Gorteria – Dutch Botanical Archives 46, 2024: 029–033
ISSN (online) 2542-8578

Georeferencing herbarium specimens of the Naturalis botanical collection using automation and crowdsourcing

L.B. Sparrius [1, 3], D.D. van der Hak[1], J.J. Wieringa [2]

Key words

georeferencing
crowdsourcing
location data
museum collection
collaborative work
hand writing
vascular plants

Abstract – Digitalized botanical collection data often lacks location data in the form of geographical coordinates, limiting a wider use of the data. Specimens of Dutch plants, mosses and fungi in the Naturalis botanical collection (herbarium codes AMD, L, U, and WAG) were initially digitized with only 3% of the records having geographical coordinates. In a two year project, coordinates were added to the specimen records by means of (semi)automated techniques and crowdsourcing. Coordinates were inferred from location data present on the specimen labels such as toponyms (including those with typographical errors and abbreviations), grid square codes and other local coordinate systems. Finally, 75% of about half a million specimens of vascular plants could be georeferenced with a precision of 5 km or better.

Samenvatting – Digitale collectieinformatie bevat vaak locatiebeschrijvingen zonder geografische coördinaten, waardoor het gebruik voor andere doeleinden vaak beperkt is. Herbariumcollecties van Nederlandse planten, mossen en paddenstoelen in het herbarium van Naturalis (met herbariumcode L en vroegere herbariumcodes AMD, U en WAG) zijn tussen 2009 en 2015 gescand en de etiketgegevens gedigitaliseerd, maar van slechts 3% van deze exemplaren waren de coördinaten op de juiste manier geregistreerd. In de vorm van een tweejarig project hebben FLORON en Naturalis zijn de (verbeterde) coördinaten aan de database toegevoegd, deels ging dat semi-automatisch en deels met hulp van een grote groep vrijwilligers. De coördinaten zijn afgeleid van de locatiebeschrijvingen op de etiketten van de collecties, zoals toponiemen (vaak met fouten vanwege slecht leesbare handschriften), atlasblokcodes en andere coördinaatsystemen. Bij afronding van het project had 75% van de half miljoen herbariumexemplaren een coördinaat met een nauwkeurigheid van ten minste 5 kilometer.

Published on 12 June 2024

INTRODUCTION

About half a million botanical specimens collected in the Netherlands are stored at Naturalis Biodiversity Center in Leiden (AMD, L, U & WAG). Most of this collection was digitized between 2009 and 2015 during one of the first mass-digitalization projects worldwide. Digitization included not only scanning herbarium sheets, but also transcription of labels (Heerlien et al. 2015). The data was curated in the BRAHMS herbarium management system (University of Oxford 2022), and published through the Naturalis Bioportal and as a dataset in the Global Biodiversity Information Facility (GBIF) (Bijmoer et al. 2022).

Location data, however, was limited to information present on the herbarium labels, which is in most cases a location name or a map grid square in a local spatial reference system (Fig. 1). Primary biodiversity data without geographical coordinates has little value, as it cannot be used for spatial analyses below country level (Townsend Peterson et al. 2018). Thus, to increase the scientific value of the digital botanical collection geographical coordinates must be assigned to specimens, a process known as georeferencing or – to be more precise – geocoding. A general approach with best practices for georeferencing has been documented by Chapman and Wieczorek (2020).

