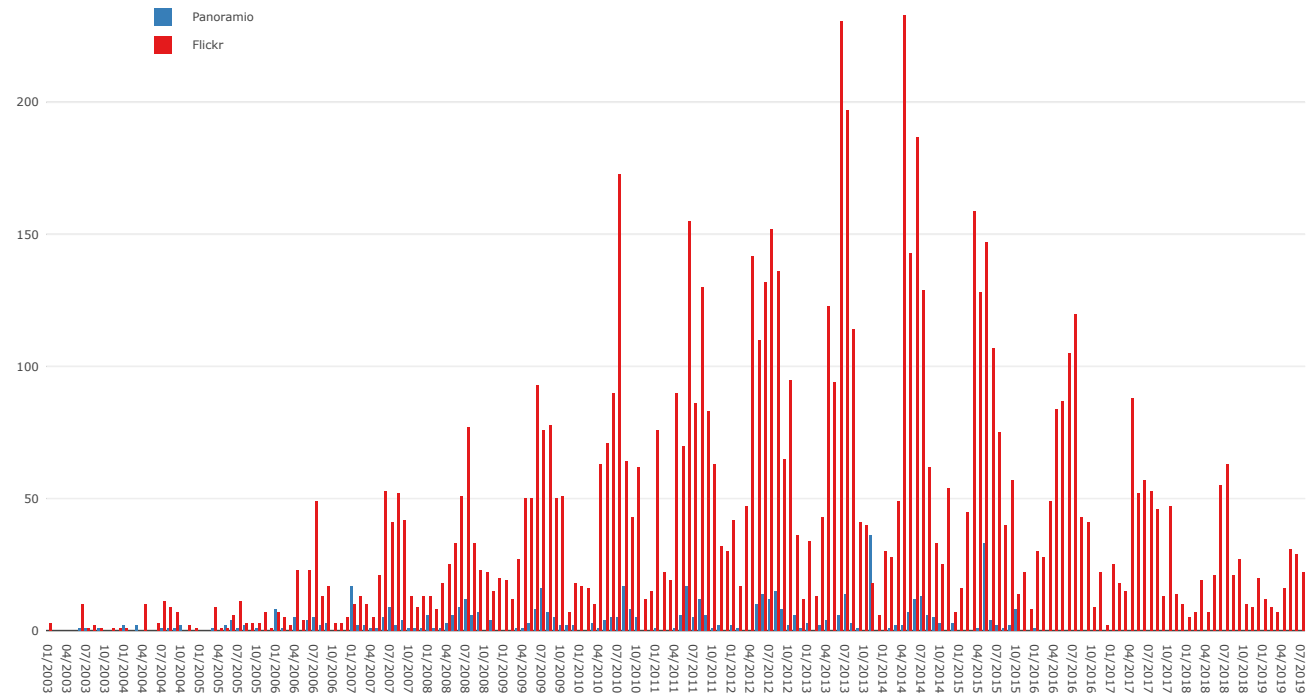


Figure 5: Date histogram by month and service, complete dataset, ignoring pictures from before 2003.

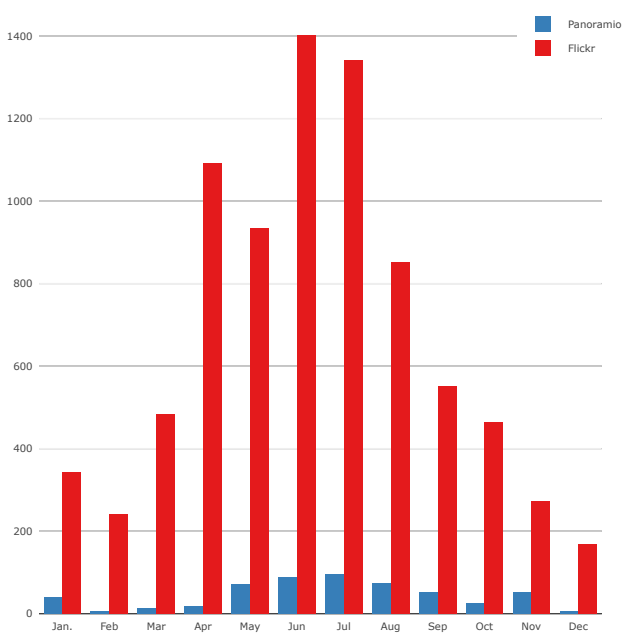


Figure 6: Aggregated date histogram, by month, complete dataset.

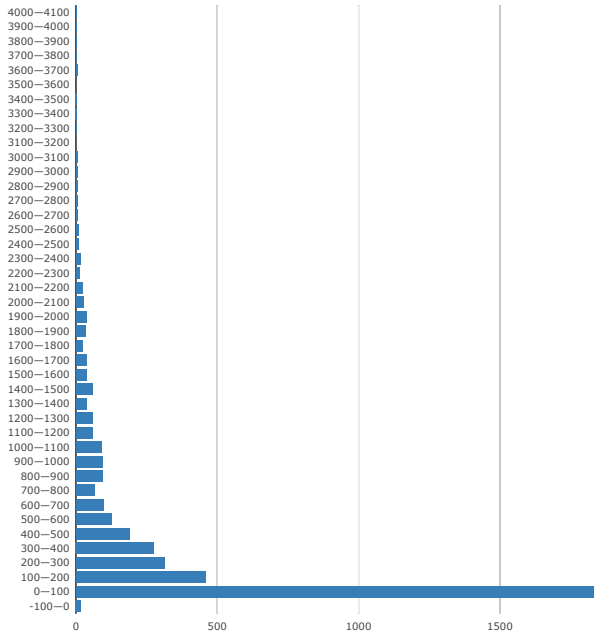


Figure 7: Elevation histogram, complete dataset.

