Draft best practice for semi-automated georeferencing

Naturalis Biodiversity Center

Result of study by
Josine Blom
Bachelor student

Wilfred Gerritsen
Information Analyst

Subjects

- What happened previously: short recap
- Results study and discussion
- Next steps: a proposal

Timeline and actions

2013 - 2014

pilot project Georeferencing Naturalis as part of multi-year digitization program

Dec. 2014

Automated Georeferencing meeting 2014 at NBC Leiden

Participants NHM, RCMA, Naturalis

Actions Exchange of detailed information on the (technical) methods the NHM, RMCA and Naturalis have developed for the automatic georeferencing of their collection objects

Conclusions NHM, RMCA and Naturalis have similar and complementary competencies and use standardized GIS tools that fit well together

March. 2015

CETAF ISTC meeting, Joensuu

Conclusions It was agreed that developments for improving geo-referencing workflows could highly benefit from collaborations between CETAF organisations. This includes not only software development activities but also the compilation of supporting data (e.g. itineraries). Naturalis will coordinate the initiative.

Similar conclusions

CETAF concluded during ISTC meeting

International best practice can contribute to an

- international standard method for (semi) automatic georeferencing and
- an infrastructure for all naturalis history collections within the EU for the future

Naturalis concluded during pilot project georeferencing

Application of the best practice - and infrastructure -
should lead to

- an enrichment of natural history databases with reliable and comparable georeferenced data

could lead to

- increase the usability and quality of digital natural history collections

Research study objectives

Main objective

..to draft a best practice for the (semi) automatic georeferencing of the digitized data collection of Naturalis

Detail objectives

the draft best practice should

a. focus on Naturalis and it's specimen collections

b. be recognizable and applicable for other interested institutions of the CETAF on the road to an international standard method and infrastructure for (semi)-automatic gereferencing for all natural history collections within the EU for the future

c. comply as much as possible to the *Principles for the best practice for Georeferencing Biological Species Data (Chapman & Wieczorek,2006).*

so

usability and quality of digital natural history collections is enhanced

Research study general information

Researcher	Josine Blom: Bachelor graduate at the ICT faculty of the Haagse Hogeschool
Tutor	Marian van der Meij, head Information Management Department Naturalis
Time period	from August 2015 until January 2016
In scope	type of georeferencing of exported data (out of CMS system); botanical, zoological en geological data sets
Techniques used	interviews, desk research central use case: Naturalis - used in various steps
CETAF participants	RCMA: Patricia Mergen, RGBE: Elspeth Haston, BGBM: Agnes Kirchhoff, NHM: Malcolm Penn
	NL-BIF: Cees Hoff and Naturalis: biodiversity researcher(s), collection managers, project members georeferencing project, bioportal developer, wikipedian in residence
	Thank you all for participating!

Conclusions: Geographic data

Geographic data can be divided into sets:

set 1: orginal locality description	set 2a: primary metadata fields	set 2b: secondary metadata fields
❏ fields: Country, State provinces, Island, locality, Station number, Full locality text. etc. ❏ allready present in most biodiversity data collections ❏ advice: important information with historic value therefore never overwritten	❏ minimum required set of fields ❏ fields that occur in all data standards, i.e. ABCD, DwC and BioGeomancer.	❏ vary in defintiion or name ❏ precise definition is institute specific ❏ geo ref process fields included ❏ recording depends on content policy institute

Geographic data have data quality issues which need looking into:

- Legacy collections often have data quality issues e.g. missing information, mispellings, historical names
- Recent collections that already contain GPS data can also have data quality issues that could affect georeferencing results, e.g. wrong sign of coordinates, number of decimals, accuracy of GPS

Conclusions: User group needs

Potential user groups of geographical data

- main application fields: research, collection management and accessibility
- broader audience for geographical data then for taxonomical data, i.e: educational institutions, amateur assocations, hobbyist and in many cases the entire public

