EXTERNE BIJLAGEN

Het (semi-) automatisch georeferencen van primaire soortgegevens

JOSINE BLOM, 12082619
Informatiedienstverlening en management Haagse Hogeschool

Algemene informatie

Versie	1.0
Titel	Externe Bijlage behorend bij afstudeerverslag: “Het (semi-) automatisch georeferencen van primaire soortgegevens”
Afstudeerder	Josine Blom Eiklaan 46 2282 AV te Rijswijk 06-105 445 51 josine.blom@hotmail.com
Studentnummer	12082619
Afstudeerperiode	31 augustus 2015 (2015-2.1) – 8 januari 2016, 2015-2.1 (2015-2.2)
Onderwijsinstelling	Haagse Hogeschool Academie voor ICT & Media Johanna Westerdijkplein 75 2521 EN Den Haag 070 445 88 88
Studierichting	Informatiedienstverlening en –management, voltijd
Afstudeerbedrijf	Naturalis Biodiversity Center Darwinweg 2 2333 CR, te Leiden 071 751 96 00
Bedrijfsmentor	Dr. Marian van der Meij Informatiemanager marian.vandermeij@naturalis.nl 071 751 793 86
Examinatoren school	Klaas Jan Mollema k.j.mollema@hhs.nl Jochem Mollema jochem.mollema@unafact.nl

Referaat

Josine Blom, 12082619
Externe Bijlage behorend bij afstudeerverslag: “Het (semi-) automatisch georeferencen van primaire soortgegevens”
Naturalis Biodiversity Center, Leiden
September 2015 - januari 2016

Dit document betreft de externe bijlage behorend bij het afstudeerverslag “Het (semi-) automatisch georeferencen van primaire soortgegevens”, van Josine Blom. In deze externe bijlage zijn de tijdens de stage opgeleverde eindproducten voor de opdrachtgevende organisatie. De externe bijlage hebben de huisstijl van Naturalis en zijn in het Engels geschreven.

Het document fungeert als aanvulling op het afstudeerverslag, in het kader van het afstudeertraject bij de Haagse Hogeschool, voor de opleiding Informatiedienstverlening en -management, voltijd.

Trefwoorden:

- Georeferencing
- Primaire soortgegevens
- Geografische informatie
- Internationale methodes
- Pakketselectie
- Standaarden
- Automatisering

Inhoudsopgave

Onderzoeksverslag

How to georeference primary specimendata in a (semi-) automated process and which tools are available and most useful?

Afstudeerstage Josine Blom

Josine Blom, 1208619
Organisatie: Naturalis Biodiversity Center
Bedrijfsmentor: Marian van der Meij

Haagse Hogeschool
Informatie dienstverlening en informatie management
Klaas Jan Mollema (Examinator 1)
Jochem Mollema (Examinator 2)

Naturalis Biodiversity Center

Darwinweg 2 T 071 751 91 02
Postbus 9517 josine.blom@naturalis.nl
2300 RA Leiden www.naturalis.nl

Management summery

One of the most important applications of the collection data of Naturalis Biodiversity Center is its use for scientific research (NBC, 2015a). Besides the taxonomic data, like specimen name, the geographic metadata of the objects gathering event is very important for this purpose. With georeferenced geographical reference points, the object data can be plotted, and the distribution of specimen can be analyzed. Besides fundamental research this allows us the also research global issues such as nature preservation and climate change (NBC, 1205b).

The geographical data could be more useful for science if the object records would be complemented with corresponding coordinates, in a unified data format (NBC, 2015b). This process, of linking a spatial location of an object to a geographic reference system (such as coordinates), so they can be plotted on a digital map, is called georeferencing (Murphey ed al, 2004). At this moment Naturalis only has validated coordinates for a very small portion of the collection. This applies to many more international institutions, collection biodiverse objects.

Some of these institutions have conducted projects or researches with various (test) methods to enrich parts of their collections with coordinates. In cooperation with several of these international institutions of the CETAF (Consortium of European Taxonomic Facilities) a group of experts on this topic was founded (the CETAF GIS experts subgroup). This group focuses on developing a standard method for georeferencing large amounts of primary specimendata.

As a starting point Naturalis has taken the initiative to conduct a study on:

- All possible resources that can be used in the georeferencing process
- How these resources can potential overlap or complement each other
- Which of these resources are the most suitable for primary specimendata.

The goal of this project was to create a manual with methods and recommendations (best practices) for georeferencing large amounts of object records in an automated way (production based). To determine these best practices the following aspects have been studied:

- **Data:** the fields that play a role in the organization of the collection management systems and for the sustainable documentation of georeference data. The recommended elements are based on a study of possible international data standards (ABCD and Darwin Core, the 'BioGeomancer Workbench Guidelines') and a use case of the data models of the collection systems of Naturalis.
- **Users:** 'Fitness for Use', is a concept that states that data quality is strongly related to use and thereby also to the users. For biodiversity collections this means that the value of geographic data is centered on improving / enriching a collection so that it is useful for all fields of application (research, collection management and accessibility of the collection). Based on interviews with these fields of application the needs and requirements could be could be determined. These requirements relate to the metadata that is necessary for the work of the users and the degree of precision which is required for research.
- **Institutions and projects**: Naturalis is not the only institution in the CETAF that experimented with (test) methods for georeferencing. Various institutions have created a tested a part of the georeferencing process. These individual components can potential overlap and complement each other to form a complete process. By engaging in dialogue with these institutions that carried out such a project or research, an inventory was made of all the experiences with georeferencing within CETAF.
- **Tools**: based on the experiences of the interviewed institutions, a list can be made of all the tools and other resources that can be used for georeferencing. The mentioned tools, methods, datasources and datacleaning methods are complemented with relevant information about the functionality, usability, manuals and availability.

All wishes and requirements of users and data quality, capabilities of the tools, experience and field specific quality standards are combined in a tool selection. This lead to a selection and combination of the most useful resources for georeferencing primary specimendata. The goal was not to come to one tool that works best. The expectation from CETAF was that several tools could be useful and could be applied in combination.

Of the relatively long list of tools, methods, resources and datacleaning methods, only a small amount of useful possibilities for georeferencing primary specimendata remains after the tool selection process. This mainly has to do with de availability of tools for this particular topic. Georeferencing is a subject that has been researched and tested by many institutions and projects. However most projects or institutions start from the beginning, without building on experiences and data that is already available. Secondly many institutions design their own georeferencing tool. Consequently, these tools and project are only available for the regarding intuitions and are not shared publicly. So, there is not that much available open source, free and online for all interested parties to use.

Based on the tool selection it can be stated that the Google Geocoding API is the most usable tool for georeferencing primary specimendata. This very clear answer makes the final recommendation easier and ensures that the recommended method in the best practice is useful for everyone. The Tool is easy to use the help of Open Refine, and perfectly matches the datacleaning method of the Visual Check with Google Maps.

Since Naturalis also used the Google geocoding API, for georeferencing their collection, the process that is described here will be very similar to the one they developed. By adding additional tools, datasources and datacleaning methods, my advice would form a complete inventory of experiences with georeferencing within the CETAF. This allows my research and advice to be used for new projects and proceeding steps towards a standard method for georeferencen large amounts of specimendata.

There is enough knowledge and tooling created for georeferencing primary specimendata. The big problem is that this knowledge and tooling disappears after completion of a project or never get s shared with the public or similar institutions. An international effort can contribute to the preservation of this knowledge and tooling, so that not every institution has re-invented the wheel.

Contents

1. Introduction

One of the most important applications of the collection data of Naturalis Biodiversity Center is its use for scientific research, for which Naturalis makes available their entire collection as well as modern research facilities (NBC, 2015a). Besides the taxonomic data, like specimen name, the geographic metadata of the objects gathering event is very important for this purpose. With georeferenced geographical reference points, the object data can be combined with other spatial data from different disciplines such as meteorology, geology, geography, social and medical sciences. Besides fundamental research this allows us the also research global issues such as nature preservation and climate change (NBC, 1205b).

During the digitization of the collection, Naturalis also disclosed the geographical information of the objects in the Collection Registration System. However for most past of the collection this is in textual form, for example: *Leiden, The Netherlands* and these locality descriptions often differ in form. The geographical data could be more useful for science if the object records would be complemented with corresponding coordinates, in a unified data format (NBC, 2015b). This process, of linking a spatial location of an object to a geographic reference system (such as coordinates), so they can be plotted on a digital map, is called georeferencing (Murphey ed al, 2004).

In 2013/2014 a pilot project FCD- Pilot georeferencing was carried out by Naturalis to research and test the possibilities for providing coordinates for a large batch of object records in a mainly automated way (NBC, 2015b).

Besides Naturalis there are more biodiversity institutions, like botanical gardens and natural history museums that require georeference data of their collections, because it gives a clear added value to the collection and can better support modern scientific research. Naturalis is therefore working together with international partners of the CETAF, the Consortium of European Taxonomic Facilities, to share knowledge about georefencing. Collectively, this research group concluded that an international best practice can contribute to an international standard method for (semi) automatic georeferencing and an infrastructure for all natural history collections within the EU for the future (NBC, 2014a). Application of the best practice should lead to an enrichment of natural history databases with reliable and comparable georeferenced data. Additionally such a standard method and infrastructure could increase the usability and quality of digital natural history collections (Arkel ed al, 2015).

During the Automated Georeferencing Meeting in 2014 georeferencing projects of different participating institutions of the CETAF where discussed. All separate surveys and pilots offer a solution to some part of the problem that come with semi- automatic georeferencen, but to prevent a multitude of manuals or guidelines these individual projects should be combined into one international the best practice. To achieve this further research is needed to explore how these different methods can complement each other and be combined in one complete process.

2. Problem Analysis

This research began with a problem analysis to describe and clarify the background and motivation of the pilot project of Naturalis in more detail. The analysis describes the main features and background of the pilot project and the meetings between the CETAF institutions on this subject and the associated research question and sub-questions (Verhoeven, 2014). In additions to this, a list will be added that includes institution that could be considered as interview partners for the research. The problem analysis was based on internal documents of the CETAF- ISTC, the Pilot project Georeferencing Naturalis and reports of the Automated Georeferencing Meeting 2014.

2.1 Problem description

To increase the quality and applicability of the Naturalis dataset it is desirable that the textual geographical object information of (parts of) object records managed by Naturalis are complemented with the corresponding geographical coordinates in a standardized process (Arkel et al, 2015).

During the pilot project FCD Georeferencing online available tools for automated georeferencing where researched and a process was developed with the most useful tool, the Google Maps Geocoding API. Here the Google Geocoding API offers large batches of location data, via Open Refine, to Google Maps and then returns extensive geo-referenced data (Arkel et al, 2015). The pilot was successful and showed that semi-automated georeferencing with Google Maps is a promising method which. Naturalis also concluded that it might be necessary to add 2 new functional metadata elements to their Collection Registration System (CRS), to store the new required georeference data: 'Georeferenced by', and 'Georeferenced when'.

Besides Naturalis there are more biodiversity intuitions, like botanical gardens and natural history museums that require georeference data of their collections, because it gives a clear added value to the collection and can better support modern scientific research. Naturalis is therefore working together with international partners of the CETAF, the Consortium of European Taxonomic Facilities, to share knowledge about georefencing. Within this organization there are several institutions which have carried out a pilot project or research about georeferencing. Together these institutions form an interest group, the GIS expert subgroup, which is willing to work together to form the needed international - best practice.

