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 - 31/10/2022
Report approved: 7/12/2022

Narrative Final report

Executive Summary

Our project proposed to update the current status of spider taxonomy on IBP and establish a systematic workflow to enrich occurrence data on spiders by leveraging the popularity of existing social media networks such as SpiderIndia. We wanted to make a semi-automated, repeatable workflow for curation of social media text posts into IBP occurrence records, validation by taxonomy experts, and data publication to GBIF. We were successful in implementing our objectives and achieving the goals of the project. We have developed a curation interface that allows the uploading of tabular data containing columns with text extracted from social media posts. This text is put through a pipeline that recognizes scientific names of spiders, place names within India, and the date of observation if present. These entities are presented in a user-friendly interface for a team of curators to examine and curate.The curation effort leads to the creation of primary biodiversity data occurrence records. The interface is replicable and reusable for similar data extraction exercises. Our curation team has successfully curated over 20,000 such records, which have also been examined and validated by a subject expert. About 70% of these records had enough information for them to be published to GBIF as Darwin Core objects. We also held two workshops, one at the beginning of our project and one near the end, to show stakeholders the workflow and explain both our goals and the need for clean baseline biodiversity data.

Progress against milestones

Has your project completed all planned activities?: Yes

Has your project produced all deliverables?: Yes

Report on Activities

Activity implementation summary

We have been able to implement all of the activities that were originally proposed. We started by scoping out the tasks and necessary components that would be needed to be wired together to extract text posts from Facebook and convert them into curated primary biodiversity records. The most challenging part of this was in extracting the required text content from Facebook. We hit numerous blocks in extracting the required content with all required values, resulting in unexpected setbacks initially. However, we were able to work past these, and find solutions to source the necessary data. Code libraries or existing open source software that were suitable were scoped out and tested individually prior to integration into a combined workflow. Some of these had to be individually modified and customised to work cohesively with our chosen backend infrastructure. Once the necessary code and software was wired in to create the backend pipeline, a user-friendly interface was designed and implemented. User permissions, logging and downloads were subsequently integrated in a phased

manner. It is now a full fledged curation interface that can be reused for similar text curation.

The curation exercise went on without much obstacles. Curators were presented with scientific names, place names and dates extracted from the text. They were tasked with selecting and verifying these from the available values or adding them from hints in the text, if the pipeline was unable to detect it. Our curators have been through over 22,000 records and curated these. A total of little over 68% of these records (15055) were marked as curated and complete, having valid scientific name, place name and date. Other records which lacked sufficient details were marked as rejected. Although each post had enough textual content for curators to work with, the interface is unable to display a preview of the original post due to Facebook's policies and blocking of cross-origin requests, resulting in curators needing to click and visit the site to access further details. Future changes to the interface could include integrating it more closely with Facebook SDKs to make it easier for users to log in and allowing Facebook users to edit their own posts on IBP or add missing information. But Facebook's rules about privacy and the fact that it constantly tries to block access to content outside of its own site make it pointless to integrate with this. A taxonomy update of known spider species names from India was carried out prior to implementing the pipeline, and taxonomic validation was enabled and carried out for all valid records.
Two workshops were held, one at the early stages of the project and one at a later stage. The initial workshop was for members of the SpiderIndia community, where they were sensitised to the objectives of data aggregation and encouraged to contribute. The workflow and pipeline for curation of textual content from Facebook, its practicality, utility, and benefits were explained to these stakeholders in both workshops and were received with much appreciation. The workshops also enhanced the capacity of members to contribute data on Facebook in more standard formats, improved their understanding of informatics, and sensitised them towards open data. In addition, participants were provided with an overview of GBIF and its role in aggregating and serving open primary biodiversity data. Workshop participants in the final workshop were also guided through the process of compiling the biodiversity around the event locality over the two days of the workshop. These records were uploaded and published to GBIF.

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: 3/10/2022 - 13/10/2022
Verification Sources: We use the Github issues feature to list out and track development ideas and progress through the Biodiv codebase on Github. Access is restricted to logged in users, but attached is a screenshot of the issues page.