User needs for geographical data in all application fields

- **biodiversity researchers consider geographical data element in species occurrence data as key; both original location data and interpreted uniform coordinates**
- **an area of 10 x 10 km around an object is accurrate enough for most types of research**
- geographical coordinates as a reference point, as they are more durable and less variable.
- a need for both the original location description as for the coordinates
  → do not overwrite original location
- additional geographic meta data
  - extra fields needed are: uncertainty, coordinate system, datum, coordinate precision and more
- most important geographic field: accurracy
  - accuracy gives indication of the usability of the data to user groups
  - accuracy should be stated clearly without room for misunderstanding by using measurable units such as meters

Overview: Inventory of georeferencing projects

Project / organisation	In scope	Tooling	Data resource used	Guidelines and best practices	Status as of sept/oct 2015
1.SPECimap Georeferencing Software / RBGE, GB	Botanical data mainly from historical collections	in-project-developed SPECimap Georeferencing Software Geoparser tool Google maps	Gazetteer of Britisch Place Names; the Fuzzy Gazetteer; Google maps; different historial maps - old survey maps; UK National Grid reference Unlock	No specific guideliness are mentioned in the report	Development - testing stage
2. StanDAP-Herb / BGBM, Germany	Botanical data	preferred tooling as of yet: GeoLocate include links to web tools	No specific data resource were mentioned in the report	No specific guideliness are mentioned in the report	Development stage Planned time period: 2014 - 2017
3. FCD Pilot Georeferencing project / NBC, Netherlands	Zoological data Collections: Invertebrates | hymenoptera Vertebrates | chiroptera	Google maps Google geocoding API Open Refine MS Excel	Getty Thesaurus of Geographic Names (TGN) GeoNames Google maps	Principles for Best Practice for georeferencing Biological Species Data" Chapman, Wieczorek, 2006)	Finished
4. HerpNET → VertNET Sevent institute project - including RMCA. Belgium and Berkeley university	Zoological data Collections: Vertebrates	GeoLocate BioGeomancer tool including Georeferencing Calculator	Digitized maps of DRC, Burundi and Rwanda GeoLocate sources BioGeomancer sources	HerpNet.MaNIS guideliness Georeferencing for dummies document	Finished
5. SYNTHESYS NA-D 3.7 "Itenerary" project / Multi institute / org project including RMCA, Belgium in BioCase	Zoological data Collections: Amphibia and Reptilia	in-project-developed algortihm to detect which data was constituent with the itinerary and which was not	Expedition itinerary data: like: field notebooks, hand-drawn maps, specimen database records, written comments, rough terrain sketches, digital maps, field number lists	No specific guideliness are mentioned in the report	Finished

Overview: Inventory of georeferencing projects

Project / organisation	In scope	Tooling	Data resource used	Guidelines and best practices	Status as of sept/oct 2015
6. Georeferecing with Google Geocoding API and R / Niels Raes - research fellow, NBC, Netherlands,	Botanical data in data poor areas, e.g. Borneo High accuracy geo ref needed	own georeferencing script (programming language R) to work with Google geocoding API	High resolution satellite images Old expedition maps SRTM digital elevation data (SRTM = Shuttle Radar Topography Mission-NASA)	No specific guideliness were mentioned in the report	Finished, part of research project
7. iCollections, the British and Irish Lepidoptera project / NHM London	Zoological data Collections: Lepidoptera (Irish and British)	Google geocoding API trial with: BioGeomancer OS place name list	Google maps; GeoNames	No specific guideliness were mentioned in the report	Ongoing
8. MITCH: Mining for Information in Texts from Cultural Heritage - part of CATCH/ NBC and two universities in NL	Zoological data Collections: animals	GeoImp with use of manually made gold standard - reference	No specific data resource were mentioned in the report	No specific guideliness are mentioned in the report	Finished 2004 -2009

Other georeferencing tools

Owner	Scope	Tooling	Data resource used	Guidelines and best practices	Operational status
Digitaal Erfgoed Nederland (=Digital Heritage Netherlands)		Histograph Historal geocoder designed for Netherlands	birth places of Dutch East India Company crew; members, monastery records and historical census data for the historical place names; GeoNames TGN for the standardized modern place names	No specific guideliness were mentioned in the report	Finished (DEN site)

Overview: Inventory data cleaning and validation techn.