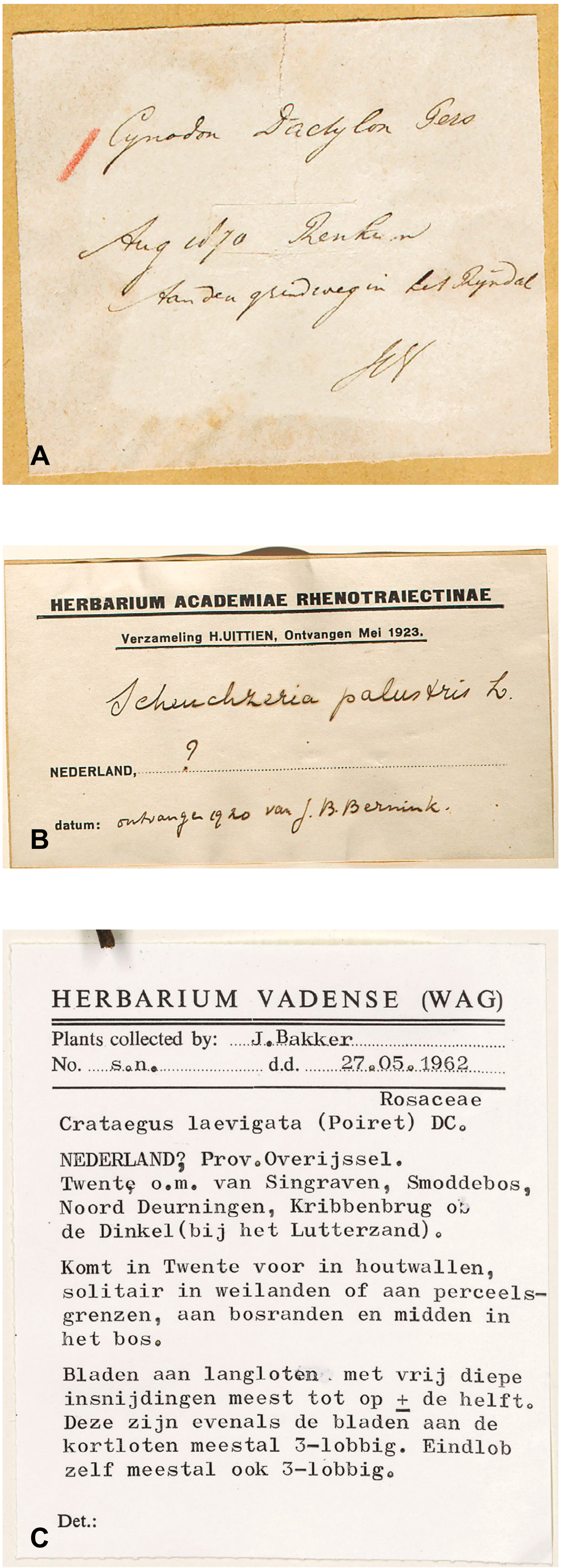
Geocoding can be a time-consuming process, as it requires the knowledge and interpretation of the many ways in which a location can be described, the variation in spatial scale and the translation of local coordinate systems. Regarding location descriptions, crowdsourcing can be useful to help interpret toponyms (Marcer et al. 2021). In this paper, we describe the process of geocoding the botanical collection of Naturalis Biodiversity Center for specimens collected in the Netherlands with help of volunteers of FLORON Plant Conservation Netherlands during 2020 and 2021.

Improving botanical data in the Netherlands helps to fill the data gap for distribution data of vascular plants identified by

[1] FLORON Plant Conservation Netherlands, Toernooiveld 1, 6525 ED Nijmegen, the Netherlands;
[2] Naturalis Biodiversity Center, Darwinweg 2, 2333 CR Leiden, the Netherlands; e-mail: jan.wieringa@naturalis.nl;
[3] Dept. of Animal Ecology and Physiology & Experimental Plant Ecology, Radboud University, Heyendaalseweg 135, 6525 AJ Nijmegen, the Netherlands;

e-mail for correspondence: sparrius@floron.nl

Sparrius et al. (2014), concerning the period 1950–1980, a period of time in which little botanical surveys were conducted.



A

Cynodon Dactylon Pers

Aug 1870 Renkum

Aan den grindweg in het Rijndal

HH

B

HERBARIUM ACADEMIAE RHENOTRAIECTINAE

Verzameling H.UITTIEN, Ontvangen Mei 1923.

Scheuchzeria palustris L.

NEDERLAND, ?

datum: ontvangen 1920 van J. B. Bernink.

C

HERBARIUM VADENSE (WAG)

Plants collected by: J.Bakker

No. s.n. d.d. 27.05.1962

Rosaceae

Crataegus laevigata (Poiret) DC.

NEDERLAND: Prov.Overijssel.
Twente o.m. van Singraven, Smoddebos,
Noord Deurningen, Kribbenbrug ob
de Dinkel(bij het Lutterzand).

Komt in Twente voor in houtwallen, solitair in weilanden of aan perceelsgrenzen, aan bosranden en midden in het bos.