Activity name: Development of an online workflow to extract and publish primary occurrence records from Facebook as secondary observations on IBP

Description: A developer working with the India Biodiversity Portal will collaborate with the Admins of the SpiderIndia Facebook group to develop a workflow for extracting Facebook posts on spiders into a tabular format on the portal for further curation. As the incoming data is unstructured, this workflow will integrate algorithms to detect scientific names, place names and dates and structure it. These algorithms will be developed and integrated into the workflow. The facility for preliminary curation by designated data curators on the website will be enabled. The ability for such data to be uploaded as a datatable on IBP, with each row being a "Secondary Observation" linking to the original permalink will be added.
Start Date - End Date: 1/11/2021 - 31/3/2022
Verification Sources: We have designed and developed a reusable pipeline for data extraction, curation, validation,validation and publication as Darwin Core record and made it available via a user interface. The listing of curated datasets are available at: https://indiabiodiversity.org/text-curation/list .Each dataset is clickable to view the records. Editing for curation purposes is available to a set of curators added to the metadata. Validators are also able to validate the records.The dataset is available to download and publish for the above individuals. The Upload interface for creating new datasets are available to a logged in user at https://indiabiodiversity.org/text-curation/create

Activity name: Curation of preliminary parsed data by curators and publishing of datatable

Description: Curators will be engaged in examining the parsed data resulting from the developed

workflow and manually curating what is necessary. The algorithms may not always be accurate or capable of handling all cases and will almost certainly require manual intervention and correction. Once cleaned up the data will be uploaded as a datatable on IBP. These tasks will be done by research fellows contracted for the purpose.
Start Date - End Date: 1/4/2022 - 3/6/2022
Verification Sources: Over 22,000 records were extracted from the SpiderIndia Facebook group and made available for curation. A team of our curators have been working to curate this data and have completed this task. The progress of curation on each dataset is indicated by a progress bar on the list page: https://indiabiodiversity.org/text-curation/list

Activity name: Taxonomic validation of preliminary parsed data by curators and publishing to GBIF.

Description: Once the data is published as secondary observations on IBP, it will undergo validation by curators with taxonomic expertise. All qualifying records that meet the data quality criteria will be marked as validated. This data will be pushed for publication to GBIF as a part of the India Biodiversity Portal publication grade workflow.
Start Date - End Date: 1/5/2022 - 30/6/2022
Verification Sources: Our partner, Siddharth Kulkarni, who is a taxonomic expert on spiders has gone through curated records that qualified as a complete record to verify and validate it. On clicking each record, each validated record includes a VerifiedBy field. This can also be seen as a column in the downloaded file. https://indiabiodiversity.org/text-curation/list

Activity name: Workshops on data mobilization, exposing and demonstrating the developed workflow to stakeholders.

Description: Two workshops will be organized during the duration of the project. The initial workshop will be for members of the SpiderIndia community where they will be sensitized on the objectives of data aggregation and encouraged to contribute. After the workflow is developed a workshop will be organized where the top contributors of the SpiderIndia community and Admins of large biodiversity groups on Facebook will be invited. The workflow, its practicality, utility and benefits will be explained to these stakeholders. It will improve the capability of members to contribute data on facebook that can be further extracted through this workflow with less manual curation. It will enhance the capacity of members to contribute data on facebook in more standardised formats, improve their understanding of informatics and sensitize them towards open data. Efforts will be made to convince other group admins to adopt the workflow to liberate and push their data for publication. This will help further mobilize data into accessible formats.
Start Date - End Date: 1/12/2021 - 30/6/2022
Verification Sources: We have successfully organised two workshops with a community of stakeholders to sensitise them towards data curation and data aggregation of primary biodiversity data on Spiders in specific and biodiversity in general. The curation interface with the curation pipeline was demonstrated to these users. A report of the two workshops are attached.

Report on Deliverables

Production of Deliverables - Summary

We have successfully completed all deliverables that we committed to in the project proposal. We extracted over 22,000 Facebook posts from the SpiderIndia Facebook group. These were uploaded to the curation interface on the India Biodiversity Portal, where the scientific names, dates, and place names were extracted from the text and presented to curators to select and verify. Place names were also geocoded through the process. The interface allows open access to view the data records, and edits are restricted to designated curators and validators. A team of curators working on the project went through each record and curated the entry by cross-verifying with the original post. The curated records were then validated by a taxonomic expert. Each action is logged as it is done and is visible on the interface. The interface allows the curated records to be downloaded as a CSV file with Darwin Core headers. When curation and verification were done, the data were sent to GBIF through the BIFA-IPT. The interface has made the curation process easy and effective, resulting in a quick conversion of textual data into primary biodiversity data records of publication quality. It is reusable for other situations that require the extraction of primary biodiversity information from text. We have been in touch with other groups that are interested in the extraction of such data and are hopeful of enabling data curation to curate more data in the near future.

Two workshops were held, one at the early stages of the project and one at a later stage. The initial workshop was for members of the SpiderIndia community, where they were sensitized to the objectives of data aggregation and encouraged to contribute. The workflow and pipeline for curation of textual content from Facebook, its practicality, utility, and benefits were explained to these stakeholders in both workshops and were received with much appreciation. The workshops also enhanced the capacity of members to contribute data on Facebook in more standard formats, improved their

understanding of informatics, and sensitized them towards open data. In addition, participants were provided with an overview of GBIF and its role in aggregating and serving open primary biodiversity data. Workshop participants in the final workshop were also guided through the process of compiling the biodiversity around the event locality over the two days of the workshop. These records were uploaded and published to GBIF.