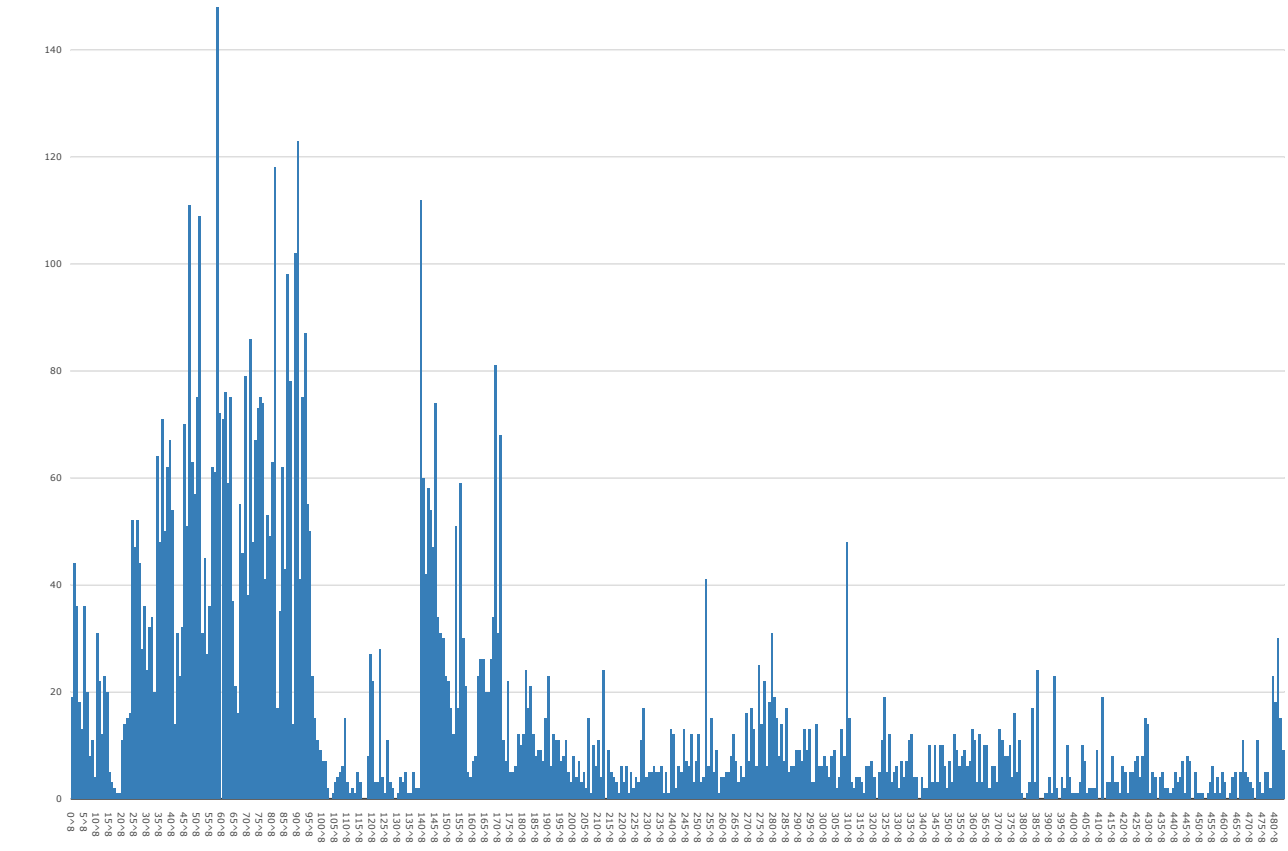
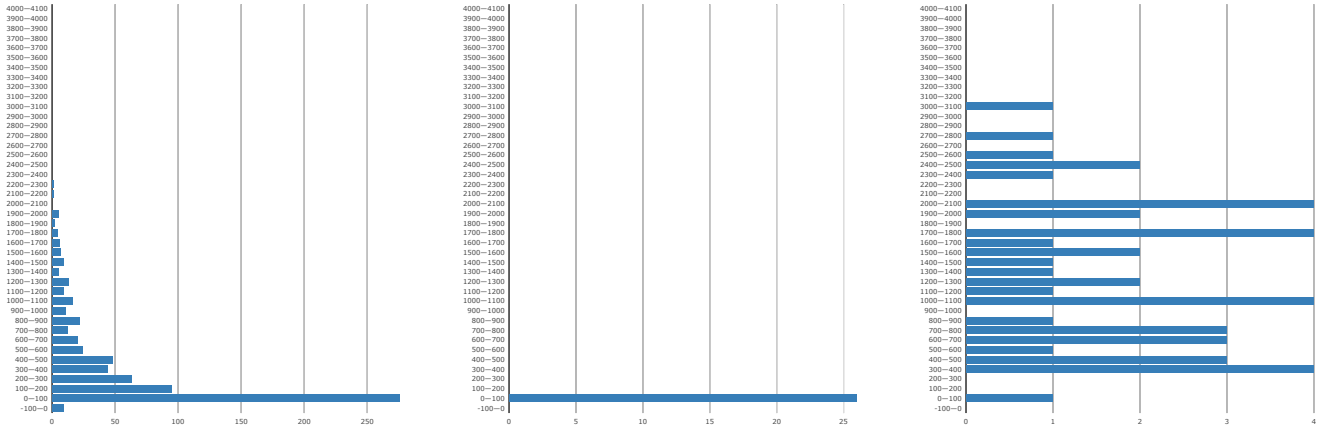
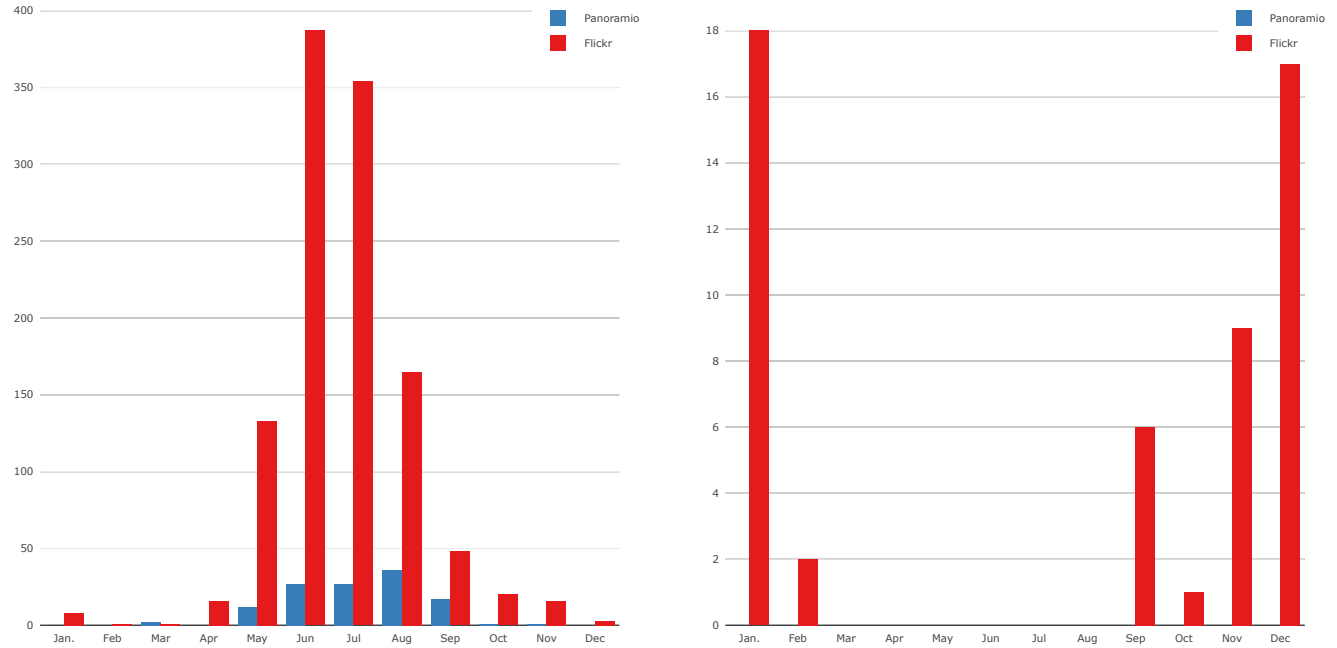


