Data cleaning of Heilongjiang Virtual Herbarium data(Focus on NEFI)

Programme:BIFA
Project ID: BIFA5_031
Project lead organization: Northeast Forestry University
Project implementation period:1/7/2020 - 31/8/2022
Report approved: 23/2/2023

Narrative Final report

Executive Summary

After more than one year of work, we have accomplished the following.
1. Understood the basic process of preparing, cleaning, and publishing data. I attended the training and got into the band.
2. Taught the knowledge and skills learned to other group members through group meetings.
3. Discussed the knowledge and skills with other experts from home and abroad.
4. Obtained specimen data from CVH, thoroughly checked and cleaned up information such as place name, scientific name, and collector, and published specimen data in GBIF.
5. Wrote a paper on the data of the plant list of Wandashan (in submission) and published the plant list of Wandashan in GBIF.
6. Wrote and published the "List of Vascular Plants of Heilongjiang Province."
7. Organized and published the place names of Heilongjiang Province in GBIF.

The critical lessons learned and best practices identified□
1. We encountered many technical difficulties when publishing, and thanks to the help of □□□ (melissaliu), the IPT Content Manager of TaiBIF and □□(Chihjen Ko),Asia Regional Support of GBIF, we finally completed the data publication. The key lesson was that we should get a standard template before the operation.
2. The difficulties encountered in cleaning the specimens were so many that it was difficult to give a standard procedure. However, it is necessary to pay great attention to the change of geographical names, be alert to the errors in the original records, rate the difficulties, and give up complex individual problems when appropriate.
The goal of "publishing a list of plants of Heilongjiang Province" was added to the implementation. Since the journal Biodiversity organizes the writing of provincial plant lists throughout China, we thought this would be an excellent opportunity to present our work.

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 obtained these data successfully and analyzed them. The main things we faced were missing place names, incorrect scientific names, incorrect spelling of collectors, and wrong collection dates. We did not include identification errors in our work. With the help of the secretariat and the dedicated person assigned by the secretary, we completed the data publication and other work.

Completed activities

Activity name: Get and inventory analysis of data sets

Description: We have obtained specimen data from CVH and get gazetteer and scientists information from the National Statistics Bureau of China, Specimens data, Floras, Manchuria historical atlas. We have inventory analysis of data sets, including the proportion of different types of information missing, the type of error information. We have complement missing information, such as city, date. Correct errors information, such as location, date, collector, names. Synonymize plant names. Check the result by GBIF Data Validator. We get technical support from the training of BIFA and some other courses on data cleaning, collection history, and other taxonomic and geographic knowledge from "Plant Systematic Community□China□.
Start Date - End Date: 1/7/2020 - 20/2/2021
Verification Sources: DOI:10.15468/aehnzp

Activity name: Get certification

Description: One member of my project team (Hongfeng Wang) received certification following the BIFA capacity enhancement workshop. technical support in the process of data cleaning, including the At the same time, we held a seminar to training on BRAHMS, IPT, TPL, TNRA, the collection history of important collectors, and other taxonomic and geographic knowledge.
Start Date - End Date: 10/3/2021 - 4/5/2021
Verification Sources: https://openbadgepassport.com/app/social

Activity name: Seminars and develop workflow

Description: On October 30, 2020, a group of 16 data experts, plant taxonomy experts, biodiversity experts, historical geography experts, and students convened a seminar to discuss the tools for data cleaning and data we obtained. 1.We compared various tools such as TNRS, LVCP, POWP, WCSP, WHP, and those provided by GBIF. 2. We discuss the characteristics of the data; how to deal with the historical evolution of administrative divisions, the conversion of geographic units to organizational units, and 3. We developed a set of workflows.
Start Date - End Date: 30/10/2020 - 30/10/2020
Verification Sources: documents “Seminars”

Activity name: publish gazetteer

Description: Village-level administrative divisions in Northeast China
Start Date - End Date: 16/1/2022 - 17/1/2022
Verification Sources: 10.15468/4x7279

Activity name: publish checklist_wanda

Description: This is the first checklist of Spermatophyta and invasive plants in Wanda Mountains, including native species and invasive species a total of 97 families 355 genera 716 species and infraspecific taxa are listed. There are 95 families, 329 genera, and 669 species and infraspecific taxa of native plants; 19 families, 37 genera, and 47 species of invasive plants. The resource data includes not only the taxa information but also the record notes.
Start Date - End Date: 16/4/2022 - 17/4/2022
Verification Sources: 10.15468/mwdhkq

Activity name: publish specimen_records_nefi

Description: There are 44642 digital records in CVH belongs to NEFI and many errors can be find in these records (about 30%). While what we've done is far from perfect, we believe most mistakes have been corrected. We can't consider this work to be completely complete, so we will update the cleanup results from time to time.
Start Date - End Date: 12/4/2022 - 12/4/2022
Verification Sources: 10.15468/aehnzp