Following the results of the Pilot Project FCD Georeferencing Naturalis approached all CETAF institutions in December 2013, asking if they have any experience with semi-automated georeferencing and if they were interested to share this information. The London Natural History Museum (NHM) and The Royal Museum of Central Africa (RMCA) in Tervuren showed their interest and Naturalis therefore organized a meeting on semi-automated georeferencen on the 2nd and 3rd of December, 2014.

During this meeting, detailed information was exchanged on the (technical) methods that the NHM, RMCA and Naturalis had developed for automatically georeferencing large batches of collection objects. The meeting revealed that NHM, RMCA and Naturalis have similar and complementary competencies and use standardized GIS tools that fit well together (NBC, 2014b). The NHM and Naturalis mainly presented tools and algorithms that are important on the background of the process, and RMCA showed the many applications and visualization tools that were developed within the framework of various projects in collaboration with other institutions.

Combining these methods from the NHM and Naturalis in the proper way, could lead to a high quality and efficient process for (semi-) automatic georeferencing. This combined method can form the basis in the future for a standard methodology and EU infrastructure for the (semi-) automatic georeferencing complete natural history collections (NBC, 2014a). Furthermore, the desire was expressed to start a follow-up research in which the georeferencing methods of the NHM, RMCA and Naturalis are compared with each other and are combined in one complete process (NBC, 2014a). This study should answer the question: until what point in the process is it possible to georeference collection objects in a (semi)

automatic way and with what tools? The results of this study should form an international best practice that can be used by any biodiversity institution to georeference their collection.

The research described in this report was Initiated by Naturalis, but was based on the meeting Naturalis had with other CETAF members, and the desire, to work together to combine all methods from different institutions and create a best practice. This means the final report knows two target groups.
Naturalis, the first target group, has already done its own research and has executed several tests. The questions that remain for this group are very specific and very much related to their own collection and collection management system (Arkel et al, 2015). The problems/ questions that remain for Naturalis are:

- In what ways is the geographic data listed in the registration and what changes would be necessary to store georeferenced data?
- At what accuracy levels do the users of the collection Naturalis want the objects to be georeferenced, and how should this be marked in the CRS?
- In what ways could we reduce the man hours needed for the heavy labor intensive data cleaning method that was tested and proved effective in the FCD pilot project from Naturalis?

The CETAF, the second group of interest requires a more global research, and that focusses more on the entire georeferencing process. They want to know what tools are available for automatic georeferencing, how they fit together, and where important parts of managing a biodiverse collection, like the historical data, accuracy, data quality etc. fit in the georeferencing process. This results in the following problems/ questions:

- What methods and tools are available to georeferencen large amounts of data in large batches with reliable results and how can these complement each other?
- How to deal with bad data quality, like missing information, misspellings and historic place names?
- Where do the terms accuracy (how close does the georeference point come to the actual gathering site) and reliability (what data is the georeference point based on) fit in the georeferencing process?

2.2 Problem definition

The central question for this research is:
“What do you have to do, as an institution; to enrich the textual geographic location markers in your collection data to the corresponding geographic coordinates at different accuracy levels in a semi-automated way?”

Sub-questions

1. What types of geographic metadata are listed in the collections that need to be enriched with geographic coordinates?
2. What levels accuracy do the users of the georeferenced data of the collection (of Naturalis) want to use?
3. What other institutions have initiated projects to enrich collections geographic data with the corresponding geographical coordinates?
    a. What is the reliability/ accuracy of the methodology/ tools?
    b. To what extent was manual correction necessary in order to improve the reliability of the georeferencing data?
    c. What resources are used to make these corrections?
4. What (semi-) automatic methods / tools are available to enrich textual location marking with corresponding coordinates?
5. What resources are available to make these corrections on a labor-extensive way?
6. What semiautomatic programs are best suited for the described collections?
7. How can these methods / tools be combined in the best way?
8. What methods / programs can be best used to enrich collect geographic data with the corresponding geographical coordinates?

2.3 Project objective

The objective of this follow-up project is to draft a best practice for the (semi) automatic georeferencing of the digitized data collection of Naturalis. This advice should focus on how Naturalis can enrich their collections geographic data with corresponding coordinates. The digitized collections of Naturalis can then be provided with new geographic data, allowing internal and external researchers to map and analyze past and present distribution data of species and to combine them with spatial data from different disciplines (NBC, 2015b).

The best practice is recognizable and applicable to other interested institutions of the CETAF, which ultimately can contribute to an international standard method and infrastructure for (semi) automatic georefencing for all natural history collections within the EU for the future. This includes guidelines for the entire process that is involved semi-automated georeferencen and that scores as high as possible on *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.

3. Datadictionary

The first step of the research is making an inventory of required geographic information of a biodiverse collection and the corresponding metadata elements. Besides the georeference points, the coordinates, there is much more additional available data created during the georeferencing process. In this chapter available standards for enclosing this additional data and the coordinates are examined. This requires the definition and qualification of fields that play a role in the organization of the collection management systems, and are used to store the geographic and georeferenced data. To create a complete image of the required metadata, it is necessary to also look at the fields with geographic data and the interpretations of this data, trough georeferencing.

Secondly this chapter describes a use case (Naturalis) to analyze the differences between the current and the required metadata elements of a taxonomic/biodiversity institution. For this case study the collection management systems and parts of the geographic information in the collection data of Naturalis was studied.

Research Methodology
Through desk research a study is done on available standards that advice metadata elements and requirements for enclosing georeferenced information and are specifically designed for biodiverse collections. For this study, the collections of Naturalis and their registration systems are used as a use case to analyze what kinds of geographical information is now stored and what will be needed in order to enclose and manage the georeference data. The inventory of the current metadata elements can be different for any institution, or even a collection. However the required metadata elements will be equal for most intuitions. This can ensure that all data is stored in a unified data format.

3.1 Standards

One of the main purposes of the georeferencing collections in a standardized way is the possibility to exchange and combine data from different collections and institutions. To make this possible, it is important that the details of the georeference data are accessed in the same way (Arkel ed al, 2015). Organizations like the TDWG and GBIF promote the use of standards and support institution in making their data available in 'the same way'. The TDWG, Biodiversity Information Standards**,** is an organization that was brought to life to create and promote the use of standards for the possibility to exchange biodiverse collection data between biodiversity institutions. The Global Biodiversity Information Facility (GBIF) is an organization dedicated to making the world's biodiversity data free and easily accessible via the internet and the TDWG. More than 50 countries publish their data through GBIF, including Naturalis (NBC, 2012). Both these organizations promote the use of Darwin Core and ABCD.

ABCD & Darwin Core

One of the most widely used standards in the international biodiversity community is Darwin Core (DwC). Darwin Core is an extension of the very general Dublin Core (DC) data standard, specifically for biodiversity-related information (GBIF, 2010). Darwin Core provides a vocabulary of terms that make it possible to find information on organisms, to retrieve and to integrate this information, for both organisms and observations in natural history collections.

The Darwin Core standard is a relatively simple standard with a total of some 180 terms, divided into different categories (TDWG Wiki, 2010b). Darwin Core is particularly useful for the core data of specimen collection events. More information can be stored in so-called Darwin Core Extensions, where additional data is linked to the "core data". One of these extension sets is an extension specifically focused on Geospatial material: the Geospatial Element Definitions Extension to Darwin Core (TDWG Wiki, 2010b).

Access to Biological Collections Data (ABCD) is a highly advanced data standard that makes it possible to record a lot of information standardized. It is often used for natural history collection. The standard includes more than 1000 terms and can be scaled for specific information needs. It is not possible or even necessary to use all terms that the ABCD standard provides, since there are so many. ABCD also knows

extensions sets to describe specific data of biodiverse collections more. For example the DNA extension or the EFG extension for geosciences, for use with paleontological, mineralogical and geological digitalized collection data. The use of the standard ABCD often requires specific software for unlocking the data. It is compatible with several other existing data standards, for example Darwin Core (TDWG, 2010 A).

ABCD and Darwin Core can be combined through the use of mappings, to match the elements from both standards so that they complement each other. In addition to the general geographic metadata elements (Higher Geography, Continent, Water Body, Island group, Island, Country, State province, county, locality) the mapping between DwC and ABCD knows the following georeference specific metadata elements (TDWG, 2007):

- ***Decimal latitude***: The latitude of the geographic center of a location where an event occurred (organism collected, observation made), expressed in decimal degrees.
- ***Decimal longitude***: The longitude of the geographic center of a location where an event occurred (organism collected, observation made), expressed in decimal degrees.
- ***Geodetic datum***: The geodetic datum to which the latitude and longitude refer.
- ***Coordinate uncertainty in meters:*** The upper limit of the distance (in meters) from the given DecimalLatitude and DecimalLongitude describing a circle within which the whole of the described locality lies
- ***Point radius spatial fit:*** A measure of how well the circle defined by the coordinates and uncertainty match the original spatial representation, as a ratio of the area of the circle to the area of the original spatial representation.
- ***Verbatim coordinates***: A text representation of the coordinate data
- ***Verbatim latitude:*** A text representation of the Latitude part of the coordinate data from its original source
- ***Verbatim longitude:*** A text representation of the Longitude part of the coordinate data from its original source.
- ***Verbatim coordinate system:*** The name of the system in which the verbatim geographic coordinates were recorded
- ***Georeference protocol***: A reference to the methods used for determining the coordinates and uncertainties
- ***Georeference sources***: A list of maps, gazetteers or other resources used to georeference the locality.
- ***Georeference verification status:*** A categorical description of the extent to which the georeference has been verified to represent the location where the specimen or observation was collected.
- ***Georeference remarks:*** Comments about the spatial description determination, explaining assumptions made in addition or opposition to those formalized in the method referred to in GeoreferenceProtocol.
- ***Footprint WKT:*** A Well-Known Text (WKT: see http://en.wikipedia.org/wiki/Well-known_text) representation of the shape (footprint, geometry) that defines the location of the occurrence.
- ***Footprint spatial fit:*** A measure of how well the geometry expressed in the footprint match the original spatial representation, as a ratio of the area of the footprint given to the area of the original spatial representation.