Figure 8: Histogram of Flickr ID bins for the identified plausible habitat images, complete dataset. Each bin contains 10^8 Flickr IDs. This plot shows some interesting patterns: We are not clear about the reason for the drop at around $120 \cdot 10^8$ and why fewer entries exist in the more recent bins since all bins are of equal size. These effects may be caused by the employed search strategies for the images using our set of keywords, or potentially it is caused by when we conducted the searches for potential picture IDs on Flickr (which we then cached for later manual inspection).



(a) (b) (c)

Figure 9: Comparison of three elevation histograms with different characteristics for (a) *Drosera rotundifolia*, (b) *Dionaea muscipula*, and (c) *Pinguicula aplina*.



(a) (b)

Figure 10: Comparison of the monthly sightings of two sundew species: (a) *Drosera rotundifolia* which occurs on the Northern Hemisphere and (b) *Drosera arcturi* from the Southern Hemisphere. The two plots clearly show the different growth periods in the Northern and Southern Hemispheres, respectively, but could also be influenced by when people travel to visit the respective habitats.

Figure 11: Our tool in *exploration mode*.

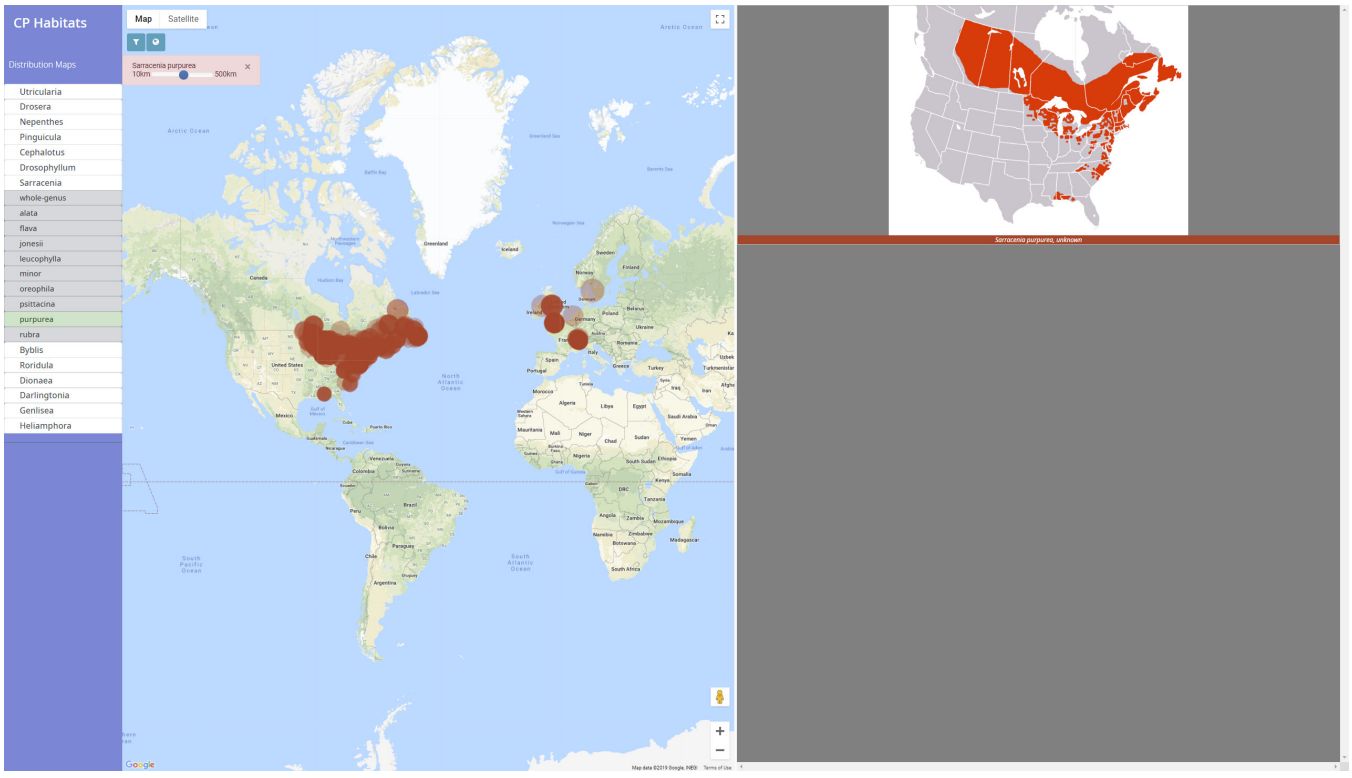


Figure 12: Our tool in *distribution mode*, showing both our distribution maps and maps from Wikipedia or other sources.