Data cleaning and validation technique	Tooling	Operational status
Manually visual check with Google maps for afdusting result georeferencing with Google Geocoding API	Google maps	Operational tool
Mismatch information from GBIF	GBIF API	Operational tool
SpeciesLink network web services Species link is a Brazilian project and is therefore only aimed at Brazilian occurrence data. Secondly this data is only available in Spanish and partially in English.		
Identify errors and standardize data	Datacleaning tool	See general remark SpeciesLink project
Detect outliers in latitude, longitude and altitude	spOutlier tool	See general remark SpeciesLink project
Converting different types of geographic coordinate systems and datums	Converter tool	See general remark SpeciesLink project
Calculating conformity score	algortihm to detect which data was constituent with the itinerary and which wasnot	Operational status is not mentioned in report
Standardize Dutch place names in a dataset.	Plaatsnamen standaardiseren	Demo status
As a datacleaning tool Open Refine can be used for several steps like: adapting signs that were copied incorrectly from the registration system to the export file, or removing offset number from the localitydescriptions,	OpenRefine	Operational
Crowd sourcing		

Conclusions: Georeferencing projects inventory

Project including tools++ inventory

- many different practices for approaching georeferencing a collection → not one is fully working and without bugs or questions
- several initiatives in CETAF institutes with project goals
- a lack of collaboration between the interviewed projects / institutes
- manual labor is always needed in georeferenced processes

Conclusions: Tool and methods selection

Tool selection

- tools that were included in tool selection are displayed red in the previous tables, e.g: GeoLocate or Google geocoding API
- georeferencing large batches can best be done with Google geocoding API
- higher accurracy needed: use Georeferencing Calculator or SPECimap

Georeferencing methods / guidelines selection

- two guideliness are in scope, i.e.: Principles for Best Practice for georeferencing and HerpNet.MaNIS guideliness
- Both guidelines discus very important (and similar) parts of the georeferencing process
- “not one that can be called ‘better’ for the focus of this research”