BioGeomancer workbench guidelines

The BioGeomancer project aimed at the development of a georeferencing tool for biodiverse data managers and was developed by the University of California at Berkeley. The project developed a tool called BioGeomancer, which uses textual location descriptions to find corresponding coordinates (BioGeomancer Consortium, 2006). In addition to the tool, the project also conducted a best practice based on the use of the BioGeomancer tool, for georeferencing biodiverse material. The project was completed in 2012, after which the tool was no longer available, but the best practices are still widely used. In the best practice, the following required metadata fields are recommended (BioGeomancer Consortium, 2006):

- ***Decimal Latitude:*** the latitude coordinate (in decimal degrees) at the center of a circle encompassing the whole of a specific locality. Decimal latitudes north of the equator are positive numbers less than or equal to 90, and those south are negative numbers greater or equal to −90
- ***Decimal Longitude:*** the longitude coordinate (in decimal degrees) at the center of a circle encompassing the whole of a specific locality. Decimal longitudes east of the Greenwich Meridian are considered positive and less than or equal to 180, while western longitudes are negative and greater than or equal to −180.
- ***Geodetic Datum:*** a model of the earth used for geodetic calculations. A geodetic datum describes the size, shape, origin, and orientation of a coordinate system for mapping the surface of the earth
- ***Maximum Uncertainty Estimate***: The upper limit of the distance from the given latitude and longitude describing a circle within which the whole of the described locality must lie.
- Maximum Uncertainty Unit: The unit of length in which the maximum uncertainty is recorded
- ***Verbatim Coordinates:*** The original (verbatim) coordinates of the raw data before any transformations were carried out.
- ***Verbatim Coordinate System:*** The coordinate system in which the raw data were recorded.
- ***Georeference Verification Status :*** A categorical description of the extent to which the georeference and uncertainty have been verified to represent the location and uncertainty for where the specimen or observation was collected
- ***Georeference Validation:*** Shows what validation procedures have been conducted on the georeferences – for example various outlier detection procedures, revisits to the location, etc.
- ***Georeference Protocol:*** A reference to the method(s) used for determining the coordinates and uncertainty estimates
- ***Georeference Sources***: The reference source (e.g., the specific map, gazetteer, or software) used to determine the coordinates and uncertainties
- ***Spatial Fit:*** A measure of how well the geometric representation matches the original spatial representation and is reported as the ratio of the area of the presented geometry to the area of the original spatial representation
- ***Georeference Determined By***: The person or organization making the coordinate and uncertainty determination
- ***Georeference Determined Date***: The date on which the determination was made.
- ***Georeference Remark***: Comments on methods and assumptions used in determining coordinates or uncertainties

3.2 Use Case: geographic metadata in the Naturalis collection

The objectives of the FES-digitization project, a project that would digitize 7 million objects and 30 million storage units from the collection of Naturalis during 5 years, required a process of industrialized digitization. For this, standardized processes were required with a focus on similarities, not differences. For that reason it was decided to register objects in accordance with the standard set of elements from the Biodiversity Information Standards, using the ABCD (Access to Biological Collections Data) and Darwin Core standards (NBC, 2015 b). The collections of Naturalis uses its own set of elements for geographic metadata, which partly corresponds to the set of ABCD and DwC to enclose the object records in the CRS and Brahms. The data model that is used in the Naturalis- collection has a number of additional elements, to the standard set, from the ABCD and DwC mapping.

3.2.1 Metadata elements in the database of Naturalis

All geographic metadata in the CRS come together under the element 'Gathering site'. This field is made up of 33 metadata elements, including the general geographic metadata elements, like water body, country and metadata elements for georeference data (NBC, n.d. B). In the diagram below, all metadata elements covering geographic data of the larger element 'Gathering site', are listed, including the definition that is associated with this field. This definition is used as a guideline for filling out the metadata elements (NBC, n.d. B).

***CRS - Gathering site**:*

- ***Geographic place names:** place name and place type combined*
  - ***Place name**: Name of the gathering site (a geographic*
  - ***Place type**: Classification categories for the class of the gathering named site (local or national subdivision levels*
- ***Bioregion:** Describe the bioregion in which the specimen was collected or observed.*
- ***Country:** Name of country or major region in which the specimen was collected or observed*
- ***State provinces**: Field to indicate the state or province in which the specimen was collected or observed.*
- ***Island:** Island name in which the specimen was collected or observed*
- ***Locality:** Field to indicate the locality full name were the specimen was collected or observed. Additional notes e.g. 10 Km South of .. must be placed in Full locality text.*
- ***Station number**: Field to indicate the station number of the locality were the specimen was collected or observed.*
- ***Full locality**: Original locality data as appearing on a label or in an original entry*
- ***Coordinates:** all following fields combined:*
  - ***Index:** In case an area is described an index number is added to the data.*
  - ***Methods**: Coordinates measuring system e.g. GPS*
  - ***Verbatim coordinates:** Original coordinate data as appearing on a label or in an original entry*
  - ***Accuracy:** Textual statement of degree of accuracy*
  - ***Distance in meters**: An estimate of how tightly the collecting locality was specified: expressed as a distance in meters corresponding to a radius around the coordinates*
  - ***Lat long***
    - ***Lat:** Latitude in decimal degrees*
    - ***Long**: Longitude in decimal degrees*
    - ***System**: Mathematical surface on which the mapping and coordinate system used for the geocodes of the record are based*
  - ***UTM***
    - ***Verbatim UTM**: Verbatim concatenated text representation of UTM coordinates*
    - ***Zone**: The numerical zone corresponding to the central meridian and origin upon which the UTM Easting is based*
    - ***Subzone**: The subzone letter corresponding to one of the 20 North-South divisions of the UTM grid system*
    - ***NS**: The hemisphere to which the UTM Northing refers (North or South)*
    - ***Easting**: The distance in meters east of the origin of the UTMZone*
    - ***Northing**: The distance in meters north of the origin for the UTMSubzone. For the northern hemisphere (UTMSubzones M through X) the origin is at the equator. For the southern hemisphere (Subzones C through L)*
    - ***Datum**: Mathematical surface on which the mapping and coordinate system used for the geocodes of the record are based.*
  - ***Grid***
    - ***Cell system**: The name of the grid system used for the gathering site coordinates*
    - ***Cell code**: The code of the grid system used to record the gathering site coordinates*
    - ***Qualifier:** A grid reference precision qualifier for the gathering site coordinates*
    - ***Latitude**: Latitude*
    - ***Longitude**: Longitude*

In the Brahms, the collection management system for botanical data, there are fewer fields to describe the geographical object information: namely 13. For a selection of the objects the botanical collection in the Brahms, there is coordinate information available this concerns about 10 % of the entire botanical collection. In the diagram below, all metadata elements covering geographic data in Brahms, are explained, including the definition that is associated with this field.

Brahms:

- ***Country:*** *chose from the look up list: 'ISO Countries or Oceans'*
- ***Major country area****: highest-order subdivision of the country' political or natural unit*
- ***Minor country area****: second highest-order subdivision of the country', like 'State/province (by Getty Geographical Thesaurus)*
- ***Locality:*** *Third order subdivision of the country. This includes city's and smaller, like a park, lake, bedding or river*
- ***Locality notes****: additions to the locality to further describe the geographic location of the object*
- ***LLunit:*** *type of notation used for the geographic coordinates, use DD-notation.*
- ***Latt:*** *Latitude: this depends on the notation used (LLUNIT), hereby N is positive and S is negative*
- ***NS:*** *North or South of the equator*
- ***Long****: Longitude: this depends on the notation used (LLUNIT), hereby E is positive and W is negative*
- ***Ew****: East or West of the Greenwich meridian*
- ***LLRES****: The accuracy of the coordinates on a scale of 1° tot 0.01'*
- ***LLorig****: Required if coordinates are on the label. In that case, fill in "sheet" or blank if there no coordinates*
- ***LLdatum****: Mathematical surface on which the mapping and coordinate system used for the geocodes of the record are based.*

3.2.2 Geographic data in the collection of Naturalis

During the FES Collection Digitization Project 8.5 million specimens objects were digitized at object level and the remaining 30 million on storage unit level (shelf, box or drawer). The information associated with the objects (such as taxonomy, collection date and collection site) was put in two central databases: one for zoological and geological data, the Collection Registration System (CRS) and one for botanical data, the Botanical Research And Herbarium Management System (Brahms) (NBC, 2015 b).

In advance of digitizing the collection there guidelines were made for the desired depth and the level of registration of the metadata of objects, following the nature, history and use of that material (NBC, 2013). A distinction was made between three levels of the depth of the registration: the 'minimum required metadata', the 'desired additional metadata', and 'if feasible metadata'. The geographic metadata belongs to the first group, and is thus recorded in the digital collection.

Not all geographic metadata fields can be filled out for all objects: and the fields that have been filled in contain major differences between the metadata. For example, the fields may contain data in text form, but also coordinates and other formats such as the Dutch Grid References, or descriptions of places, exist in the collection. For the majority of the metadata fields, there is verbatim data recorded as well as a thesaurus match to correspond with this.

During the FES digitization project it was also agreed that the people who registered the objects that: information written in the past by curators on object- tags or labels, were not to be interpreted but were recorded as verbatim in the CRS or BRAHMS (NBC, 2013). For example: there are many personal preferences for ways to spell a name scattered through the collection, such as the ways 'Den Haag' can be quoted:. '*Hage*', '*s-Gravenhage*', '*The Hague*', '*in den Hage', 'Den Haag', 'Schravenhaegen', 's Gravenhaege', 'Gravenhage'*. The spelling is entered unchanged and then linked to a thesaurus with standardized values.

In this way, the verbatim transcription preserves any historical names of the original object-label, but the modern names (with the notation) can be used for geographical research (Arkel ed al, 2015). For example, objects of which the verbatim locality is "s' Hage" can be linked to the standardized term "Gravenhage". The Getty Thesaurus is the geographical thesaurus used by the CRS as a leading thesaurus for geographic data (NBC, 2013).

The CRS and BRAHMS contain the metadata fields ‘Locality' and 'Full locality' that give the opportunity to further specify the described locations of the gather site, with a heading (north or south) an extent distance or a degree of elevation etc. These fields are free text fields and therefore often contain very detailed information, which can vary greatly in form, language, or length (NBC, 2013). Below you can find some examples of verbatim ‘locality’ and ‘full locality’ information from both the CRS and from the Brahms:

- *“MUNNEKEZIJL: LAUWERSMEER: KOLLUMERWAARD: BREDE SLOOT 05.27. AC 211.735-543.209“*
- *“HORST-AMERICA: SAAR BIJ SPOORLIJN“*
- *“Gem. Terneuzen Nederland (Z.) Braakman N (MV 1) Westgeul, ES 5186“*
- *“ES 4491 Nederland (Z), (MV) Inlaag Hoofdplaat West Langeweg“*
- *“Noord- Thailand“*
- *“NEDERLAND (N.H.) IJsselmeeroever bij Broekerhaven 2 km Z.W.v.Enkhuizen“*
- *“Nederland (Z), (MV) Kuststrook Cadzand t.o.camp. Hoogduin“*
- *“Z.-Spanje, prov. Malaga. Langs pad“*
- *“wegrand op half-open grond, Kr. wit Pfafleralp (Löhuchental)“*
- *“Prov.: Antalya. Lar (± 10 S.E. of Antalya)“*
- *“Fron an evident by bind swarm in areble land on the lower of River Copse, above Inkpen, Berks“*
- *“Lusitanica, Rotswand bij Penacora”*

Georeference tools, like the Google Geocoding API, find it hard to recognize these descriptions (Arkel ed al, 2015). For example, such a tool can see the numbers as addresses or postal codes instead of distances, but these additions give the opportunity to georeference object with a better accuracy. For example: if the locality description is “*2 km Z.W.v.Enkhuizen”*, the Google API can georeference the city of Enkhuizen, but leaves out the 2 km ZW, which means the coordinates are 2 km of from the actual collection site.

A similar variety of data is also visible on the already known coordinates. Today, researchers are often going into the field with a GPS tool that directly registers the GPS data of new incoming objects (NBC, 2013). But errors can easily occur in this process:

- The plus and minus of the coordinates are reversed, making the location description and coordinate data not compatible
- There are differences in the number of decimals that get recorded, making the accuracy of the specimen collection differ from each other
- Accuracy of the GPS device can differ from the decimals recorded
- "0.00 / 0.00" is entered when the coordinates are unknown. Resulting in wrongly matched locations and coordinates.
- The coordinate system or datum sometimes not recorded or the wrong datum or system is recorded. Making it harder to use with other spatial data, of which the datum or system is different.
- The latitude and the Longitude are reversed, making the location description and coordinate data not compatible

3.3 Conclusion

Geographic information regarding species occurrence data can be divided into two sets of elements. The first sets of geographic data elements regards the original locality description, as written on the label, in the field notebook or in the CMS. This set usually contains fields like: Country, State provinces, Island, locality, Station number, Full locality etc. These metadata elements are already present in most biodiversity collections, and are therefore not part of this best practice. It is however important to mention: that georeference data must be added to existing records, so that the existing data (the data as written on the label or in the field notebook) is not over written. This is important information and has historic value and should therefore never be overwritten by standardized data.