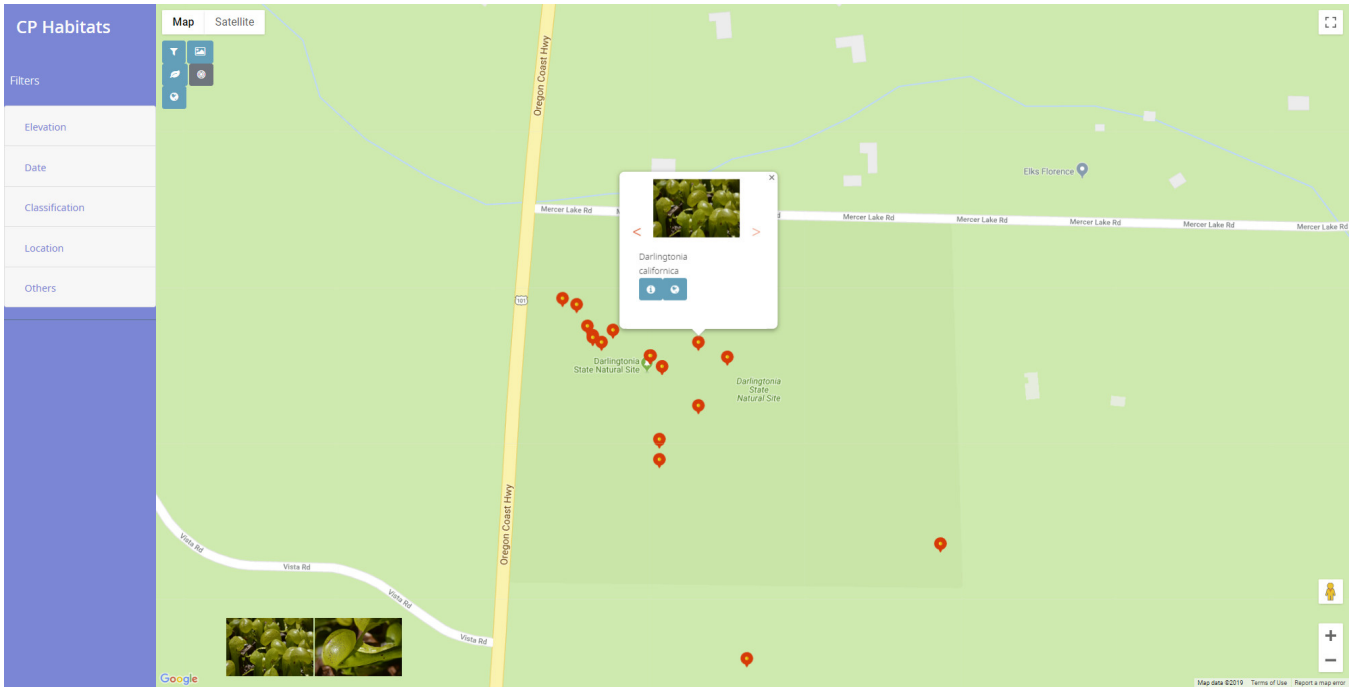


Figure 13: In *exploration mode*, narrowed in onto a particular site, with pictures of the site shown below.

Figure 14: In *exploration mode*, looking at a particular picture from the site shown in Fig. 13. The image is Flickr image 28504089402 by Jeremy Riel (CC BY-NC 2.0).

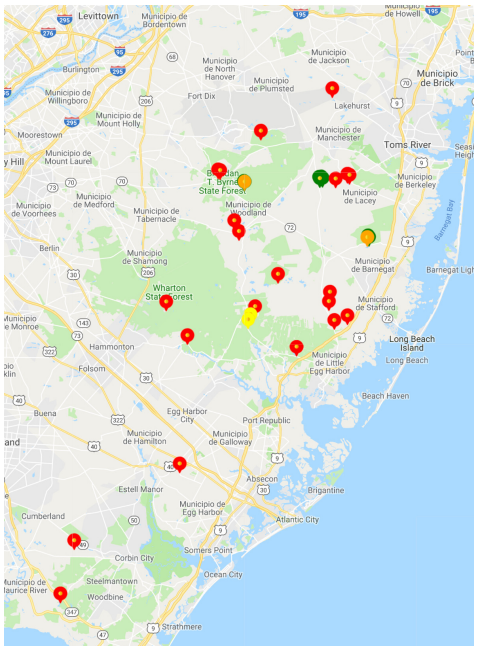


Figure 15: In *exploration mode* with the trust mapping enabled, a site in the United States that shows the different categories of verfication for a 1 km search radius (red: not verified by other users; orange: other plants by other users; yellow: other plants of same genus by other users; and green: other plants of same species by other users). Same map section as in Fig. 16.

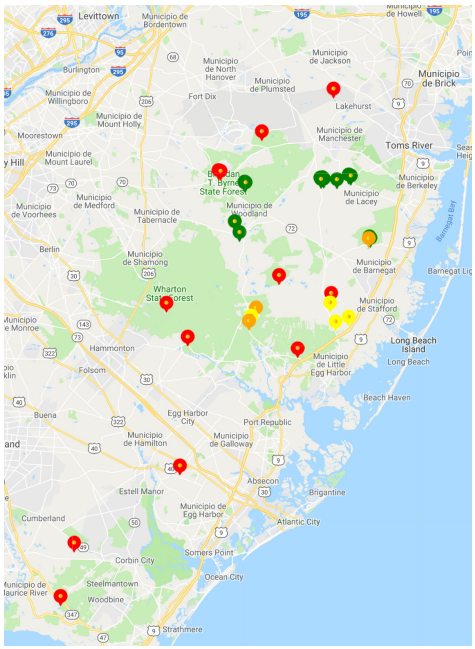


Figure 16: In *exploration mode* with the trust mapping enabled, a site in the United States that shows the different categories of verfication for a 4 km search radius (red: not verified by other users; orange: other plants by other users; yellow: other plants of same genus by other users; and green: other plants of same species by other users). Same map section as in Fig. 15.