Bladen aan langloten met vrij diepe insnijdingen meest tot op ± de helft. Deze zijn evenals de bladen aan de kortloten meestal 3-lobbig. Eindlob zelf meestal ook 3-lobbig.

Det.:

Fig. 1. Examples of labels of herbarium specimens with: A. unambiguous and complete information or taxon, location, collecting date, and collector name (L.3076286); B. unknown location and collecting date (U.1593783, with only the country name); C. ambiguous location data (WAG.1503626, with many different location names).

METHODS

Data management

Specimen data was exported from BRAHMS into a plain csv file, containing the following spatial information: latitude and longitude (EPSG:WGS84), data in a local coordinate system (Amersfoort/RD New EPSG:28992), IVON grid square notation (format e.g. M3.43.45), kilometer grid square notation (format e.g. 51.23 for 5 km square, or 51.23.34 for a 1 km square) and toponyms (province, city, and/or site). Digital maps of the two Dutch grid systems are available as open data (Sparrius 2022).

Coordinates in local grid systems were not always registered in the correct data columns, but sometimes added to the location name. For this reason, site descriptions were searched for grid squares or local coordinates by using regular expression operators for the various ways in which they were written in the labels, e.g. 51.23, 51-23, 140-345, 140/345.

Assigning tentative locations

The next step was to assign tentative locations to specimens, based on automated attribution of geographical coordinates based on other coordinate system or text extraction.

Specimen records containing coordinates and valid grid squares were converted to latitude and longitude (EPSG:WGS84). The spatial precision, mostly 1 or 5 km, was also noted. Spatial precision of 1 and 5 km were most frequently used, matching the grid square sizes of mapping grids commonly used by botanists.

Location descriptions (text field) were scanned for matching location names on the topographical map of the Netherlands (TOP-10NL) and the Google Geocoding API (Goldstein et al. 2014).

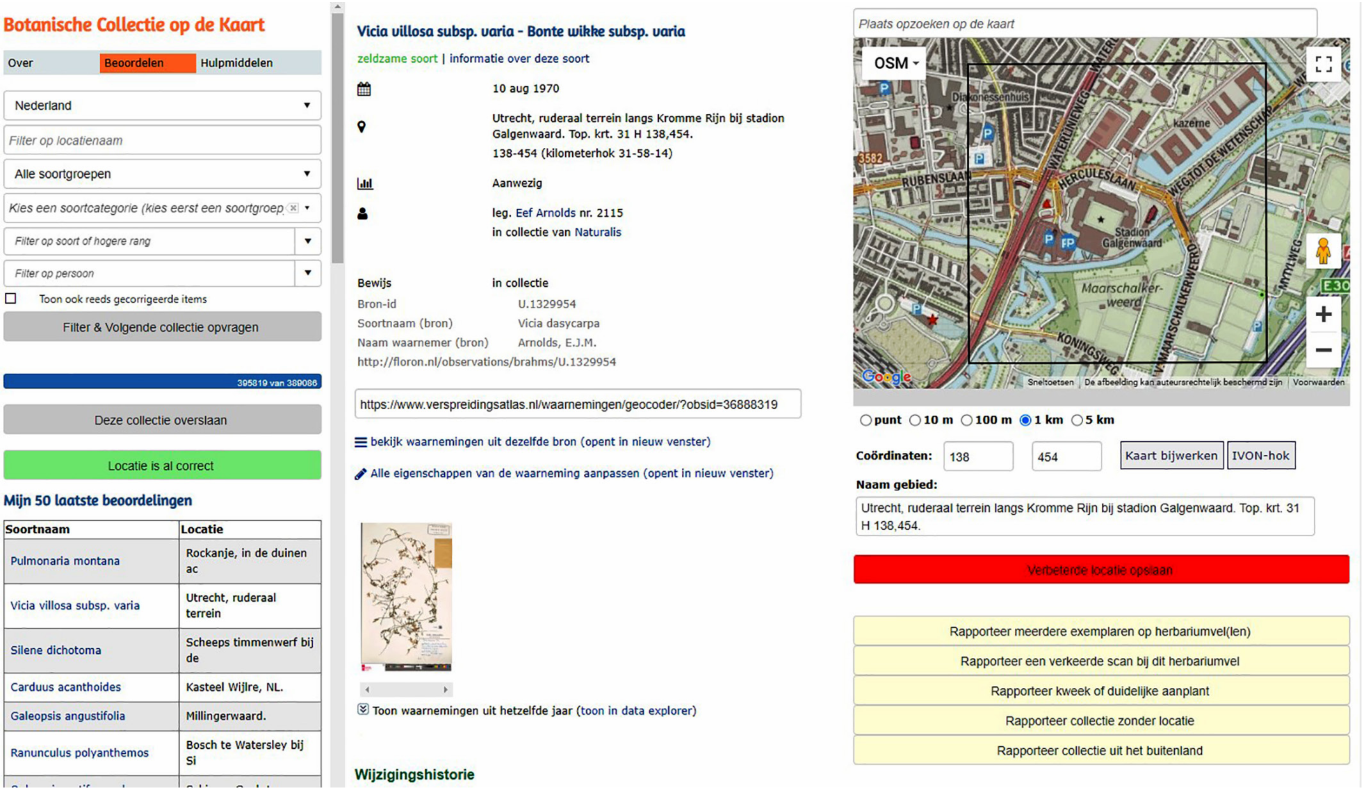
Many toponyms, however, contained typographical errors or had historical names that are no longer used on current maps. For these records, names have been corrected manually and the corrected names were again scanned for matches. All remaining records with toponyms (about 10%) were not assigned to a tentative location. Many different approaches were used to interpret variants of toponyms and location names with abbreviations and typographical errors.

