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 - 31/12/2021
Report approved: 12/5/2022

Narrative Midterm report

Executive Summary

1. We have completed the geographical name collection of the northeastern region. This place-name collection includes a list of place-names in Northeast China.
2. We have conducted a preliminary cleaning of the specimen records and corrected most scientific names.
3. 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."

We regularly classify various errors in specimen records, estimate the proportion of different types of errors, and evaluate our actual work progress and speed accordingly.

Progress against milestones

Has your project published at least one dataset through GBIF.org?: Yes

Dataset published:

Dataset	DOI
checklist_wanda	10.15468/mwdhkq

Has at least one member of your project team received certification following the BIFA capacity enhancement workshop?: Yes

Name of the workshop participant: Hongfeng Wang

Certification obtained: Basic Badge

Report on Activities

Activity progress summary

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□."

One member of my project team (Hongfeng Wang) received certification following the BIFA capacity enhancement workshop

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”

Report on Deliverables

Deliverables progress 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.

Progress towards deliverables

Title: gazetteerofnechina

Type: Dataset
Status update: First edition
Dataset scope: Northeast of China
Expected number of records: 554797
Data holder: Hongfeng Wang
Data host institution: China Civil Affairs Bureau
Sampling method:
% complete: 100
DOI: 10.15468/4x7279
Expected date of publication: 2021-02-01

Title: checklist_of_wanda

Type: Dataset
Status update: published in January 19, 2022
Dataset scope: Wanda Mountains, one of the main mountainous regions in the east of Heilongjiang Province, is facing Russia across the Ussuri River (latitude 44°51′13″ to 47°10′30″ North, longitude 129°30′20″ to 134°10′10″ East), with a total area of 4486,000 hm2
Expected number of records: 691
Data holder: hongfeng wang
Data host institution: Northeast Forestry 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: 10.15468/w2fy2u
Expected date of publication:

Events

Get and inventory analysis of data sets

Dates: 2020-07-01 - 2021-02-01
Organizing institution: CVH
Country: China
Number of participants: 5
Comments:
Website or sources of verification: https://ipt.taibif.tw/resource?r=specimen_records_nefi&v=1.0

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

Monitoring and evaluation findings

About 60% of errors can be resolved quickly, and 15-20% of the remaining errors can be resolved in batches by developing software algorithms. However, 20-25% of errors are more complicated and need to be resolved manually. Nevertheless, this part can be completed in a tolerable time.

Impact of COVID-19 pandemic on project implementation

The COVID-19 pandemic has caused some impact on visits to the herbarium, but we are now assessing that our project should not need to change the plan.

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.


GBIF