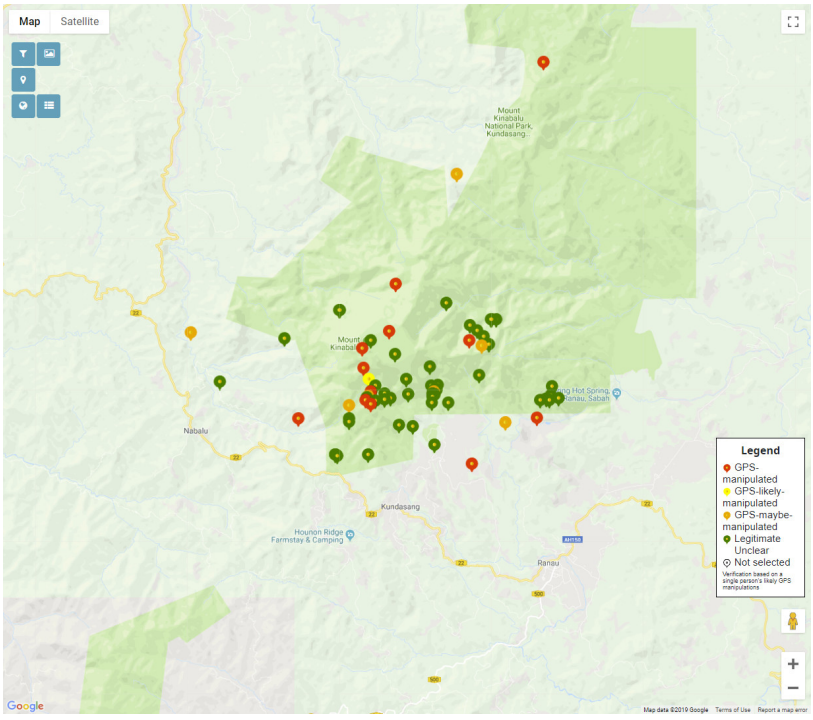
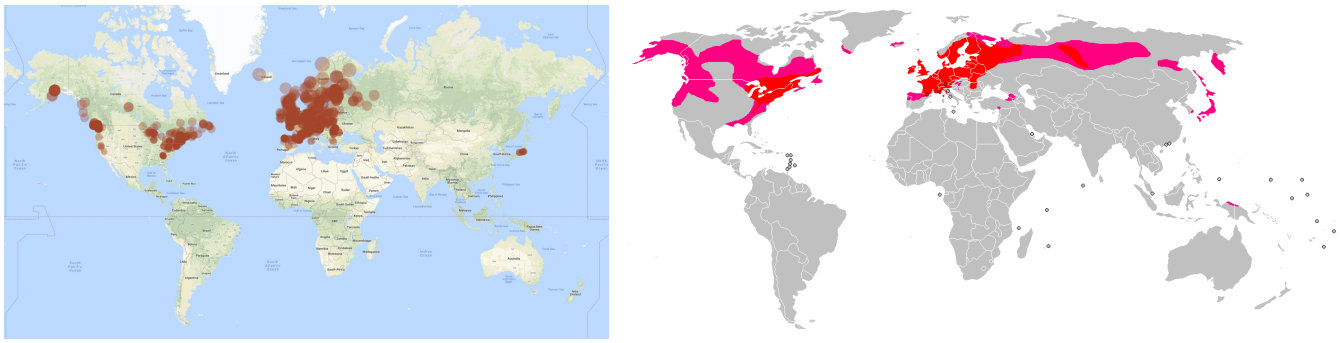
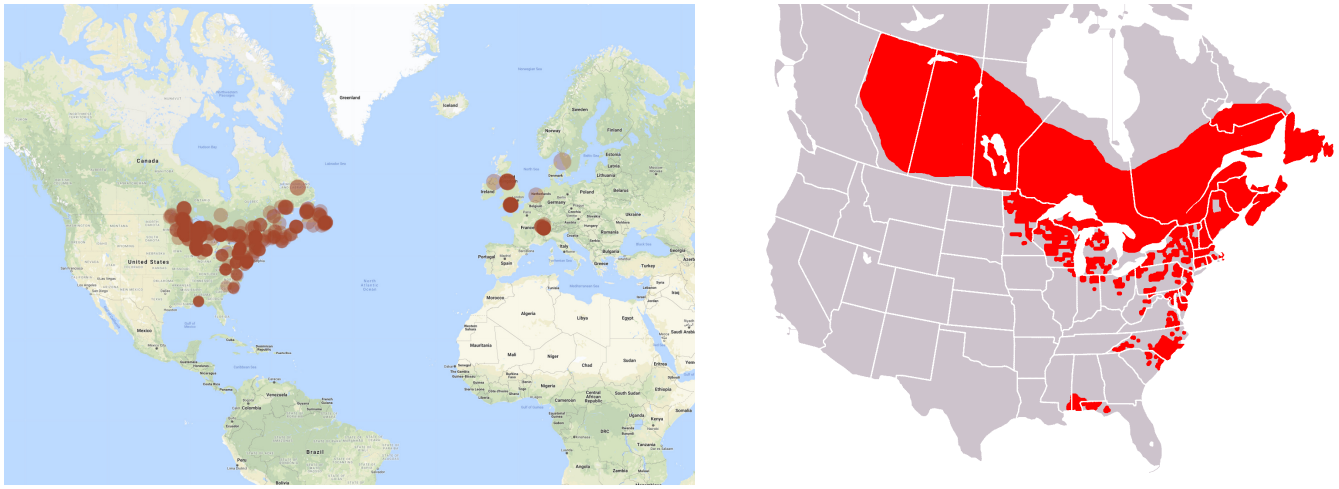


Figure 17: Verification based on the images of a single poster. If four or more images of the same poster are at exactly the same GPS coordinates, we assume the GPS coordinates to be manipulated (red). If three images are at the same location, then the coordinates are likely manipulated (orange). If two images have the same location, then the coordinates are possibly manipulated (yellow). Otherwise the coordinates are probably not manipulated or we cannot make a good judgment (green).



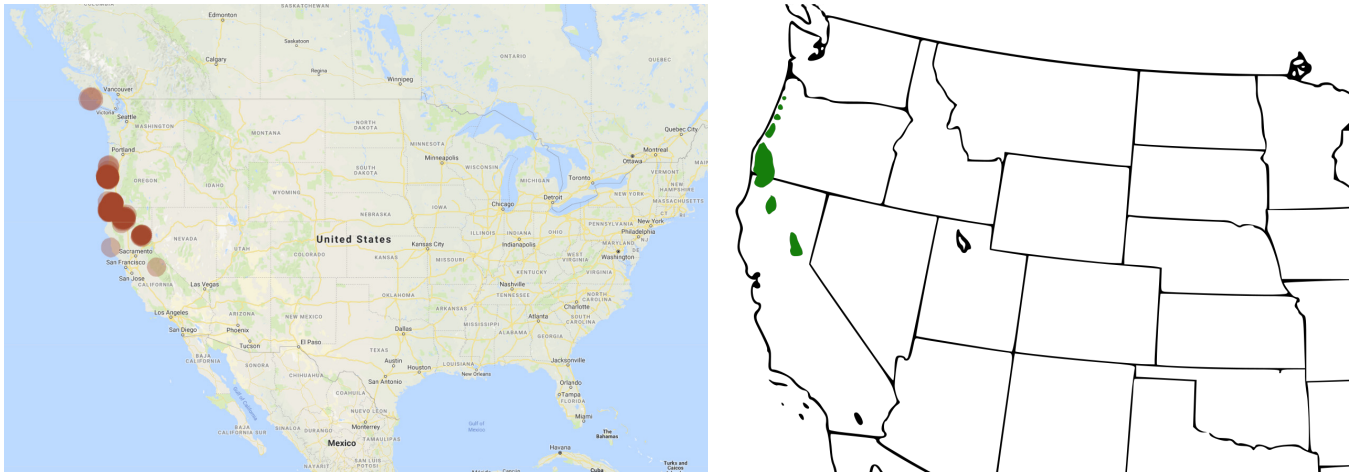
(a) (b)

Figure 18: Comparison of distribution maps for *Drosera rotundifolia*: (a) based on our social media data, (b) map from Wikipedia (Ⓢ).