Some locations have similar or identical toponyms, such as the cities and villages 'Zwolle' and 'Hengelo'. Specimens were manually tentatively assigned to the biggest city. In the case of 'Tienhoven' a few specimens were assigned to the right location based on their ecology (rich fens versus sandy river banks).

Crowdsourcing tool

Crowdsourcing was performed in a webtool as part of the National Database Flora and Fauna Atlas (www.verspreidingsatlas.nl), build on the ASP.NET platform and a MySQL database. The crowdsourcing geocoding tool 'Botanical Collection on the Map' (Fig. 2) has the following specifications:

- The user gets a random specimen or selects specimens based on a filter on year, person, or province. The tentative location is shown on the map. The user can also zoom in on the digital herbarium specimen with the original label and annotations.
- The user interprets the location and may choose between the options 'Location is correct', 'I don't know', and correcting the location by clicking on the map.



Botanische Collectie op de Kaart
Over
Beoordelen
Hulpmiddelen
Nederland ▾
Filter op locatienaam
Alle soortgroepen ▾
Kies een soortcategorie (kies eerst een soortgroep ▾
Filter op soort of hogere rang ▾
Filter op persoon ▾
Filter & Volgende collectie opvragen
Deze collectie overslaan
Locatie is al correct
Plaats opzoeken op de kaart
OSM ▾
Sneltoetsen
De afbeelding kan auteursrechtelijk beschermd zijn
Voorwaarden
Kaart bijwerken
IVON-hok
Verbeterde locatie opslaan
Rapporteer meerdere exemplaren op herbariumvel(len)
Rapporteer een verkeerde scan bij dit herbariumvel
Rapporteer kweek of duidelijke aanplant
Rapporteer collectie zonder locatie
Rapporteer collectie uit het buitenland
Wijzigingshistorie

□ Toon ook reeds gecorrigeerde items

395819 van 389086

Mijn 50 laatste beoordelingen

Soortnaam	Locatie
Pulmonaria montana	Rockanje, in de duinen ac
Vicia villosa subsp. varia	Utrecht, ruderaal terrein
Silene dichotoma	Scheeps timmenwerf bij de
Carduus acanthoides	Kasteel Wijlre, NL.
Galeopsis angustifolia	Millingerwaard.
Ranunculus polyanthemos	Bosch te Watersley bij Si

Vicia villosa subsp. varia - Bonte wikke subsp. varia

zeldzame soort | informatie over deze soort

10 aug 1970

Utrecht, ruderaal terrein langs Kromme Rijn bij stadion Galgenwaard. Top. krt. 31 H 138,454.
138-454 (kilometerhok 31-58-14)

Aanwezig

leg. Eef Arnolds nr. 2115
in collectie van Naturalis

Bewijs	in collectie
Bron-id	U.1329954
Soortnaam (bron)	Vicia dasycarpa
Naam waarnemer (bron)	Arnolds, E.J.M.