The second set of geographic data elements are those actually describing the georeference data and process. What is the best way to record this in a collection management system, so that it can be preserved in a uniform format, was the question for this part of the research? To answer this I looked at different

biodiversity metadata standards, ABCD and Darwin Core, and a best practice by Chapman, originating from the BioGeomancer Workbench, since exchange and corporation are very important aspects of data storage and digitization in the biodiverse domain.

When you compare the standards of ABCD and Darwin Core, it becomes clear that there are many fields that are occur in all standards and are therefore probably highly valuable. For this research these are considered the minimum required (primary) metadata field for the recording georeference data in a biodiverse collection.

In addition to the primary fields, there are some fields, which occur in multiple standards, but differ in form a little such as a different definition or a different name (secondary). These fields can add value to the georeference data and contribute to a standard method for enclosing it, but the precise definition, is institution specific. This can be adjusted based on the original standard that is used, the project or the functionalities of a CMS. The secondary date elements also contain some fields that describe the georeferencing process. These can be included in the collection management system, but will primarily contribute to the documentation of the development of the collection and to the reliability of the georeferencing process, and less for the exchange of geographical information (Chapman, 2006). Whether or not this is recorded depends fully on the content policy of an institution.

The following table shows which of the metadata fields from the standards are considered as primary and secondary fields. The fieldnames and description that appear in table are derived from the ABCD standard element Shema. For institutions that prefer to use the Darwin Core standard, the mapping found here: http://rs.tdwg.org/dwc/terms/history/dwctoabcd/, can be used.

Primary metadata elements	
Fieldname per standard	**Description/definition**
LatitudeDecimal	The latitude of the geographic center of the locality expressed in decimal degrees.
LongitudeDecimal	The longitude of the geographic center of the locality expressed in decimal degrees.
SpatialDatum	The geodetic datum to which the latitude and longitude refer.
CoordinatesText or Coordinates UTM/UTMText	A text representation of the coordinates. This can be one element for latitude and longitude, but can also be separate elements
CoordinatesGrid/GridCellSystem	The name of the system in which the verbatim geographic coordinates were recorded.
CoordinateMethod	A reference to the methods used for determining the coordinates and uncertainties.
coordinateErrorDistanceInMeters or AccuracyStatement	A measure of the area in which the described locality must lie. Usually expressed by the distance from the coordinates to the upper limit/outer corners of the polygon/radius, in meters. A free text statement of the degree of accuracy of the latitude and longitude coordinates.
Secondary metadata elements	
Fieldname	**Description/definition**
FootprintSpatialFit	The overlap between the actual locality and the georeference polygon/ point
GeoreferenceSources	A list of maps, gazetteers or other resources used to georeference the locality.
GeoreferenceVerificationStatus	The status of the georeference data; “What validation steps have been conducted?”
GeoreferenceRemarks	Comments about the recorded georeference data
Georeferenced By	The person or organization making the coordinate and uncertainty determination.
Georeferenced Date	The date on which the determination was made.

The last two fields in the table above are derived from the field set from the Naturalis collection (use case). The set that is used in the CRS of Naturalis complies with the standards, but is much more extensive. It contains most of the minimal required data fields and is therefore an option to use this as an example on how to make the standards comply with your CMS.

4. Information needed by users of georeference data

The quality of data is a concept that has many definitions, and often depends on the purpose and the origin of the data. In the geographical world one definition is widely accepted: fitness for use (or potential use) (Chrisman 1991). Which states that data quality is strongly related to use and thereby also to the users. The Fitness for use cannot be determined without analyzing the users and their goals. (Chapman, 2005). For biodiversity collections this means that the value of geographic data is centered on improving / enriching a collection so that it is useful for research, collection management and accessibility of the collection.

For example If you are conducting a research on the spreading of a bird specimen in the northern part of Europe, the precision of a georeference code can be quite large, like 20 km. But if you want to conduct a research on the spreading of that specimen in the northern part of Norway, you will want a georeference code that has a smaller precision, like 5 km. If the available dataset has an accuracy of 10 km, it is fit for use for the first research, but not fit for use for the second research.

Use case
Naturalis, in this case the users of their collection, will again be approached as a use case in this chapter, but the results will be classified by three general fields of application that benefit from enriching the geographical information: collection management, research and accessibility. These are fields of application that will occur in most taxonomic/biodiversity institutions with a collection.

Within the collection of Naturalis there are several groups that benefit from enriching the collection with georeference data. Most of these users currently make use of the currently available geographic data (textual location descriptions) in the collection, but this is often not sufficient enough. Geographical reference points and additional metadata can help these groups to perform their work better and more efficient and provide new opportunities.

To determine the information needs of these three areas interviews were held with the users of geographic data in the collections of Naturalis, from each field: collection managers, in-house researchers, BioPortal developers and a Wikipedian in residence. This way it was possible to determine at what levels or how the current geographical data collection should be enriched, to assist in their work. The information gathered from these interviews was converted to a summary of the information needs: wishes and requirements that are to be met by the georeference data. Based on this information, a tool/method will be selected that fits these need best.

4.1 Collection Management

The collection management is responsible for the management, conservation and documentation of the collection. Since the digitization project of last year, the management, conservation and documentation of digital collection has become a larger part of their work activities, since the digital collection has grown enormously. Additionally, they support scientific research, by answering questions about the collection material. The metadata, including the geographical data, associated with the collection specimen is therefore used and consulted by collection management.

With digitization there is always a battle between quality and quantity, and for collection management digitization mainly centers primarily on quality (Personal communication Luc Willemse 2015). The big push behind digitization is to make the collection more accessible, but for collection management this also includes placing and identifying (determination) the object specimen properly in the collection. The georeferencing of objects is an aspect that is also a part of it, but in both parts the reliability of the data is the most important (Personal communication Luc Willemse 2015). Coordinates can certainly contribute greatly to the work of collection management, provided that they are correct.

Secondly with the addition of uniform coordinates the collection could be browsed based on geographical data. The collection can for example be sorted by certain geographical areas. For the collection management this could save allot of time to answer the questions of in-house researchers and in some cases it can become easier to gather the required data collection (Personal communication Luc Willemse, 2015).

4.2 Research

Precise geographic information is essential for harvesting collections for research on biodiversity. The taxonomic data (such as a species name), and the coordinates of objects are the main types of information needed by a researcher to research the distribution and occurrence of species. This is also the most widely used data of the collection (Personal communication Luc Willemse 2015).

During the digitizing project the geographic data there is copied verbatim from labels and tags, to maintain historical data. The interpretation of the verbatim data takes place outside the CRS. Whilst digitizing the verbatim data a researcher can link a thesaurus term to the verbatim text. The data that is recorded verbatim gives some direction but a coordinate is much more useful for research. In some cases there may also be a verbatim coordinate recorded in the collection. For these cases it is also necessary to check if any interpretation or georeferencing is necessary.

Linking the corresponding coordinates to the verbatim location descriptions can increase the usability and accessibility of the collection. With coordinates you can directly plot species on maps, research the changes of the occurrence of the species and combine it with other data such as climatic or environmental data to examine the impact of this species.

The most important aspect of georeferenced data in a Biodiversity collection is the accuracy of the coordinates. Any kind of research has a different need for a degree of accuracy of the georeference points. However, this accuracy is jeopardized by various difficulties occurring in the textual location descriptions:

- **Missing information**: For some species, only the country of the collecting event is known
- **Misspellings:** because the labels are transcribed verbatim during digitizing project, the spelling mistakes that were on the labels are also adopted in the digital collection.
- **Historical names:** Some collections are up to 200 years old and contain names of sites that no longer exist in modern gazetteers.
- **Changes in the landscape:** because of changes of river banks, the growth of cities and the disappearance of settlements, it can be difficult to find some locality descriptions on modern maps.
- **Unclear descriptions:** some descriptions consist of only a direction, distance, or a landmark, making it difficult to find a precise location.
- **Location descriptions of data poor areas:** where few name-bearing landmarks exist, such as in tropical areas.

Regarding an area of 10 to 10 km around an object is accurate enough for most types of research and feasible with the geographic information available in the main part of the collection of Naturalis (Personal communication Luc Willemse, 2015). There are, of course, situations where this in not sufficient: for example:

If an object is found in a mountainous area, a georeference code with an accuracy of 10 km is not sufficient, because 10 km in such areas can mean the difference between a mountaintop and a mountain valley. For ecological conditions, this is a very big difference, making the species that occur there also significantly different.

According to both collection management and the researchers is interviewed a good solution to the accuracy problem is to give an indication of what the accuracy of a coordinate is (Personal communication Luc Willemse, 2015). This must be done in a way that is understandable and unambiguous for each user, so that there no pretense of a small accuracy occurs. If, for example a rating scale from 1 to 10 is used to indicate the accuracy of a coordinate, you pair a subjective value judgment to the coordinates, which can have a

different meaning for each user (Personal communication Luc Willemse 2015). The most objective way, is to indicate an accuracy in meters, and where possible also to indicate what this is based on (Personal communication Niels Reas, 2015). Based on this information users can decide whether or not the coordinates are usable enough for their research goals.

4.3 Accessibility

The accessibility of collections is important to support scientific research as much as possible and to make the collection visible and searchable. Accessibility and searchability of the collection is important not only for internal purposes (such as collection management and research), but also for the wider public and schools (Personal communication Ruud Altenburg, 2015).

Parts of the collection of Naturalis is showcased to the public in the museum, but since the digitization of the collection, you can also bring it to the public digitally. This will increase the audience of collection data from a select group of specialists to the whole community. Geographic information is very valuable fort his objective. Where the collection is often arranged systematically (taxonomically) for internal purposes, it is accessible for the public if objects can be viewed on a map, or can be searched by area.

The digitized collections of Naturalis are accessible through various channels for the wider public: the BioPortal, Gbif, Wikipedia, Europeana and various apps. These products also have an interest in high quality geographic and georeference data.

The BioPortal

The BioPortal at Naturalis is a channel that makes the collection accessible and searchable via the web through an API, for anyone who is interested. At this moment it is possible to search the collections in the BioPortal based on geographical data, but this only concerns the collections that have useful geographic data: coordinates. In the geographic search system you can by country, city, or area search for specific species, but you can also select an area on a map. In this respect, this is an addition to CRS, where this is not possible, and it is very useful to the wider audience, and for internal use (Personal communication Ruud Altenburg, 2015). The BioPortal only displays object that contain latitude and longitude coordinates, so if there are more collection that are enriched with coordinates, there more collections that can be accessed by the BioPortal (Personal communication Ruud Altenburg, 2015).

For the users of the BioPortal it is important that an indication of the accuracy gets marked in the collection management system. The latitude and longitude of an object can be very accurately described in degrees, minutes and seconds, and thereby give the impression that they are very accurate, even though this is not so. This can be misleading for users, who do not know what the coordinates are based on. As a result, there is the possibility that a pretense of the accuracy occurs. By using an additional metadata field for the accuracy of the coordinates, this can be made understandable for all users of the data.