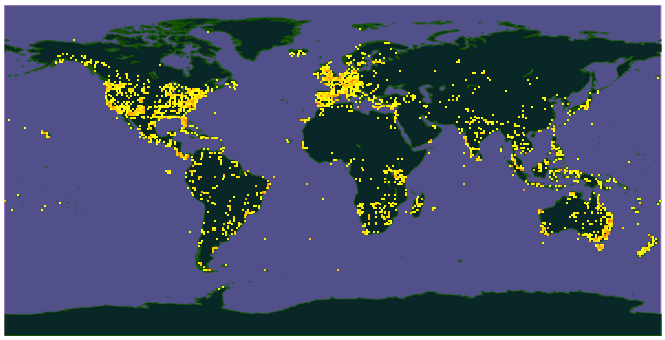
(a) (b)

Figure 19: Comparison of distribution maps for *Sarracenia purpurea*: (a) based on our social media data, (b) map from Wikipedia (Ⓢ).

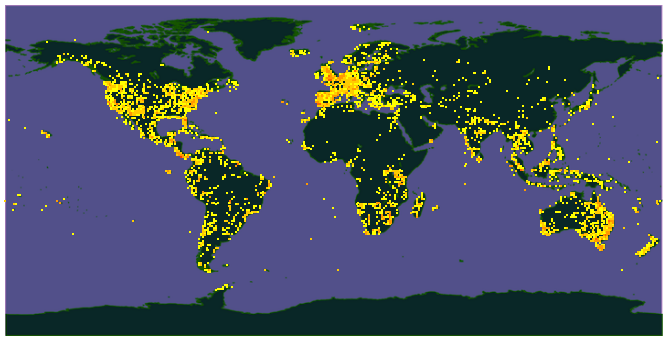


(a) (b)

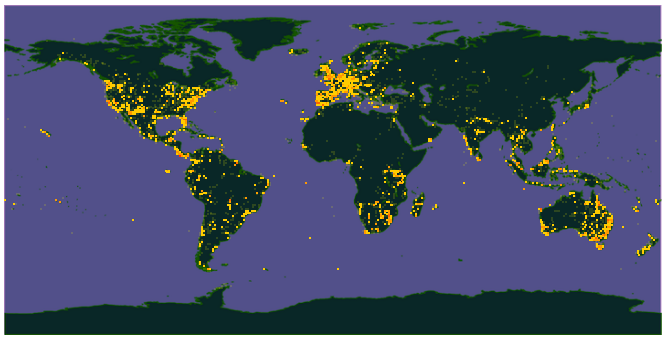
Figure 20: Comparison of distribution maps for *Darlingtonia californica*: (a) based on our social media data, (b) map from Wikipedia (Ⓢ).



(a) Heatmap of the geographic distribution of all 135,578 posts in the *Encyclopedia of Life* group by 2012, providing an estimate of the expected global likelihood for habitat image posts. Image by Roderic Page (CC BY 4.0).



(b) The same heatmap generated based on the currently available data of 349,011 posts up to July 2019, generated with the same script by Roderic Page. The overall distribution pattern did not change compared to 2012.



(c) Same image as in (b), only with the less-than-ten-images-per-region class (formerly yellow) shown transparently, to better emphasize the locations with multiple image posts.

Figure 21: Analysis of geographic distribution data based on the *Encyclopedia of Life* Flickr group which collects images and videos of animals, plants, fungi, protists, and bacteria: Since they cover a similar subject matter but for all species in general, the maps can be used as an indication of where it is more or less likely to get a good distribution coverage (see `https://iphylo.blogspot.com/2012/06/where-is-in-crowdsourcing-mapping-eol.html`). Color map: yellow: 1–9 images; light orange: 10–99 images; medium orange: 100–999 images; dark orange: 1,000–9,999 images; red: more than 10,000 images. Notice that in those regions that have no color assigned (i. e., the background map is visible) not a single image was posted to the *Encyclopedia of Life* Flickr group.