http://floron.nl/observations/brahms/U.1329954

https://www.verspreidingsatlas.nl/waarnemingen/geocoder/?obsid=36888319

bekijk waarnemingen uit dezelfde bron (opent in nieuw venster)

Alle eigenschappen van de waarneming aanpassen (opent in nieuw venster)

Toon waarnemingen uit hetzelfde jaar (toon in data explorer)

○ punt ○ 10 m ○ 100 m ● 1 km ○ 5 km

Coördinaten: 138 454

Naam gebied:
Utrecht, ruderaal terrein langs Kromme Rijn bij stadion Galgenwaard. Top. krt. 31 H 138,454.

Fig. 2. User interface of the geocoding tool ‘Botanical Collection on the Map’ with selection tools (upper left), recent changes made by the volunteer (lower left), specimen info and herbarium sheet (middle column), location editing (upper right), and buttons for submitting changes and reporting digitizing errors (lower right).

The tool was also used to make a notification about digitalization errors, such as multiple specimens or locations on a sheet, specimens with no location information at all, or specimens from other countries.

When the location is corrected, approved, or ignored, the next specimen is shown. Users are able to see a progress bar, a list of the specimens that they approved or amended, and a list of top-ranking participants.

The project ran from April 2020 until April 2022. The circa 300 participants included volunteers with a background in botany and/or history and were mainly sourced among FLORON volunteers. The tool is still running at a slow pace and may also be used for georeferencing new datasets.

Assigning tentative locations and final quality checks

Finally, records were checked for coordinates outside of the Netherlands, or in large waterbodies. This resulted in a small number of manual corrections and specimens attributed to other countries.

For very rare species, occurring in fewer than 16 locations, locations were adjusted if they have been reported from neighbouring km grid cells in the same year in a reference database of digitized plant records from botanical literature . This is done to avoid the superfluous appearance of two grid squares next to each other on distribution maps.

For all other species, records were validated according to NDFF validation procedures. This includes a check against validated records from a different source, where the species was recorded in the same 5 × 5 grid square with maximum of 10 years apart.

Publishing in GBIF

In the geocoding tool the location is represented as a polygon (grid squares and circles), within which an observation has been made. The first step to import data in BRAHMS was to convert polygons to centroids with a radius of uncertainty (Wieczorek et al. 2004). In BRAHMS, uncertainty is stored as the number of decimals of the latitude and longitude (BRAHMS field LLRES). The geocoding method (e.g. manual, Google Maps API or grid cell) was also stored in BRAHMS.

The specimen records in BRAHMS were updated with the new latitude, longitude, and uncertainty. The BRAHMS data was then published in Bioportal (bioportal.naturalis.nl) and subsequently in GBIF (Bijmoer et al. 2022). A copy of the dataset, with the original polygons assigned by the geocoding tool, is included in the National Database Flora and Fauna (www.ndff.nl). The use of BRAHMS identification numbers makes it possible to update data from BRAHMS to NDFF and vice versa.

Cost of the project

The project was funded by Netherlands Biodiversity Information Facility (NLBIF). The total costs for the project were € 50,000, or € 0.13 per record, which is slightly less than georeferencing in the digitization project iCollections at the Natural History Museum, London (NHMUK; Blagoderov et al. 2017). Moreover, the developed webtools are available for future projects.

RESULTS

The type of location data retrieved from location descriptions were mostly toponyms and a smaller number of geographical coordinates and grid cells (Table 1).

Table 1. Different approaches for geocoding and numbers of specimens. * Toponym was often present together with one of the coordinate fields.

Location data retrieved from text fields	Number of records
Amersfoort/RD New (EPSG:28992) coordinates	774
Latitude and longitude (EPSG:WGS84) coordinates	5
Kilometer grid square notation	3,850
IVON grid square notation	29,814
Toponym*	42,5084

Crowdsourcing resulted in direct approval of the location for 138,568 records, while 256,923 records were either corrected, or had their precision improved by volunteers. Furthermore, 8798 digitalization errors were found. The most common error was the presence of multiple specimens from different locations on one herbarium sheet with only one location record present in the collection database.