Wikipedia

Naturalis started a project in early 2015 to make a part of the collection of images and other multimedia, such as videos, 3D models or sound recordings accessible through Wikimedia for the wider public. In addition to the multimedia files, they also donate the corresponding metadata of the species which can be found in the images or video's, such as the taxonomic data and the geographic descriptions of the distribution of the species. This includes geographic coordinates. For multimedia this concerns the coordinates of the species visible on the multimedia object and of the location where the image of video was taken/recorded (Personal communication Hans Muller, 2015).

For species that are on the red list of the IUCN, the geographic metadata to gets left out at some level (NBC, 2013). With these species, only the country and possibly the city or region is indicated (Personal communication Hans Muller, 2015). With this particular situation Naturalis deliberately chooses for a bad accuracy to protect the concerning species. This complies with the Open content policy of Naturalis and the guide to best practice for generalizing sensitive species occurrence data, by GBIF. This last document is an international best practice discussing the sensitivity of the exact localities of rare, endangered species or species of commercial value and how to deal with this data (Chapman, Grafton, 2008).

GBIF

The Global Biodiversity Information Facility (GBIF) is an informational open data infrastructure for digital biodiverse material. The organizations works with and helps biodiversity institutions worldwide to share their collection data for the use of research. By promoting international standards like Darwin Core and ABCD, it improves the possibility of combining biodiverse material from different collections and with other disciplines. Naturalis is one of the 50 institutions that publish their data through the GBIF portal. At this moment 6.5 million objects, and at the end of 2015 7 million, of the Naturalis collection will be available on through the GBIF infrastructure.

The Naturalis data on GBIF can be used by researchers and institutions worldwide and can be combined with data from other institutions that place their data in GBIF. Researchers can view the collection data individually, in datasets or as a total. Every dataset includes a creative common license.

Naturalis has one representative working together with GBIF, and the national node NLBIF, to enhance the quality of the data that Naturalis uploads to GBIF and to check if the data complies with the standards that GBIF advices.

4.4 Conclusion

Fitness for use is the most important aspect when it comes to collecting georeference data for a biodiversity collection. For biodiversity collections this means that the value of geographical data is centered on the information needs of the application fields: research, collection management and accessibility.

Within biodiversity collections there are several application fields that benefit from the enrichment of collections with coordinates. These fields in some cases use the currently available textual verbatim geographic data for their work, but this is not always sufficient. For all these application fields it means that there is a clear need for geographical coordinates as reference points, as these are more durable and less variable. Coordinate systems/ datum's don't change very often, and when you have a coordinate it will (almost) always refer to the same place, regarding the map or tool that is used. Another important advantage of coordinates is that they offer opportunities to conduct activities more efficient / better.

To make georeferenced data fit for use for the mentioned application fields, it is necessary to record more than the coordinates in the collection management systems, like: uncertainty, coordinate system, datum, coordinate precision etc. The most important additional metadata field is the 'accuracy' of the coordinates. Accuracies and Inaccuracies highlight the limitations of the application of the data. These limitations determine the ultimate quality of the data, and thus the quality of the results, and how a researcher or collection manager should interpret them. So by recording this in the collection it can be made clear for what purposes the data is fit for use.

For example, an accuracy of 10 km is sufficient for the majority of the researches carried out for Naturalis. Of course this will not cover all types of research conducted at an institution like Naturalis, but sometimes it is just simply not possible to find more accurate results based on the original data.

However, the most important thing is that the accuracy of the coordinates has to be indicated clearly and without the possibility of misinterpretation. This could be done, for example, with a field such as *'uncertainty in meters*', '*the maximum error distance'* or '*maximum uncertainty'*. Based on this researchers or other users can determine whether or not the coordinates are useful for his or her goal. Defining the measurement of accuracy is best done in in measurable units, like meters that is understandable and unambiguous for each user. This way there is no possibility for misinterpretation or false precision.

Georeferenced data makes the collection accessible based of geographic information, for internal use and for the general public, making it accessible to users without knowledge of taxonomy, to browse the collection. For geographic data from species the user group is much larger than the user group of taxonomical data, such as educational institutions, amateur associations, hobbyists, and in many cases even the entire public. For this the coordinates, and original location description is enough. It is therefore again important that the original geographic data and the new georeference data exist complementary to each other, and that the georeference data does not overwrite the other one.

5. Georeferencing projects

During the Semi-Automated Georeferencing Meeting 2nd - 3rd December, 2014, detailed information was exchanged on the (technical) methods the NHM, RMCA and Naturalis have developed or tested for automatically georeferencing their collection objects.

The meeting revealed that NHM, RMCA and Naturalis have similar and complementary competencies and use standardized GIS tools that fit well together (NBC, 2014b). the expectation arose that there could be more institutions within the CETAF that executed georeferencing projects and used tools or methods that could complement the discussed methods form the NHM, RMCA and Naturalis.

There was a common need, amongst the institutions present at the meeting, for a study to explore the next question: until what point in the process is it possible to georeference collection objects in a (semi) automatic way and with what tools? To answer this main question, I interviewed different international biodiversity institutions that had experience with georeferencing, from their own projects. In the interviews is spoke to representatives from the institutions about recent projects they did, the tools/ methods they used, how they worked, what way's they used to reduce the amount of man-hours needed for the project and how accurate and reliable the created georeference data was. All of the institutions I interviewed, I spoke via Skype or Google Hangouts. Some of the interview partners send me a few links, or documents afterwards, and some even showed me, trough screen share on Google Hangouts how the tool they used, works.

In this chapter you can find a description of the 7 projects of all 6 institutions I interviewed for this part of the research.

5.1 FCD-Pilot Georeferencing, Productiematig Georeferencen

Naturalis Biodiversity Center

In this project a process is developed based on Google Maps (Google Geocoding API) for semi-automated georeferencing, to see what the possibilities are for matching locality descriptions to corresponding coordinates in a largely automatic fashion. During this process a large batch of location data is offered to Google Geocoding API, via Open Refine, and comes back with detailed geo-referenced data. The process had to meet the ''*Principles for Best Practice for georeferencing Biological Species Data"* (Chapman, Wieczorek, 2006).

Semi- Automated georeferencing via Google Maps (Google Geocoding API) is the only georeference method currently available that provides the ability to georeferencen large amounts of data in batches. This is a necessary part of the process, considering the size of the collection.

During the pilot, two datasets were used to test the georeference methods: the Hymenoptera dataset object with 18 227 records from the bumblebee- collection and Chiroptera dataset with 23 704 records from the bat-collection. The object records were exported by the department Information Management from BRD and CRS to an Excel file. The georeferencing process was applied to the exported files.

Google Maps can be used to find latitude and longitude of a location, and display these on a map. The results can, however, also be returned as text to the user. This way the found georeference data can be exported and added to the collection database.

The Google Geocoding API communicates with Google Maps by an 'HTTP' request (URL) that can be put back in the search bar of an Internet browser. The API returns the results in the form of a string. To offer

large amounts of geographic data with HTTP requests automatically to the Google Geocoding API and after running it collecting and adding the results to the original file, the program Open Refine[1] is used.

The results of the semi-automated georeference method with the Google Geocoding API are promising: 39% of the tested object records from the Hymenoptera and 50% of the Chiroptera object records gets a coordinate after a first run of the data through Google Maps. For the Hymenoptera records the percentage found coordinates rises from 39% to 90%, when bad/not found results are reprocessed by Google Maps, after data cleaning. Records that after this second run are not yet matched with coordinates could be manually matched by an expert of georeference data or the specimen.

To increase the number of results with a full match by Google Geocoding API improving data cleaning was required. This project looked at how the geographic data of the datasets can be improved so that it generates more and more reliable georeference results. A distinction is made between light labor-intensive, heavily labor intensive, very heavy labor intensive and software based data cleaning.

Light labor-intensive data cleaning could mean three different forms that are applied before an initial run Google Maps, like:

- adapting signs that were copied incorrectly from the registration system to the export file,
- removing offset number from the locality descriptions, because google can see the as a part of an address (like a house number or postal code).
- Deleting double location names. This step reduces the amount of unique locations and reduces the chance of double matches for a locality description.

Heavy labor-intensive data cleaning concerns the mapping and improving of ambiguous geographical data notations that could generate wrong coordinate's matches or 'unfound matches'. This data cleaning process takes a lot of time.

An efficient way to remove data that does not provide a response (such as "West" Or "loft") is a visual data cleaning based on Google Maps. The strings in Excel are changed by a formula into links which can visually show what the outcome of the string is in Google Maps. By making manual adjustments in the search strings in Google Maps, you can visually search for a string that gives a good result.

A first run through Google Maps has been delivering 40-50% results and currently costs 30 man hours for 10,000 object records. An investment of 10 extra man-hours in order to carry out data cleaning, including a second run by Google Maps, delivers an increase of 40 to 90% for the Hymenoptera dataset. With the current method, an average of 250 records per man-hour, including data cleaning, can be provided with georeferentiedata. This is 40,000 records per month.

It is expected that after further optimization and automation of the method described here, about 5 hours less are needed to run 10,000 records through the whole process.

5.2 SPECimap Georeferencing Software

Royal Botanical Garden of Edinburgh

In collaboration with the School of Informatics at the University of Edinburgh, the Royal Botanical Garden of Edinburg developed its own georeferencing tool: called SPECimap that allows users to curate botanical specimen objects by adding or complementing automatically generated georeference data to existing geographical locations of the collection. The tool automatically plots textual and numerical (coordinates) localities on a map, but requires manual input to further create accurate georeference codes. Therefore SPECimap combines the accuracy of manual annotation with the efficiency of automatic processing.

Process
The tool takes certain field from the collection database and uses text strings from these fields for text mining

[1] http://openrefine.org/

on a known gazetteer like the Gazetteer of British Place Names[2] and the Fuzzy Gazetteer[3]. Based on this the tool plots the localities that were found with text mining on a map. The tool than gives back different types of geographic metadata, like latitude and longitude, country, ALTM, National Grid reference, region, habitat. the found coordinates are shown underneath the map on which the object is plotted. This can be adopted in the collection data. This output can then be corrected or complemented in order to specify and extract georeference data of where the object was collected.

The map that the tool uses is based on Google maps[4], but is combined with different historical maps (old survey maps) and the UK National Grid reference. These maps are layered on top of each other and with the help of an API, created by the Botanical garden of Edinburgh; you can move through the different layers of maps and fade them out. This way you can see the changes in costal arrears or city, and place the plotted locality more accurately.

For example
If you have a plant specimen that *was found in 1750 on the northern borders of the city of Edinburgh, you can pin this object on the current border of that city with the help of a modern map. This may seem very accurate, but over the last 250 years, the city of Edinburgh (just like most cities) grew in size. So the plotted locality will likely be less accurate than initially thought or even be plotted in a completely different grid number. By using the SPECimap tool to view the plotted locality on a map that shows the city of Edinburgh in the 18th century, you can see where the border was in that time and replace* the plant specimen *to where it was collected and create more accurate results.*

SPECimap can be used for georeferencing with 3 different objectives, because the localities can be plotted based on the textual locality description, national grid numbers and already known coordinates. Regarding the input it the tool generates all the remaining metadata elements. The tool works very well for georeferencing object specimen that have difficult locality descriptions. For example botanical samples from legacy collections with historic place names or samples from difficult geographic locations, like mountains or coastal areas where a small distance could make a great difference in ecological circumstances.