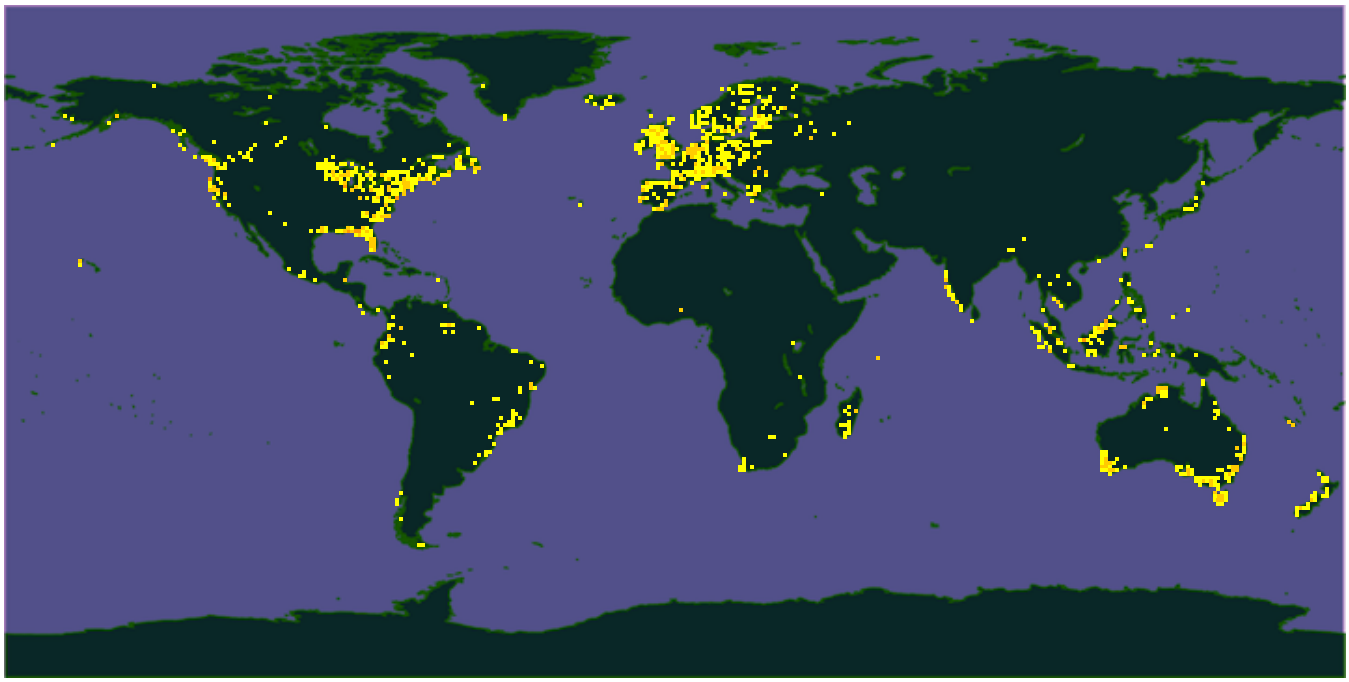


Figure 22: Heatmap for our data, generated with the same script by Roderic Page. Same color map as in Fig. 21.

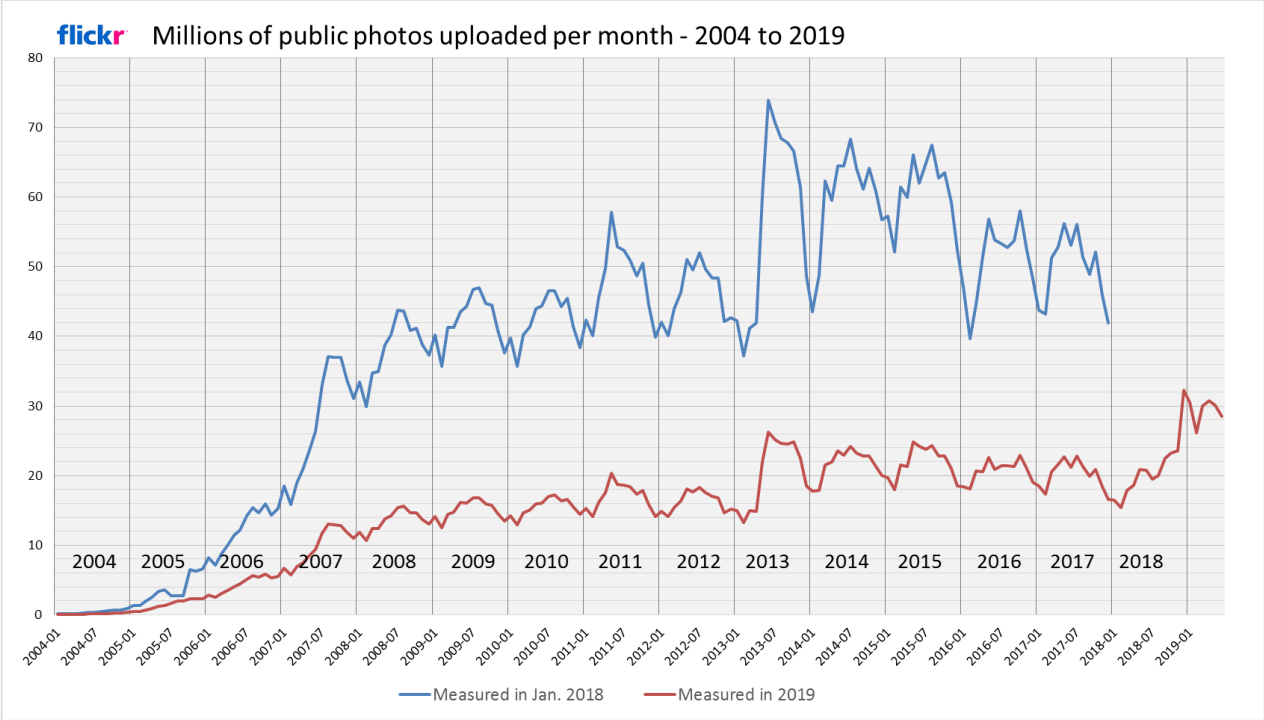


Figure 23: Flickr's changed upload policies from January 2019 apparently have let quite a number people to delete their images from the service. This fact may have let us to "loose" some images in the process (see Fig. 4) since we do not continuously scape the Flickr database but instead query it only in irregular intervals. Flickr image 6855169886 by Franck Michel (CC BY 2.0).

(a) (b) (c)

Figure 24: Examples for images which our search found but which we manually excluded because they do not show plausible plant habitats: (a) fake/artificial plants (Flickr image 9479368060 by Craig Moe; CC BY-NC 2.0), (b) botanical gardens or similar (Flickr image 5086658928 by Jane Nearing; CC BY-NC 2.0), and (c) plants kept at home (Flickr image 10753181585 by Mike Linksvayer; CC0 1.0). Other reasons for rejecting entries include name collisions with geographic place names, GPS locations at unsuitable locations (e. g., farmland, streets), and images that do not show any or not the right plants (e. g., name collisions with common or scientific names, general textual habitat descriptions in the comments of the postings).