Activity name: publish data paper

Description: Checklist of tracheophyte in Heilongjiang Province
Start Date - End Date: 30/6/2022 - 30/6/2022
Verification Sources: https://www.biodiversity-science.net/CN/10.17520/biods.2022184

Report on Deliverables

Production of Deliverables - Summary

We have initially completed the "Gazetteer of Northeast China."
Through this project, we have additionally obtained the firstChecklist of Xiaoxing'anling in Heilongjiang Province, which has been uploaded to GBIF. We publish a data paper “Checklist of tracheophyte in Heilongjiang Province” and a database of 44642 digital records in CVH belongs to NEFI

Production of deliverables

Title: Gazetteer

Type: Dataset

Status update: Uncertain
Dataset scope: place names appearing in the specimen of NEFI}
Expected number of records: 56633
Data holder: Ministry of Civil Affairs of China
Data host institution: Ministry of Civil Affairs of China
Sampling method: Heilongjiang Province
% complete: 100
DOI: 10.15468/4x7279
Expected date of publication:

Title: checklist_wanda

Type: Dataset

Status update: Uncertain
Dataset scope: Wanda Mountains}
Expected number of records: 691
Data holder: Hongfeng Wang
Data host institution: Northeast Fforestry University
Sampling method: From 2018 to 2020, we set 25 line transect and 420 quadrats, 60 for arborous plants (25 m × 25 m), 120 for shrub plants (5 m × 5 m), and 240 for herbaceous plants (1 m × 1 m) in Wanda Mountains while undertaking the project “The risk of invasive plants in Heilongjiang Province.” All species, both in quadrats and line transect, have been recorded.
% complete: 100
DOI: DOI10.15468/mwdhkq
Expected date of publication:

Title: specimen_records_nefi

Type: Dataset

Status update: Uncertain
Dataset scope: NEFI}
Expected number of records: 44642
Data holder: Hongfeng Wang
Data host institution: CVH
Sampling method: Specimen records from NEFI
% complete: 100
DOI: 10.15468/aehnzp
Expected date of publication:

Title: Checklist of tracheophyte in Heilongjiang Province

Type: Data Papers
Description: Data paper published in the journal "Biodiversity"
Sources of verification: https://www.biodiversity-science.net/CN/10.17520/biods.2022184

Impact of COVID-19 pandemic on project implementation

The COVID-19 pandemic has caused some impact on visits to the herbarium, but did change our plan.

Events

Get and inventory analysis of data sets

Dates: 2020-07-01 - 2021-02-01
Organizing institution: Northeast Forestry University
Country: China
Number of participants: 5
Comments:
Website or sources of verification: https://ipt.taibif.tw/resource?r=specimen_records_nefi&v=1.5

Events

Seminars

Dates: 2020-12-05 - 2020-12-12
Organizing institution: Northeast Forestry University
Country: China
Number of participants: 7
Comments:
Website or sources of verification:

Communications and visibility

We have completed the geographical name collection of the northeastern region. This place-name collection includes a list of place-names in Northeast China and the history of place-name changes in the past 100 years. We have conducted a preliminary cleaning of the specimen records and corrected most scientific names. We have checked the records of thousands of specimens in the herbarium one by one and corrected the error in "collection person," "collection time," and "location."

Monitoring and evaluation

Final Evaluation

About 2,000 specimen records from northeastern China were included in Gbif before; upon completion of this project, the total number of records will reach nearly 50,000, an increase of more than 20 times. Although it is still far from meeting the needs, it has dramatically improved data availability. There are still about 200,000 digital specimens in the region and an estimated 400,000 undigitized specimens. Much similar work is needed to improve the quantity and quality of digital specimens.
CVH was very supportive of this project, and after the completion of the project, CVH leaders wanted us to introduce the relevant experience to our counterparts in China, and we intended to share some of the workflows with you. In the later stage of the project, the secretariat assigned a dedicated facilitator, which was very effective and greatly improved our efficiency in troubleshooting technical difficulties. We hope that the secretariat will continue to post facilitators to help us publish relevant technical guidance materials to enhance the dissemination of knowledge and technology.

Best Practices and Lessons Learned

The difficulties encountered in clearing specimens are so numerous that it is difficult to give a standard procedure. However, it is necessary to pay great attention to the change of geographical names, be alert to the errors in the original records, rate the difficulties, and give up some issues in due course.
The goal of "publishing a list of plants of Heilongjiang Province" was added to the implementation. Since the journal Biodiversity organizes the writing of provincial plant lists throughout China, we thought this would be an excellent opportunity to present our work.
The biggest challenge in the implementation of the project was the identification of the original information in the specimens, and the most challenging part was the technique of publishing the data.

Post Project Activity(ies)

After the project, we plan to present our workflow to my colleagues in China.

Sustainability plans

There are still roughly 200,000 digital specimens in the region and an estimated 400,000 undigitized specimens. We plan to continue to promote the digitization and cleaning of specimen records.

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.