The second way to use the tool is to check objects that already contain coordinates, created with a different georeferencing tool. The tool delivers very accurate results, mostly with an accuracy of 5 km, but because it requires a lot of manual work and it is very time consuming it not suited to georeference big batches of object specimen, but it can be used as an extra quality check or to fine-tune georeference data that was collected with another tool/method. This could be done, for botanical samples that delivered questionable coordinates, or multiple possible locality matches, derived from the data in the original collection.

Because the tool delivers such accurate results, it could also be used to georeference collections, or parts of collections, that contain specimen samples that require very accurate results to be usable for research. For example botanical samples with occurrence data in geographic localities with very specific ecological circumstances, like mountains of coastal areas.

Even though the tool can create very accurate georeference data it is still very important to mention the accuracy in the collections database and public website. Mainly because the accuracy between specimens that were georeferenced with this tool can differ greatly from specimen that were georeferenced with tools like GeoLocate of Google Maps geocoding API.

Currently the tool is almost ready to be used. After some tests there are still some small bugs that need to be fixed, before the RBGE starts a two month project for georeferencing parts of the collection, using SPECimap. Once the tool works properly the tool will be available as open source.

[2] http://www.gazetteer.org.uk/
[3] http://isodp.hof-university.de/fuzzyg/query/
[4] http://www.google.com/maps

5.3 StanDAP-Herb - a standardized and optimized process for data acquisition from digital images of herbarium specimens

The Botanical Garden and Botanical Museum, Berlin-Dahlem

Herbarium sheets usually contain labels with handwritten or type machined metadata glued to the sheet. These labels contain important data like plant name, collector, occurrence data etc., which is vitally imported for preserving the collection and using it for research. When the herbarium sheets are digitized the data on the labels becomes visible on the images taken of the objects, but not available as machine readable data. This is the ultimate goals of the digitization.

Extracting this data so that it is machine readable is a very time consuming task, since most botanical collections contain are very large amount of herbarium sheets, like the collection of the Botanical Garden and Botanical Museum of Berlin that contains about 22 million herbarium specimens.

The Berlin-Dahlem Botanical Garden and Botanical Museum started a project called StanDap- Herb[5]. The project aims to develop a way to extract information on labels automatically and with that replace the very labor- intensive manual process that is used in most institutions. For the project the Botanical garden and Botanical Museum of Berlin specifically choose to use herbarium sheets because they have very similar information and labels in every collection.

The StanDAP-Herb Project is developing a standard process for (semi-) automatic detection of meta-data on the Herbarium sheets. The aim of the project is to create a workflow to extract this type of data and connect different tools or methods to this workflow that can be used to verify data quality, facilitate data discovery and enhance the application of collection data in research (GitHub, 2011). The goal of this extracting is to eventually enrich the data and to complete the collection.

The project uses existing software, and evaluates and enhances it, to comply with standard interfaces and integrates it into open software architecture, hosted at GitHub. At this moment, the project is still under development, and only in the beginning stages, so there is nothing available yet, but when finished this is the intention of the project.

Object detection software detects objects such as labels or barcodes on the scans of the herbarium sheets and classifies them in different categories like taxonomic, collection or geographic information. The text objects are transformed into structured information with the use of text mining algorithms. Many botanical collections contain historical objects that only contain handwritten data, for this the workflow attempts to use author identification and handwriting recognition to classify data on the labels

A part of herbarium collections contain geographic information, like occurrence data, on the labels of the herbarium sheets. When the information is extracted using the StanDAP-Herb workflow tools this information is also detected and classified by the text mining algorithms/ software.

Georeferencing this extracted geographic data creates a more uniform data standard like coordinates and enhances the application possibilities of a herbarium collection. This is an important part of enriching the extracted data and expanding the collection to create more usable data for research.

The project StanDap- Herb does not actually build or use a georeferencing tool yet, but it wants to include links of tools and methods in the workflow, to completely cover the extraction and enrichment process. Currently the focus of the project on a suited georeferencing to lies on GeoLocate[6], because this fits best within the workflow.

Since the project only started a short year ago, and will take 3 years to complete, there are no current test results for the georeferencing step available.

[5] http://standap-herb.server.de/servlet/is/11/

[6] http://www.museum.tulane.edu/geolocate/

5.4 HerpNET & SYNTHESYS NA-D 3.7 "Itinerary" project,
Royal Museum of Central Africa, Tervuren

The last years the Royal Museum of Central Africa in Tervuren has done several georeference(-like) projects like the EDIT Cyber Taxonomy Platform, HerpNET and most recently they are participating in EU-BON. In these projects they have used several tools and gazetteers, mainly of African geographical data, like: BioGeomancer, GeoLocate, the Georeferencing Guide for Dummies document (Spencer et al, 2005). Within the EDIT platform project and the EU-BON project there was no actual georeferencing done. This project did however contain different applications and examples of possibilities using georeferenced data.

In the following summaries of the projects, I only discussed the project that contained actual georeference activities.

HerpNET

HerpNET[7] was established in 1999 alongside FishNET2, MaNIS and ORNIS to create a global network of biodiverse collection material. The databases where created by a collaborative network of natural history museums and taxonomic institutions as a reply to the growing international demand for quick and reliable access to this type of data to research climate change. The four databases existed until January 2015, when they were merged tot VertNET[8], which now includes data from 171 international collections.

Georeferencing the available collections is an important part of the HerpNET and VertNET project, to ensure the possibilities of combining biodiverse material with climate data and analyze this on maps. For georeferencing the collections that are available in the database, HerpNET used the GeoLocate tool, the BioGeomancer tool and the HerpNET/MaNIS Guidelines[9].

The RMCA collaborated with 7 other institutions and the University of Berkley and VertNET in 2005, to georeference and add 7 extra collection databases to the HerpNET database. For this project the collaborating institutions received the GBIF DIGIT Seed money Award.

During the project a few prototypes for georeferencing herpetology were tested with the BioGeomancer tool and the Georeferencing for dummies document. For this project the RMCA donated its entire documented amphibian collection to GBIF and HerpNET. This collection contains ca 3000 unique localities from DRC, Burundi, Cameroon, Ivory Coast and Togo, from specimen with linked locality names, station numbers, and collector coordinates. The collection was georeferenced according to the HerpNET protocol, which includes the checking of coordinates and the addition of 'radius' (maximum error radius). With the use of GIS the localities where georeferenced and the extent and error where calculated.

The first step of the project was to digitize the maps, from DRC, Burundi, Rwanda so they could be used as a background layer to georeference the objects. On this layer they visualized the localities of the objects and identified the following cases for each locality: urban area, remote named place, near a named place. The next step was to calculate the extent and the error (maximum error radius) for the georeferenced localities, with the point and radius method. All georeferenced localities were given an ID that could be connected with the specimen data.

In total the HerpNET project georeferenced 645,669 localities, for 1,799,255 cataloged specimens. This is 69% of the total number of unique localities for all institutions connected with HerpNET. Since January 2015 VertNET launched a new data portal combining the 4 previous databases, where all the data from the initial project sites can be searched and used. On the portal the project also promotes several best practices[10] in georeferencing in collaboration with GEOLocate and iDigBio's georeferencing working group, to share useful georeferencing data resources, gazetteers, and tools.

[7] http://www.HerpNET.org/
[8] http://VertNET.org/
[9] http://manisnet.org/GeorefGuide.html
[10] http://georeferencing.org/index.html

SYNTHESYS NA-D 3.7 "Itinerary" project

The RMCA has been collecting objects and data since 1898, which means many of their collection object belong to legacy collections that contain outdated and historical geographical object data. For more recent projects they needed coordinates, instead of textual locality descriptions. To georeference these descriptions the museum analyzed information from the expeditions during which the objects where collected.

Many collection objects were collected during research expedition trips over the past 200 years. These expeditions were sometimes documented very poorly, and are therefore not always useful for research, but sometimes there is more data on the expeditions available in literature like: field notebooks, hand-drawn maps, specimen database records, written comments, rough terrain sketches, digital maps, field number lists etc.

During the SYNTHESYS NA-D 3.7 "itinerary" project (Biocase 2005), the RMCA and The Global Biodiversity Information Facility (GBIF) and the Biological Collection Access Service for Europe (BioCASE) used this literature to attempt to detect certain patterns in the expedition itinerary data, to discover errors or inconsistencies within the geographic data.

The project developed an algorithm to detect which data was constituent with the itinerary and which was not (Meganck et al, 2006). To assess whether or not a record belongs to an expedition the algorithm makes a Boolean (yes- or no) decision, based om a conformity rule. The tool focusses on the coherence of the complete dataset instead of individual objects points. At the most likely pathways and routes where during the expedition and if there are any objects in the dataset that fall outside of these pathways. The sometimes corrects the found anomalies automatically, in other cases it will list the found errors, along with possible causes. These lists can be checked and corrected by experts (Meganck et al, 2006).

For the project the Lang and Chapin expedition to (then) the Belgian Congo, 1909-1915 dataset, was used to test the developed algorithm.

5.5 Georeferencing with google Geocoding API and R

Niels Raes, Naturalis Biodiversity Center

Niels Raes is a research fellow at Naturalis Biodiversity Center, specialized in botany, who developed its own georeferencing script to work with the Google Geocoding API. For many of his research projects he uses distribution models, for which coordinates are needed. Unfortunately only a small part of the Naturalis collection contains coordinates.

The focus of Niels his Research lies on botany in data-poor areas, like Borneo. The problem with these types of areas is that most available georeferencing tools and gazetteers, do not know many localities from these types of areas. The localities in collections from data-poor areas often consist of small settlements, creeks, rivers and hills and local names, whereas known gazetteers like the BioGeomancer and the La Tierra gazetteer focus on larger settlements, rivers, mountains, villages etc. in the more western parts of the worlds.

For one of his researches he needed coordinates with a high accuracy, but couldn't find a right tool to help him with this. That is when he decide to write his own tool to perfectly match his specific needs. The tool is built with R (program language) with the use of the following R packages:

- GGMaps (http://cran.r-project.org/web/packages/ggmap/ggmap.pdf)
- Dismo Package (http://cran.r-project.org/web/packages/dismo/dismo.pdf / http://rpackages.ianhowson.com/cran/dismo/) (Schnörr, 2014)

The data that is needed for the tool comes from GBIF or the collection management system of Naturalis. This is automatically offered to the Google geocoding API in large batches and georeferenced according to the same method, the pilot project of Naturalis used (project 1). The tool links the specimen objects to coordinates, which the researcher can them approve manually. There is also the possibility to check the

coordinates tot land borders, before accepting them. The tool then uses the output from the API and calculates the accuracy of the found coordinates.

High resolution satellite images, old expedition maps and SRTM digital elevation data is also incorporated in the tool, because of the historical nature of some of the collections. These might contain collection sites, like river bends, that could have changed location during the past decades. By checking old expedition maps and satellite images you can discover the original site of these locations, and create more accurate georeference codes.

5.6 Data validation check GBIF

Global Biodiversity Information Facility

GBIF, the Global Biodiversity information Facility[11], is an informational open data infrastructure for digital biodiverse material. The organizations works with and helps biodiversity institutions worldwide to share their collection data for the use of research. By promoting international standards like Darwin Core, it improves the possibility of combining biodiverse material from different collections and with other disciplines.

Their web services provides different guidelines on how to prepare data to comply with the standards of GBIF. To guarantee high quality data GBIF has a backbone of different databases and thesaurus regarding taxonomy, like the Catalogue of Life[12], and geography, like the Getty thesaurus[13] (link). Any dataset that is published on the website is checked against these backbone databases and thesaurus for mismatches.