Production of deliverables

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

Type: Dataset

Status update: Over 22,000 observations from the SpiderIndia Facebook group were extracted into tabular data and uploaded through the pipeline that we developed to parse and extract scientific names, dates, place names with latitude and longitude coordinates, and a permalink to the original post. This data was put out on a curation interface on the India Biodiversity Portal with user permissions for a set of curators to work on and curate the data. The data was manually curated over several months to generate a dataset with valid records. Incomplete records were marked as rejected. All records were verified by a taxonomic expert to qualify as research-grade data. The curated data was downloaded as a CSV file with Darwin Core headers and published to GBIF via the GBIF IPT.
Dataset scope: Spider occurrence data from India between 1982 -2022}
Expected number of records: 15055
Data holder: India Biodiversity Portal
Data host institution: India Biodiversity Portal
Sampling method: Opportunistic occurrence data were uploaded by users on the SpiderIndia Facebook group, with or without identification of the organism. The date and location of sighting were input as strings within the descriptions of each post. Each post also had one or more images of the spider. Identification of the uploaded organism was done either by the uploader or by members of the SpiderIndia community. Either in the post's description or in the comments, the ID was posted as a string. For each record, the description and comment text, as well as a link to the original post, were extracted. This was uploaded in CSV format to a pipeline on the India Biodiversity Portal, which used the GNRD or Flashtext algorithm to extract scientific names and locations and a Python date library for dates. The Pelias search engine was used to fetch place names within India, and these were used as entities for recognition within the text. Pelias was then used to geocode the locations and add latitude, longitude values. Scientific names, places, and dates were verified manually by curators, and the scientific name was validated by an expert.
% complete: 100
DOI: 10.15468/7jz829
Expected date of publication:

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

Type: Other
Description: We have designed and developed a reusable pipeline for data extraction, curation, validation, and publication as a Darwin Core record and made it available via a user interface. A group of curators can make changes for curatorial purposes. Validators are also able to validate the records.
Sources of verification: https://indiabiodiversity.org/text-curation/list

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 were held, the first at Auroville, Pondicherry for members of the SpiderIndia community, and the second for a larger community of stakeholders across taxa, at Sunderbans, west Bengal.
Sources of verification: Reports from the two workshops are attached. A dataset of the flora and fauna observed during the second meet is published as a dataset at: https://www.gbif.org/dataset/a255c20a-10de-4bb7-8ba8-10ba413a9845

Impact of COVID-19 pandemic on project implementation

Many of our team members have been afflicted by COVID-19 during the project implementation. These

have led to temporary setbacks in timelines but we have been able to recover from these and continue with the project. There have been no significant delays or impacts due to COVID-19 .

Events

1st stakeholder meeting of SpiderIndia participants

Dates: 2021-12-17 - 2021-12-20
Organizing institution: Nature Mates Nature Club
Country: India
Number of participants: 15
Comments: We successfully organized a workshop for members of the SpiderIndia community in December 2021. 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. On the sidelines of the meeting, the whole project team got together at the venue to talk about project development, the curation strategy, and how it will be used in the future. A summary summarizing the meeting is included with this report.
Website or sources of verification: Attached is a report of the meeting.

Events

Final Stakeholder Meeting and First DiversityIndia Meet (2022), Sundarbans, West Bengal, India

Dates: 2022-04-16 - 2022-04-19
Organizing institution: Nature Mates Nature Club
Country: India
Number of participants: 33
Comments: The first Diversity India Meet 2022, (16th-19th April) was conducted in the mangroves of the Sundarban Tiger Reserve. The purpose of this meet was to inculcate an understanding of biodiversity data mobilization and to document the biodiversity of Sundarbans by involving students and experts on various topics related to biodiversity documentation. There were over 35 participants, all of whom came from different places and had different interests. During the meet, the participants documented both the flora and fauna of the area. In all the regions we have visited, 114 different types of plants have been spotted during the meet. In fauna, students have documented animals from both the vertebrate and invertebrate sections. Among invertebrates, they have seen many arthropods, molluscs, and among vertebrates, they have seen many birds, reptiles, and mammals. The compiled biodiversity of the region was published as a dataset to GBIF.
Website or sources of verification: Attached is a report of the meeting

Communications and visibility

During the project implementation phase, we conducted two meetings, the first in Auroville, Pondicherry, and the second in Sunderbans, West Bengal, where invited members of stakeholder communities were invited. The ongoing project and the pipeline that we built were presented and demonstrated at these meetings, and feedback was sought. The reports of these meetings are attached here. We have also initiated a paper to be written up and published in a reputed biodiversity informatics journal. When completed, this will give the project widespread attention and publicity. We have also been attempting to reach out to like-minded groups and organizations that could benefit from utilizing the workflow and pipeline, so that it can be reused in other data mobilization exercises. We are positive in extending the use case and publishing further data.