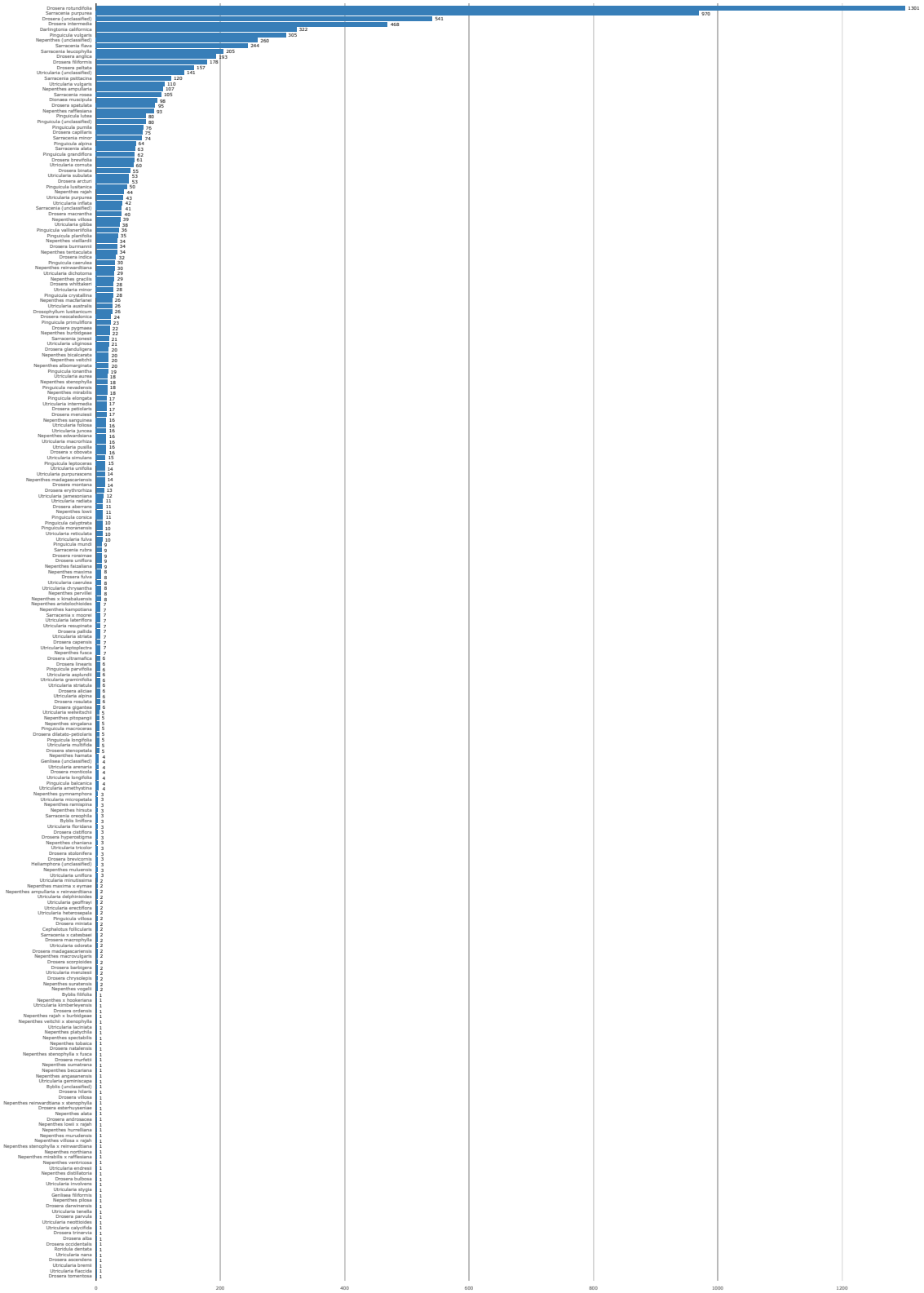


Figure 25: Species count in complete database of all species that we found, sorted by rank.

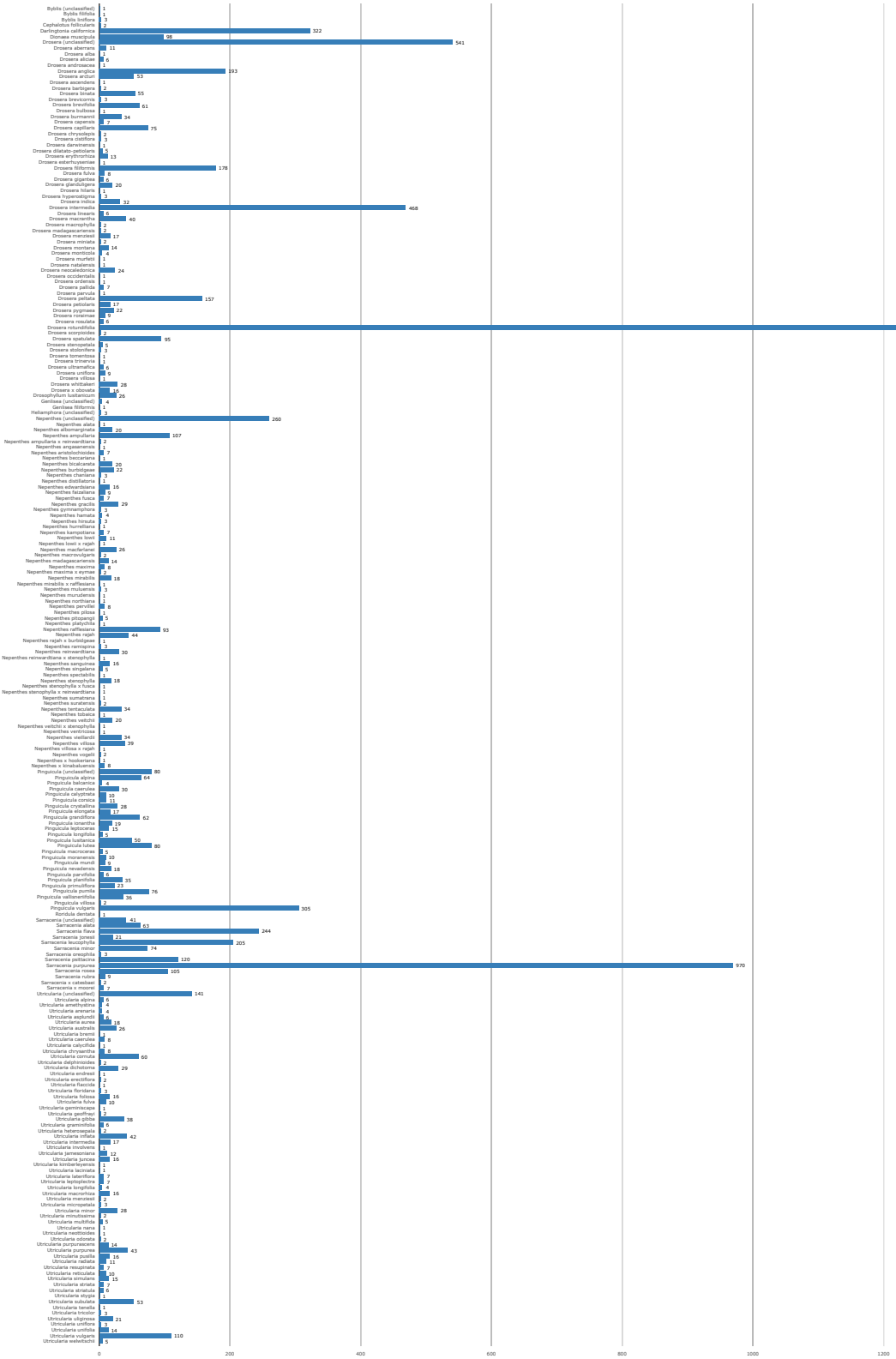


Figure 26: Species count in complete database of all species that we found, sorted by species name.