For geographic data GBIF has a list with all country and country borders for example. If a dataset is published with object that were collected in the Netherlands and there are a few object with coordinates outside this country, the web services gives back a mismatch, called: 'country coordinate mismatch'.

The web services works for datasets with coordinates, datasets with locality descriptions and datasets with both. If a dataset has coordinates, the backbone links the matching locality descriptions to the objects, but this does not work the other way around (matching coordinates tot locality descriptions).

Another mismatch that is found quit often is the switch of the positive and negative coordinates. When plotting data with this is match you can sometimes see the outlining of a country, or area on the other site of the world map. The third thing that GBIF checks are al objects that were given 0.00, 0.00 coordinates. In most datasets only a very few objects where actually found on the 0.00,0.00 coordinates, but there is much larger amount that were assigned these coordinates, there for most of them were probably georeferenced wrong.

GBIF does not change the mismatches itself. This is a task that the data-owning institution has to do. It is however possible to export the checked datasets, including an extensive report on all the mismatches. This way publishing data on GBIF could be used as an extra data quality check for institutions with georeferenced data in their collection. Recently the web services launched a Google Chrome extension that makes it possible to better visualize this type of export data.

5.7 iCollections, the British and Irish Lepidoptera Project

Natural History Museum, London

In the iCollections project the NHM has created a semi-automated georeference method based on the Google Geocoding API. During the iCollections project, the Natural History Museum, London, is aiming to digitize the British and Irish Lepidoptera collections, which includes circa half a million specimens. The main aim is to capture the label data, not the per se the image of the specimen. The georeferencing part is done in an interface that is built by the museum itself and is based on the Google Geocoding API. This method might be used where the first run of the Google Geocoding API method from NBC ends.

[11] http://www.gbif.org/
[12] http://www.catalogueoflife.org/
[13] http://www.getty.edu/research/tools/vocabularies/tgn/index.html

The Google Geocoding API is used as a standard in this project, and not only as a tool. The tool does not deliver 100% accurate or right results, it gives the same type results, for all object records. By mentioning the tool as a 'method' or a 'resource' in the additional metadata fields in a collection database, it creates a unified standard for the georeference data. If in the future a better tool is developed, the google data can be updated.

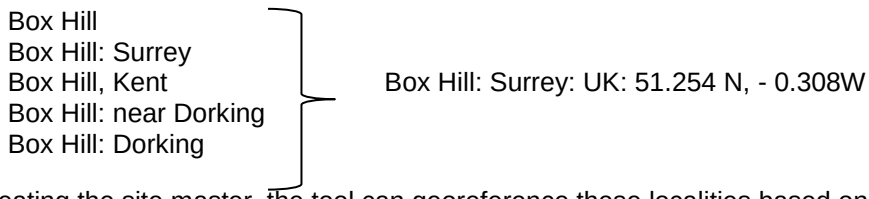
The interface itself was built around Microsoft Access forms, but was only created it to make the process faster and easier to use. The concept of how to georeference was more important during the project, because this can later be adopted in any other project, institution or collection. The interface was just an extra layer, to make the concept easier.

The museum also did two trials with other open source georeference approaches. The trial with de BioGeomancer resulted in 10% georeferencing rate acceptance. Using the OS place names list, in house, resulted in 11% georeferencing rate acceptance. For the Lepidoptera project the NHM used the Google Geocoding API, which results in 12% georeferencing rate.

Data quality is a very important part of the project, to unsure the fitness for use for research, collection management- and preservation and public engagement. The NHM wanted to have full control over the data that eventually will be imported back into their collection management system (KeEMu), and not be pushing large batches of unchecked data into their collection, to have to redo a quality check later on in the process. Therefor the museum has decided to improve the data quality whilst digitizing and transcribing the specimen records and before the data is imported back in their collection management system. To do so, the NHM will first image the specimen and transcribe the label data. This transcribed label data is harmonized with taxonomical datasources and the locality descriptions are georeferenced, to increase the data quality.

The first step of the georeferencing process is to reduce the number of localities by selecting only the unique locality descriptions and linking all variants of one site, to one master site. In a biodiversity collections there are very many site variants that differ in accuracy, description way, spelling etc. By rationalizing you get much more accurate representations of these sites, without duplicating efforts.

The Lepidoptera collection contains around 9681 unique site variants, but these can still contain overlapping sites, due to duplicates, miss- spelling, historical spellings etc. By linking all site variants of one named place tot a 'master site', as the NHM calls it, the number of localities that need to be georeferenced is reduced even further with 65%, to 3400 master records. In the scheme below you can see an example of different site variants of the same named place, Box Hill, but with different larger named places. All these sites variants refer to the same place and therefore belong to one corresponding site master:



After creating the site master, the tool can georeference these localities based on the Google Geocoding API. In the tool, the locality is plotted on a map and can be viewed both visually and in text.

In the tool, the user can search for a locality in the search bar. The tool finds matching locations that are known in Google and the other data resources, such as Geonames. The tool than automatically fills in the following fields: continent, country, province, county and settlement. The user has to select the right location, if there are more than one hits found. The user than has the opportunity to move the pinned locality of the map to a more precise location. When this is done, the tools also fills in the latitude, longitude, extent, methodology and source of the georeference code.

The georeferenced data that is found with the interface is later stored in de KeEMu, with additional metadata besides the coordinates (latitude, longitude), like: datum, method, source, date, by (employee), to ensure sustainability of the collation data.

The tool require quite allot of manual input and is therefore not very suited to georeference large batches of specimen data at once, but it can be used to georeference object records very accurate. Because the tool combines, all the data from different site variants in one site master, this master location contains more data, than one variant on its own. Besides this, this tool offers the possibility to hand select the most suitable locality that is found on google maps, and uses more than one data source than google maps, like Geonames.

5.8 MITCH: Mining for Information in Texts from Cultural Heritage

University of Tilburg, University of Amsterdam, Naturalis Biodiversity Center and ILk

Before the Pilot project Georeferencing Naturalis cooperated with the ILK Research Group of the Tilburg center for Creative Computing at the university of Tilburg and the University of Amsterdam on the MITCH program (Mining for Information in Texts from Cultural Heritage), that ran from 2004- to 2009. MITCH was part of CATCH (Continuous Access to Cultural Heritage), a project with the goal to develop and test information technologies, like information retrieval and data mining, to disclose new data and knowledge from digital Dutch cultural heritage collections (Tilburg University, 2005). ALL CATCH programmers particularly focus on semi- structured data, digitized and stored in databases like collection management systems.

In the context of the project the language technologists of the university of Tilburg developed a system, that help the researchers of Naturalis query the collection faster and easier. The information technology system was designed to help biologists and researchers at Naturalis improve the data quality, derive new information from the database and search the collection easier (Tilburg University, 2005).

One of the processes of collection management that is made difficult due to bad data quality is georeferencing (Tilburg University, 2005). The lack of geographic information usually plays a large role in this. Many specimen records are not annotated with enough geographic information to gather accurate coordinates. Secondly many generic gazetteers and other geographic data resources are not fully useful for biodiverse material (van Erp et al, 2014). Generic resources only contain generic names of locations, whilst locality descriptions in biodiverse collection can also contain local names and other data, like offsets and headings.

By using domain knowledge about species geographical distribution from the online Global Biodiversity Information Facility, like specimen occurrence data, the MITCH project attempts to tackle this process from a different perspective than most georeferencing tools do. Naturalis teamed up with the computer science department of the University of Amsterdam to use data mining techniques like this to georeference biodiversity collections. For georeferencing animals' specimen datasets with data mining techniques, the developers created and tested a prototype at Naturalis (van Erp et al, 2014).

The prototype resulted in the online demo: Geolmp 16[14] as a front end that can be used to georeference single object records or multiple object records in batches, as a csv file. The biologist of Naturalis, the target and test group, could enter queries in one or more fields in the online demo, that would give back georeference location data visually (on a map) and in coordinates. The more query terms a biologist used, the les ambiguous the process was (van Erp et al, 2014). Besides the coordinates the tool would also calculate a confidence score based on the information that was entered. This confidence score is based on a number of decisions like: if the country is known and if the location names can be found in the used gazetteers etc. (van Erp et al, 2014). Based on this the system can point out an object with a bad confidence score to be checked by an expert. The biologists and researchers at Naturalis could then manually add more information to create a better confidence score.

To test the system first a gold standard was created, in which the georeferencing was done manually by two annotators using the MaNIS/HerpNET/ORNIS Georeferencing Guidelines[15]. Georeferencing 120 records for

[14] http://semanticweb.cs.vu.nl/geoimp

[15] http://manisnet.org/GeorefGuide.html

the gold standard, took the annotators about 2 half days. Despite following the guidelines only 60% of the object records were georeferenced with not more than a 1 km difference between the two annotators. For 26% of the object records tested there was no georeference code found (van Erp et al, 2014).
The actual system test showed that using domain specific knowledge, like occurrence data, could lead to more accurate results (van Erp et al, 2014). In the table below you can see the percentage of records georeferenced in the gold standard versus the full system and the improvement the system brings. Using the prototype system the amount of records that could not be georeferenced went down to less than 10%.

Records with correct match within	**Gold standard**	**Full system**	**Improvement**
5 kilometers	38.9%	61.7%	22.8%
25 kilometers	47%	74%	27.5%
100 kilometers	58.4%	79.9%	21.5%

5.10 Conclusion

During the meeting between the NHM, RMCA and Naturalis detailed information on their georeferencing projects and methods were discussed. The expectation arose that there could be more biodiversity institutions within the CETAF that executed georeferencing projects and used tools or methods that could complement the discussed methods form the NHM, RMCA and Naturalis.

In total 7 projects from 6 different institutions within the CETAF where researched and interviewed. Some of these projects developed their own georeferencing methods and some used similar or complementing methods to the ones the NHM, RMCA and Naturalis used.

In conclusion I can state, that there many different possibilities for approaching georeferencing a collection, but not one is fully working and without bugs or questions. It is a verily new subject and there are no readymade methods available. Therefore many institutions have started projects based on research, testing or a trial and error process and created their own methodology for georeferencing.

The BioGeomancer tool was mentioned many times as a very useful and suited tool to georeference biodiverse collections specifically, but unfortunately this is no longer available. Some institution therefore decided to develop their own tool, like the Royal Botanical Gardens of Edinburg. They created a tool that has features specifically for their collection, suited to their data quality and wishes. The result was a tool that can georeference objects mostly with an accuracy of 5 km, and can be used for historical data.

Projects like the SYNTHESYS NA-D 3.7 "Itinerary" project (RMCA) and MITCH targeted georeferencing from a completely different angle: text mining. Where different techniques like name entity recognition, tokenization and conformity rules are used to detect anomalies and find matching coordinates in datasets with geographic metadata.

Thirdly there were several institutions, Naturalis, The Natural History Museum London and The Berlin-Dahlem Botanical Garden and Botanical Museum that used open source georeferencing tools: the Google Geocoding API and GeoLocate. These tools are all available online, and are free of use, but the most important corresponding feature these tool have: is the option to georeference large batches of objects at once.

None of the interviewed institutions collaborated with one another to share experiences, tools, methods, or even georeferenced data. Every institution has its own project with its own project goal. Some focus on georeferencing large batches at a general accuracy level, some georeference specific collections as accurate and precise as possible and some use it to improve the data quality of collections for specific digitalization or research projects. The common conclusion of the interviewed institution was that it is desired, but not that easy to georeference objects with various types of localities in an automatic way or less time consuming, and that even with the help of several tools there will probably always be a part that has to be executed manual. An overview of available and useful tools, methods and projects that could aid in the georeferencing process would therefore be very useful. The overview would help shed light on all parts of the process and which of these could be done automatically and with what tools.

