Small animals, big data: Mobilizing citizen science observations to establish the largest database of spiders in Asia

Programme:BIFA
Project ID: BIFA6_029
Project lead organization: Strand Life Sciences
Project implementation period:1/9/2021 - 31/8/2022
Report approved: 16/3/2022

Narrative Midterm report

Executive Summary

We began the project by building a team that would collaborate across and within our respective institutions. The project's goals and criteria were set, and a developer was assigned to undertake a background study, test multiple implementations, and choose appropriate algorithms and software. The Facebook APIs and other approaches were researched to access the required data. Several methods for extracting scientific names, place names, and dates, as well as geocoding software solutions, were put to the test. Sample data was prepared and shared with team members before the project was completed. Online sessions between partners were scheduled as needed to clarify and get feedback.
We organized a workshop for members of the SpiderIndia community in Auroville, Tamilnadu, between the 17th and the 20th of December 2021, which was attended by all participants of this project as well as 15 members of the SpiderIndia community. Presentations on the project's goals and objectives, as well as the GBIF, were given at this meeting. This time was spent by the team evaluating the project's progress and fine-tuning plans for future implementation. This report is accompanied by a workshop report.

Progress against milestones

Has your project published at least one dataset through GBIF.org?: No

Dataset published:

Dataset	DOI

Has at least one member of your project team received certification following the BIFA capacity enhancement workshop?: No

Rationale: Two of our team members attended the course and completed the evaluation exercises. We are yet to receive an update on the certification.

Report on Activities

Activity progress summary

We have put together a team to work on the project. The development team at Strand consists of a developer who will be led by the Project lead -Thomas. The curation team will consist of two curators, who will be guided by the other project partners. The development team has assessed a number of implementations and solutions in order to meet the project's goal of extracting and parsing postings from social media in order to extract scientific names, dates, and placenames, as well as geocode

them. In GitHub issues, all essential development activities were documented and scoped out for implementation. We began by looking into Facebook APIs in order to gain access to SpiderIndia's Facebook group posts. However, due to Facebook policy constraints, previous postings are now unavailable via APIs, and new posts necessitate the building of a custom app, which each member of the Facebook group must grant access to. To get over this significant limitation, we opted to forego the APIs and extract the Facebook post text straight from the browser. We were able to extract all 20,000 posts. We used several natural language processing and named entity recognition techniques to extract the items of interest after obtaining the post text. Some of the solutions tried and evaluated for location extraction included spaCy, FlashText, and locationtagger. The GNRD scientific name extraction tool was used to extract scientific names. For date recognition and extraction, several Python modules were used. We've set up and tested a localized Pelias installation, populating it with data from OpenStreetMap, Who's on First, OpenAddresses, Geonames, and custom imported data. This installation will be used to recognize placenames and as a source for geocoding them. The preliminary data was then shared with the curation team, and their feedback was taken into account. The chosen algorithms are presently being turned into APIs within a microservice and linked up to a user interface for the curation team to use.

We successfully organized a workshop for members of the SpiderIndia community in December. The event took place at Auroville from December 17th to December 20th, 2021, and around 15 people from all over India attended. Identification, field investigations, and current research were among the subjects covered in the session. Members were given presentations on both GBIF and the goals of the current endeavor to promote awareness of both subjects. The curators for this project were picked from a pool of candidates. All members of the project team met on the margins of the meeting at the venue to discuss project development, curation strategy, and future implementation. A summary summarising the meeting is included with this report.

Completed activities

Activity name: Project team setup, project management infrastructure, scoping of frameworks, software and background research.

Description: The first month will be utilized to set up the team, conduct necessary background research and scoping, finalize software and development setup, formalize project management infrastructure and acquaint the partners involved.
Start Date - End Date: 1/9/2021 - 31/10/2021
Verification Sources: The github account is is restricted through user login, but a screenshot of it is attached.

Report on Deliverables

Deliverables progress summary

We had to extract the textual content, as well as permalinks and other metadata, from the SpiderIndia group's Facebook postings before we could generate the target collection of spider occurrence data. We considered using Facebook APIs to accomplish this but came into some difficulties. The Facebook APIs and authorization framework are highly restricted, preventing such content from being extracted freely. We looked at other options, such as Python libraries, and were able to extract all of the text data as CSV files that contained all of the necessary information. The software team then ran this raw data through a series of pipelines and algorithms, fine-tuning the parsing technique for scientific names, dates, and placenames.
A subset of the data was supplied to the curation team so that they could review it and provide feedback. This subset has been curated as sample data by the curation team, and it is now available for publication. Before release, this data was examined on a GBIF test IPT to finalize the structure and make any necessary adjustments. As stated in the deliverables, the curating process will begin soon, and it will be available for publication in June.
We successfully hosted the first workshop for members of the SpiderIndia community. The event took place in Auroville from the 17th to the 20th of December 2021