Of the original BRAHMS records, less than 3% contained coordinates, half of which with high precision (< 1 km). After georeferencing, roughly 75% of the records contained coordinates, 35% of which with high precision (< 1 km) (Table 2).

Excluded and non-georeferenced records

The dataset still contains about 142,000 non-georeferenced records. These include 12,000 records with no location at all, and 130,000 records that could not be processed in the geocoding tool due to missing specimen barcodes, missing dates, cultivated plants in gardens, or ambiguous data. The last category also includes specimens from ‘herbarium Perin’ collected between 1840 and 1860, which contain a large number of falsified specimens with species that probably never occurred at the given location (Vuijck 1901: 629). Although this has nothing to do with the barcoding process, we wanted to only georeference data that is suitable for biodiversity analyses. We estimate that about half of the remaining non-georeferenced records can be referenced in future. Those records might contain sufficient location information, but require much more expert time to decipher or interpret.

DISCUSSION

Assigning tentative locations and manual georeferencing

There are many different ways in which coordinates and toponyms are written on specimen labels. In the subsequent collection digitization process, new errors have been made in transcribing the label data. This makes it hard to describe the exact steps of the georeferencing procedure. As explained in the methods section, many different approaches for the parsing of toponyms and coordinates were used. This also implies that fully automated georeferencing is practically impossible. Especially hand-written pre-1900 specimens are a great source of transcription errors, with vague toponyms and abbreviations of collectors and missing collecting dates. Recently collected herbarium specimens are usually geocoded with coordinates printed on the label and easier to read.

Crowdsourcing and data validation

There are many different ways in which coordinates and toponyms are written on specimen labels. In the subsequent collection digitization process, new errors have been made in transcribing the label data. This makes it hard to describe the exact steps of the georeferencing procedure. As explained in the methods section, many different approaches for the parsing of toponyms and coordinates were used. This also implies that fully automated georeferencing is practically impossible. Especially hand-written pre-1900 specimens are a great source of transcription errors, with vague toponyms and abbreviations of collectors and missing collecting dates. Recently collected herbarium specimens are usually geocoded with coordinates printed on the label and easier to read.

The crowdsourcing tool has mostly been used by experienced botanists who would restrict their work to areas that they were familiar with. However, during data validation, we came across the following common issues, which are also mentioned in Chapman & Wieczorek (2020).

The georeferencing process is partly iterative: in the first iteration specimens are located at or near the supposedly correct location based on specimen label data. In a second iteration, specimens were located at a higher precision. This means that the georeferencing tool should allow multiple changes to a record. For example, a user can filter on a known ambiguous location name and go through all the specimens to assign them to the right location, even if they had already been assigned a location.

In the dataset we included specimens that already have fairly precise coordinates, either from GPS or manually looked up on

Table 2. Presence and spatial precision of geographical coordinates before and after geocoding.

Precision	Before geocoding (numbers of specimens)	After geocoding (numbers of specimens)
< 20 m	5,328	8,095
< 200 m	1,818	5,009
1 km	1,508	133,423
5 km	464	264,936
> 5 km	69	0
unknown	5,211	0
No coordinates / Left to be processed	550,505	142,774
Total	**564,903**	**564,903**

a map by the collector and subsequently added to the label or database. In the crowdsourcing process, volunteers tended to try to make such coordinates even more precise, which is undesirable since they have not more knowledge than the collector did, and such original data should never be altered. This can be solved by carefully selecting the records for improvement in the georeferencing tool.

After georeferencing, about 50 % of all records were validated against reference data in the National Database Flora and Fauna. The other 50 % still requires manual validation by experts before it can be used as a reliable source for biodiversity mapping.

CONCLUSION

Georeferencing is a strong improvement of the quality of mass-digitalized natural history collections. Using crowdsourcing, collection institutes can source knowledge about toponyms and the ecology of plants from a wide audience in a cost-effective way.

DATA AVAILABILITY

The improved specimen dataset is published and regularly updated on GBIF (Bijmoer et al. 2022). The Dutch grid system used for botanical mapping surveys is published as open data (Sparrius 2022). The geocoding tool can be visited on https://www.verspreidingsatlas.nl/waarnemingen/geocoder, but the code is proprietary as the tool is fully integrated in another web platform.