Data resouces

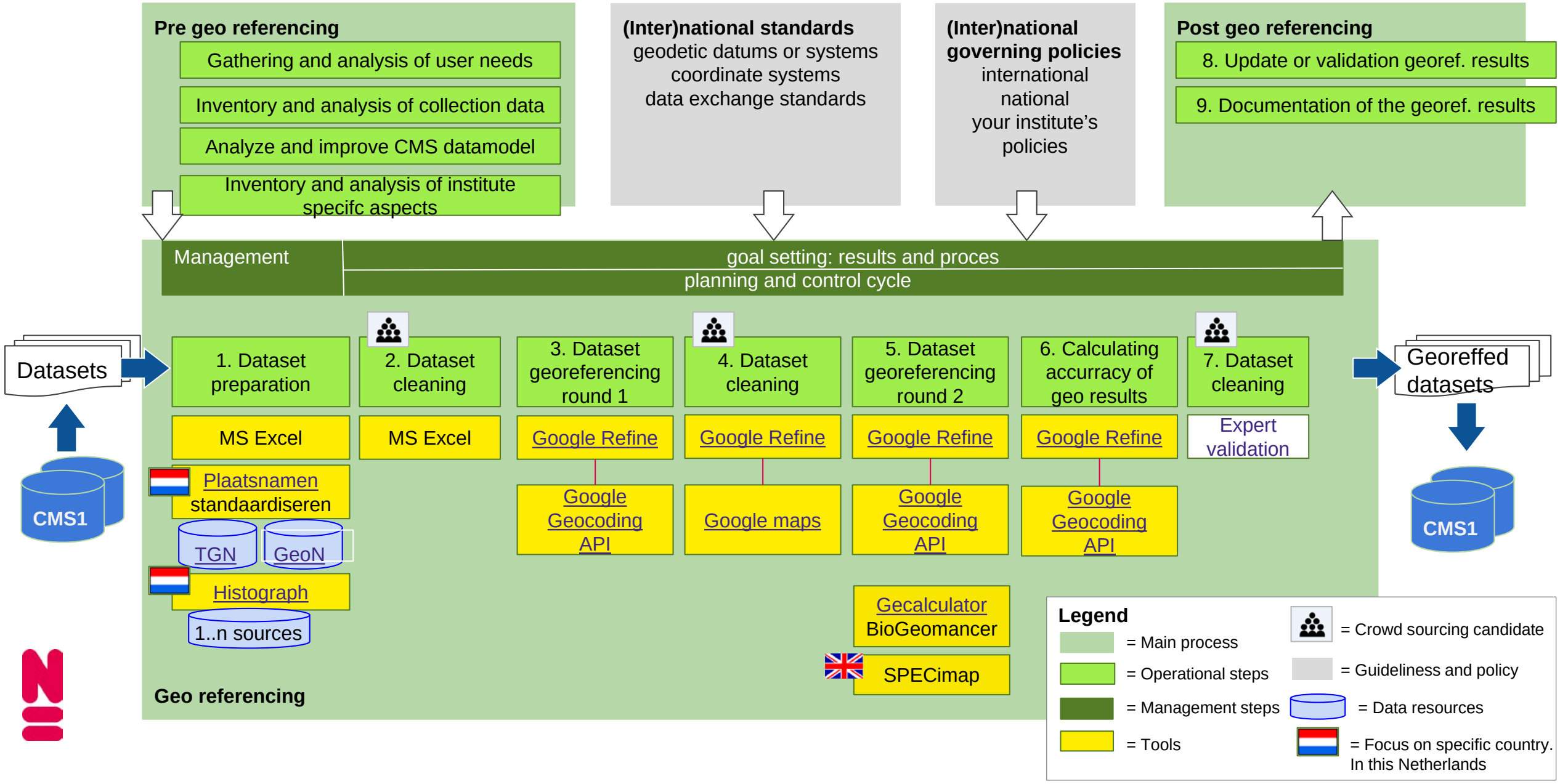
- all described data resources are all online available databases and resources containing geographic information that is useful for georeferencing, like coordinates and place names.
- the TGN and Geonames are only useful for datasets with standardized place names because, they don’t handle deviations in spelling
- Google Maps, is useful for collections without standardized place names

Conclusions: Tool and methods selection

Data cleaning methods

- the question regarding tools to ease the process of datacleaning and reducing the amount of man-hours needed for this, cannot be fully answered
- there are no tools available that only require a very small amount of time and work with large datasets, handle fuzzy data and contain worldwide geographic information
- the GBIF data check, Open Refine and the Visual Check by Google, can however ease the datacleaning process to some degree
- crowd sourcing is ...a way to reduce the amount of man-hours needed to do so

Draft best practice - at a glance



Final conclusion and discussion

Final conclusion

- biodiversity researchers consider geographical data element in species occurrence data as key
- overview georeferencing projects including tooling, data resources used and guideliness / best practices
- overview and selection of tools, georeferencing methods, data resources, data cleaning and validation
- draft best practice georeferencing species occurrence data which includes steps to be taken and rules of thumb

Discussion

- your ideas? reactions?
- could SPECimap be part of the tool solution?
- collaboration and learning

Next steps?

Current situation

- first draft best practice
- next step: collaborative action in order to bring draft to maturity

How

- joint project: parallel testing and improvement draft best process on a tool workbench

- light-weight project approach with main roles:
  - best practice owner — role: process owner, test manager, project lead
  - workbench manager — role: workbench owner and developer
  - data resouce manager — role: data resource expert and supplier
  - workbench support — role: support and environment manager
  - test case and testers: perform tests — individual CETAF institutions who offer data sets and

- Naturalis offers to implement initial workbench with international acces.
  It can act as workbench support.