Monitoring and evaluation

Final Evaluation

Our project was conceived with the intention of creating a reusable pipeline to curate pre-existing unstructured biodiversity data from social media posts, utilizing the case of SpiderIndia as a test scenario to produce a large dataset of curated primary biodiversity data for publication. The goal of the project was to make a pipeline that could be used again and again and would be useful beyond the

scope of the project. This would lead to more data curation activities that would produce similar data in other situations. The concept was largely theoretical when it was first put forth, and we were unaware of any other initiatives of a similar nature. We were able to create a workflow for data curation on the India Biodiversity Portal using a variety of software elements and algorithms that can be applied repeatedly. However, our chosen data source proved to be problematic due to Facebook's restrictions on data-sharing and access. We were able to work around this to extract the desired data, but have realized that future attempts to extract data from Facebook may be challenging and difficult. Our interface was able to load, parse , extract, and display the original text from the data source, along with links to the original source, and detect scientific names, place names, and dates. Our curator team then worked within this interface to examine the parsed data alongside the original text and curate it by verifying the detected entities and marking records as either complete and curated or rejected. The overall experience of the curators was positive. However, viewing the original data required clicking through and visiting the Facebook post in order to verify because Facebook blocked a preview of the data at source. This resulted in some delays. Less than 70% of the 22,000 records had all of the species, location, and date information that was needed for curation. This meant that the remaining records were missing information like the place and date of the observation or the name of the organism. A large number of these could have been used if members were made aware of the need to mention place and time details while asking for help in identification on the Facebook post. As a result, we spent a lot of time and effort in the two meetings we organized with stakeholder communities, raising awareness of these issues and drawing attention to how members could contribute high-quality primary biodiversity information on social media sites or structured biodiversity data aggregation portals. Our presentations were well received, and there was much interest in our efforts as well as in the subject in general. We have been in discussion with like-minded Facebook communities such as Dragonflies of Kerala, Ants of India, etc., who have expressed an interest in carrying out similar curation exercises to extract data from Facebook groups. However, the success of such initiatives depends on the availability of curators and verifiers who are prepared to invest their time and energy. We will make further attempts to find curators who are willing to invest time and effort into these activities, either voluntarily or through further funded projects. We are also looking into the potential for utilizing non-Facebook groups, like email groups that focus on biodiversity discussions, to mobilize such data. Further refinements and fine-tuning to the pipeline, its usability and enhancements will need to be carried out as and when such opportunities arise.

Best Practices and Lessons Learned

The curators' efforts are the key element in producing clean data, even though the interface is simple and convenient to use. Curation efforts require devoted time and effort and are frequently challenging to sustain without financial support. Future initiatives that provide curators with stipends could significantly encourage the process.
Facebook groups, which we identified as our target data source, proved to be very difficult to use because of their strict data sharing policies and active mechanisms for preventing data extraction. This made it difficult for us to gather the data we needed. Although Facebook groups continue to be a valuable source of data, developing a reliable and repeatable method of data extraction from them will be difficult. The interface, however, has been created to be sufficiently general to accept any tabular data as a CSV file containing textual occurrence data.
In retrospect, a workshop could have been planned specifically for admins of specific Facebook groups where the similar data is available. It's possible that time and travel restrictions during our final workshop limited participation overall and, consequently, exposure to the curation interface. However, we can still accomplish this after the project duration.
Even though some groups have expressed interest in using the interface for similar data extraction exercises, they have also stated that it would be desirable to be able to support curators and validators with at least a minimal financial support.

Post Project Activity(ies)

Collaborating with organizations and communities that have been compiling textual or social media-based data on different taxonomic groups to show the platform's utility and ease data curation
Enhancing the curation user interface to make it universal for all data sources.
The technical aspects of the workflow, the curation efforts, and the finished product will be published as a paper in a suitable journal of biodiversity informatics.
obtaining additional funding to expand the features and functionality of the pipeline as well as to conduct additional data mobilization exercises.

Sustainability plans

The infrastructure of the India Biodiversity Portal has been used to build the curation interface. It is now a part of the Biodiv codebase and will continue to receive backend support in the near future so

that it will be available on IBP. We plan to further extend its usage by collaborating with organizations and communities that have been compiling textual or social media-based data on different taxonomic groups to show the platform's utility and ease data curation.

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.