Progress towards deliverables

Title: A dataset of spider occurrence records curated from the SpiderIndia Facebook community

Type: Dataset
Status update: The raw data has been extracted from Facebook. A subset has been tested through

the preliminary pipeline to generate sample data that has undergone scrutiny by the curation team. This has also been verified for publication on a demo IPT installation. https://www.gbif.org/tools/data-validator/87131010-436b-44a5-b938-c0ab25c9afb5. The larger data is now due for curation as soon as the pipeline is made available and will be published by the due dates stated via the BIFA IPT.
Dataset scope: Spider occurrences
Expected number of records: 15000
Data holder: SpiderIndia
Data host institution: Strand Life Sciences
Sampling method: Raw data extracted as text from Facebook posts, put through scientific name, date and location parsers, curated through curator intervention, validated by a taxonomic expert , geocoded and published.
% complete: 40
DOI:
Expected date of publication: 2022-07-31

Title: Two workshops for members of the SpiderIndia community and other group administrators on mobilizing data and showcasing the interactive workflow for extracting occurrence data from Facebook groups

Type: Other
Description: Two workshops for members of the SpiderIndia community and similar groups on Facebook. The workshops will focus on sensitizing users on the need for aggregating data on spiders and encourage contribution.
Sources of verification: Report on the meeting is attached

Title: A generalized replicable, online, interactive workflow for extracting occurrence data from Facebook groups.

Type: Other
Description: A workflow to integrate algorithms to detect scientific names, place names and dates and preliminary curation by designated data curators on the website will be enabled.
Sources of verification: We tested out natural language processing and NER algorithms to extract Scientific names, place names, and dates. spaCy, FlashText, locationtagger, GNRD scientific name extraction and several Python libraries were tested. A localized Pelias installation was populated with data from multiple data stores for placename recognition and geocoding. The extracted data was shared with the curation team for feedback. APIs and are currently being wired up to UI for testing and is expected to be completed by March-end. Screenshots of the user Interface developed so far and the Github links to code are attached.

Events

First workshop for members of the SpiderIndia community

Dates: 2021-12-17 - 2021-12-19
Organizing institution: Naturemates
Country: India
Number of participants: 15
Comments: In December, we successfully hosted the first workshop for members of the SpiderIndia community. The event took place in Auroville from the 17th to the 20th of December 2021, and it drew over 15 people from all around India. The workshop covered a variety of topics related to spiders, such as identification, field investigations, and current research. To raise awareness of these themes, members were given presentations on both GBIF and the goals of the current initiative. This project's curators were chosen from a pool of candidates. On the sidelines of the meeting, all members of the project team met at the venue to discuss project development, curation strategy, and future implementation. This report is accompanied by a report summarising the meeting. The second workshop is scheduled for April later this year.
Website or sources of verification: Report attached

Communications and visibility

From the 17th to the 19thth of December 2021, we organized the inaugural workshop for members of the SpiderIndia community. Vijay Barve, a GBIF data ambassador, gave a presentation on GBIF to inform participants on the organization's role as a global biodiversity repository and how it may be used to store biodiversity data. Thomas Vattakaven described the current project and intentions for data extraction, as well as how members may help structure their Facebook data to improve data quality.

Monitoring and evaluation

Monitoring and evaluation findings

Based on our current assessment of the project's development, we believe it is on track to meet the project's timetables and deliverables. All project development activities are documented and tracked in Github. Daily meetings between the developer and the PI are held to keep track of progress, and the scoping phase, as well as preliminary testing, have been accomplished. The construction of the user interface and the wiring up of backend programs are presently underway. The information we gleaned from the Facebook group isn't as extensive as we had hoped, with only roughly 20,000 posts in all. Only around half of the records will satisfy publication standards after curation. We will, however, continue to add records that have come in after the data was first collected.
The curation team meets online every week to discuss issues and feedback. The other team members have regular online meetings as and when there are updates or issues to clarify. The first workshop of community members was successfully organized and plans for the project were presented to them for feedback.
We have also decided to publish the data through the BIFA IPT, as this will be the most straightforward and avoid data duplication and clash with IBP’s regular occurrence dataset.

Impact of COVID-19 pandemic on project implementation

COVID positive tests have been found in some project team members and close family members at various phases. One of our developers has recently been affected, However, the symptoms were minor, and only minor disruption was caused due to it. While organizing workshops, we had to consider the number of attendees, as well as travel and lodging constraints imposed by the epidemic and current government standards. In the current scenario, we anticipate sticking to the established timeframes and completing the project on time.

GBIF leads the Biodiversity Information Fund for Asia (BIFA), a programme funded by the Ministry of the Environment, Government of Japan. The programme provides supplementary support for activities addressing the needs of regional researchers and policymakers through mobilization and use of biodiversity data.