Acknowledgements – The authors thank 300 FLORON volunteers for their help, especially those who contributed with checks of over 5000 specimens: Dick Veenendaal (68324 checks), Christine Tamminga (56456), Ben Bruinsma (34190), Egbert de Boer (23505), Willie Riemsma (21923), Rens van der Knoop (21470), Petra van der Wiel (21079), Hinko Talsma (19291), Lieuwe Haanstra (14126), Jeanette den Herder (13315), Pieter Stolwijk (12373), Koos Ballintijn (9894), Jan Pellicaan (7942), Bert Geerdes (7540), Frank van Gessele (7165), and Wim Vuik (5006). The project was funded by the national GBIF node, the Netherlands Biodiversity Information Facility (NLBIF). We received a grant from Google for using the mapping and geolocation API. We thank Erik Slootweg for his comments on the manuscript.

REFERENCES

Bijmoer R, Scherrenberg M, Creuwels J. 2022. Naturalis Biodiversity Center (NL) - Botany. Naturalis Biodiversity Center. Occurrence dataset. GBIF.org. (https://doi.org/https://doi.org/10.15468/ib5ypt).

Blagoderov V, Penn M, Sadka M, Hine A, Brooks S, Siebert DJ, Sleep C, Cafferty S, Cane E, Martin G, Toloni F, Wing P, Chainey J, Duffell L, Huxley R, Ledger S, McLaughlin C, Mazzetta G, Perera J, Crowther R, Douglas L, Durant J, Scialabba E, Honey M, Huertas B, Howard T, Carter V, Albuquerque S, Paterson G, Kitching IJ. 2017. iCollections methodology: workflow, results and lessons learned. Biodiversity Data Journal 5: e21277. (https://doi.org/10.3897/BDJ.5.E21277).

Chapman AD, Wieczorek JR. 2020. Georeferencing Best Practices. GBIF Secretariat, Copenhagen. (https://doi.org/https://doi.org/10.15468/doc-gg7h-s853).

Goldstein ND, Auchincloss AH, Lee BK. 2014. A no-cost geocoding strategy using R. Epidemiology 25: 311–313. (https://doi.org/10.1097/EDE.0000000000000052).

Heerlien M, van Leusen J, Schnörr S, de Jong-Kole S, van Hulsen K. 2015. The Natural History Production Line: An Industrial Approach to the Digitization of Scientific Collections. J. Comput. Cult. Herit. 8, 1, Article 3. (https://doi.org/10.1145/2644822).

Marcer A, Haston E, Groom Q, Ariño AH, Chapman AD, Bakken T, Braun P, Dillen M, Ernst M, Escobar A, Fichtmüller D, Livermore L, Nicolson N, Paragamian K, Paul D, Pettersson LB, Phillips S, Plummer J, Rainer H, Rey I, Robertson T, Röpert D, Santos J, Uribe F, Waller J, Wieczorek JR. 2021. Quality issues in georeferencing: From physical collections to digital data repositories for ecological research. Diversity & Distrib. 27: 564–567. (https://doi.org/10.1111/DDI.13208).

Sparrius LB. 2022. Dutch grid map layers for biodiversity surveys. (1.0) [Data set]. Zenodo. (https://zenodo.org/records/7640805).

Sparrius LB, van Strien AJ. 2014. Het berekenen van jaarlijkse trends van planten op basis van verspreidingsgegevens [Calculating annual trends of plants based on occurrence data]. Gorteria 37: 31–40.

Townsend Peterson A, Asase A, Canhos DAL, de Souza S, Wieczorek J. 2018. Data Leakage and Loss in Biodiversity Informatics. Biodivers. Data J. 6: e26826. (https://doi.org/10.3897/BDJ.6.E26826).

University of Oxford. 2022. BRAHMS. Management of natural history. Available from: https://herbaria.plants.ox.ac.uk/bol/content/software/v8/BRAHMS_Manual.pdf; last accessed on 22 May, 2024.

Vuijck L. 1901. Prodromus Florae Batavae, ed. 2. F.E. MacDonald, Nijmegen.

Wieczorek J, Guo Q, Hijmans RJ. 2004. The point-radius method for georeferencing locality descriptions and calculating associated uncertainty. International Journal of Geographical Information Science 18: 745–767. (https://doi.org/10.1080/13658810412331280211).