6. Georeference Methods

Derived from the projects that were interviewed for the previous chapter I made an overview of all the tools, methods, data sources, guidelines and datacleaning possibilities that were mentioned. To complete this overview I added information about the functionalities, usability and availability of all these resources. This information is based on an analysis of relevant secondary literature and includes references to manuals for tools, reports and articles describing pilots and tests with the methods or tools and experiences from professionals and previous projects. This overview contains all mentioned tools, methods, guidelines, datasources and datacleaning possibilities regarding if there are available or not.

6.1 Tools

The tools described below can be used to transform textual locality descriptions into geographic reference points like coordinates, and generate additional georeference data like uncertainty's, extents etc. All tools that are available are provided with links and references to user manuals.

a. GeoLocate Georeferencing Software

The Geolocate Tool[16] was created during a project by the Tulane University's Museum of Natural History that aimed at developing a tool for translating textual locality descriptions into geographic coordinates, specifically for biodiverse collections (Tulane University, 2015). The tool provides users an interface to help georeference object records, visualize georeference data and correct calculates coordinates and determine the polygon error. This can be done one object record at a time, simply by typing in the locality description or in batches by uploading CSV, XML or TXT files (Tulane University, 2015). The georeferenced data is returned with the following elements: Latitude, Longitude, and Uncertainty in meters and Polygon error. The tool uses the WGS84 datum and the coordinate precision is exact to the nearest second (Rios, N.E. & Bart, H.L, n.d.).

GeoLocate offers three different ways to use it: 1) as a web application, 2) as a standalone application and 3) as a collaborative georeference application. The common features of the three application types are:

- Import of existing locality data via XML, CSV, or TXT files.
- Batch processing for unattended georeferencing
- Detailed visual map display, including street level data for the United States
- Drag and drop correction of georeferenced coordinates
- Polygon error determination

In the collaborative georeference application users can create communicates and work on shared datasets in which they can retrieve and visualize results, make corrections to existing data, leave comments and mark errors. This is designed for georeferencing very large amounts of object records or collaboration projects between different institutions (Rios, N.E. & Bart, H.L, n.d.).

GeoLocate works specifically great for finding road or river intersections because it uses different georeferencing heuristics, like legal land descriptions and river miles and applies additional linear features to gazetteers (Tulane University, 2015). Therefore the tool can generate more accurate results. These features, however, are only available for the United States, Canada and Mexico. Besides this the tool only works with textual data in English, and does not support any other languages (Rios, N.E. & Bart, H.L, n.d.). The tool also has trouble with misspellings or historic names.

[16] http://www.museum.tulane.edu/geolocate/

b. Google Geocoding API

The Google Geocoding API[17] is an online tool that can be used to find latitude and longitude coordinates of a location. This can be done by the Google Maps application[18], which displays the result on a map. The results could however also be returned in the form of text strings, by using the Google Geocoding API (Google Developers, 2015). The Google Geocoding API uses http requests to communicate with the Google Maps application. The API returns the results in the form of text strings that can be placed in the search option of web browser (Arkel et al, 2014).

Example of a text string:
http://maps.googleapis.com/maps/api/geocode/json?sensor=false&address=rivadavia%203250,%20Ciudad%20de%20Buenos%20Aires,%20Argentina (Google Developers, 2015)

The output of the Google Geocoding API contains the following fields (Arkel et al, 2014):

1. **Address components:** this shows which components are used to find the address; like land, political borders, postal code, house number etc. Not all parts of the original locality are used in all strings.
2. **Formatted address:** this contains the textual address that was found and the latitude and longitude of the found address
3. **Viewport:** The Viewport shows the accuracy of the found results, in coordinates. It indicates an area, shaped like a rectangle, in which the results are located. This is based on the outer southwest and northeast corners of the smallest geographic unit of the address component the Geocoding API found like; city or postal code area. The viewport is created to suggest a possible polygon for displaying it to users.

The Google Geocoding API can handle objects individually or in batches (Google Developers, 2015). To georeference batches you would need an additional program like Open refine too serve the batch files to the API. For both individual objects as for batches the API can return the following types of results:

1. **1 location, Full match:** there is one match of coordinates found based on the information that was entered with the HTTP request. The output of the API contains the found coordinate, the address components, a formatted address and a Viewport.
2. **Multiple coordinates found:** sometime an http request can lead to multiple coordinates that match the location information in the request. In this case the output contains the found coordinates, the address components, a formatted address and a Viewport of all the matches that were found.
3. **1 location, partial match:** the API finds one coordinate that is based on a part of the location information in the request. The output in this case contains the found coordinates, the address components, a formatted address and a Viewport, and in the formatted address the data that resulted in the match is highlighted.
4. **No coordinates found**: the API cannot find coordinates that match the location data in the request. This is called zero results.

Georeferencing with Google Maps (Google Geocoding API) takes processing time. For a search of 1000 object records the software takes at least an hour. If you have basic Google Account you can do up to 2500 free requests a day, after this you will have to pay $0.50 USD for an additional 1000 request. This can go up to a maximum of 100.000 requests daily. If you have a premium account to Google the maximum requests depends on the contract you have (Google Developers, 2015).

To use the Google Geocoding API there are several options that can aid in offering batches of data into the tool for georeferencing, like Open Refine and an R-script.

[17] https://developers.google.com/maps/documentation/geocoding/intro
[18] www.google.com/maps

1. **Open Refine**[19] is a tool that can be used to offer large data batches to the Google Geocoding API. This open source tool is a standalone desktop application that can be used in extension to the Google API and also to improve your data quality of a datasets (see paragraph 6.4) (GitHub 2015). The tool looks like a spreadsheet, but is in fact more like relational database with rows of data and cells under columns (Enipedia TU Delft, 2015). Open refine can handle many different formats imports and export: TSV, CSV (or values separated by a custom separator you specify), Excel (.xls, xlsx), XML, RDF as XML, JSON, Google Spreadsheets, RDF N3 triples (Enipedia TU Delft, 2015).

   After importing data you can start a new project for which open refine automatically imports the dataset into a table view. For georeferencing a dataset with textual geographic data you can create an extra column in the dataset for the georeference data (Enipedia TU Delft, 2015). With open refine you can create URL's that are served to Google in repeated calls to the API. The returned data (read the part about the Google Geocoding API, to see what data is returned) is placed in the column that was created. By using the URL fetch feature you can parse the retuned data to extract the particular elements you want, like the Coordinates (Enipedia TU Delft, 2015).

2. **R** is a programming language that can be used mainly for statistical software and data mining. There is not a ready to go R-Script available at this moment. Housed at Naturalis an in-house researcher created his own script for a research subject 2 years ago. This is not available for the public, but is housed internally.

   The programming language is however extended with packages created by users for specialized statistical techniques, graphic plotting and importing and exporting possibilities (Wikipedia, 2015a). There more than 7000 packages available online, but only a core set of these are included in the installation of R. The rest can be added manually. For creating a script that can be used for the georeferencing process the following packages could be used (Raes, 2014):

   - GGMaps (http://cran.r-project.org/web/packages/ggmap/ggmap.pdf)
   - Dismo Package (http://cran.r-project.org/web/packages/dismo/dismo.pdf / http://rpackages.ianhowson.com/cran/dismo/)

   To create an R- script as such, technical knowledge of programming and this particular language is necessary. The main benefit from designing your own script with R is that you can specify exactly what you want or need based on the georeferencing tool, original data and the collection management system of the institution. With an R-script any desired task can be performed, such as: linking specimen objects, to coordinates with the Google API, manually approving data by experts, check the returned coordinates against land border data, etc.

Besides tools to import data into the Google geocoding API, there is also a way to make processing of the output easier, an interface. This can be done in the same way, the Natural History Museum of London did. They created an interface, with Microsoft Access forms, as an extra layer to use on top of the Google Geocoding API (Personal communication, Malcolm Penn). In a tool like this, the user could search for a locality in a search bar. They would then find matching locations that are known by Google and possibly other data resources, such as Geonames (Hine, n.d.). The tool can than automatically fill in the desired fields, like continent, country, province, county etc. needed for the documentation of the data. The user can select the right location, if there are more than one matches found (Hine, n.d.), after which the user can move the pinned locality of the map to a more precise location, if necessary. When this is done, the tools also fills in the latitude, longitude, extent, methodology and source of the georeference code.

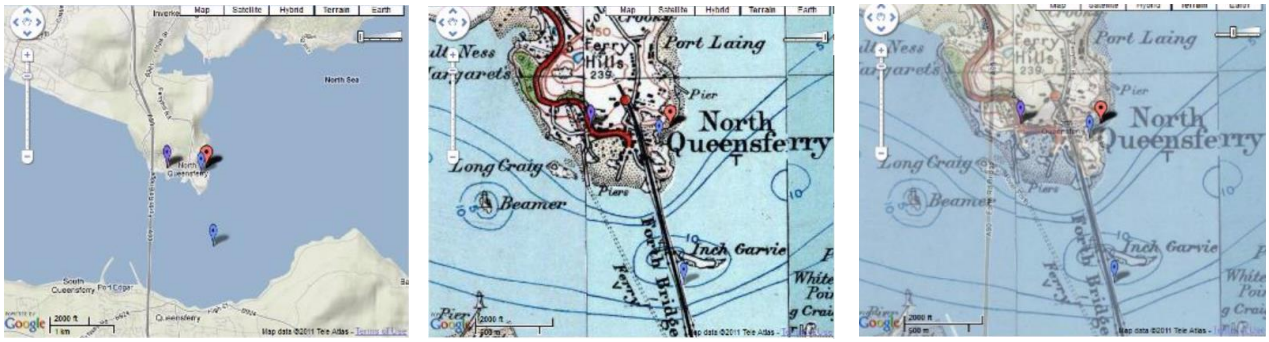
Building your own interface to make the georeferencing process easier and user friendly also makes it possible for more users to help with georeferencing. For a user friendly interface a user would not have to need the same knowledge or experience they would need for tools like open refine or the Google geocoding API. Building you own interface also provides the ability to choose the fields and functionalities that are most important to your institution or project.

[19] http://openrefine.org/

c. SpeciMap

SpeciMap is semi- automated georeferencing tool, designed by the Royal botanical Garden of Edinburgh and the School of Informatics of the University of Edinburgh. It allows users to add georeference codes to textual localities or complement already existing coordinates, from botanical object records (personal communication, Elspeth Haston). The tool combines different text analysis and data mining heuristics and a georeferencing approach, which users can access via a web- based interface (Llewellyn et al, 2012). SpeciMap combines the accuracy of manual annotation with the efficiency of automatic georeferencing, and could therefore be described as semi-automatic.

The base of the interface is a combination of different maps, including Google maps and various historical maps, survey maps and the National Grid Reference. These maps are layered on top of each other and with the help of the interface users can move through the different layers and fade the maps out. This way you can see the changes in costal arrears or cities (Llewellyn et al, 2011). Below you can see a plotted object on three different maps from the SPeciMap tool:



Screenshots of the map display in the SpeciMap Tool.

The tool takes certain fields from the collection database and identifies text strings as places with the help of the Geotagger, Place Name Recognition, a part of the Geoparser tool[20]. This last one, the Geoparser tool, is an open source web tool that can be accessed and used freely at: https://www.ltg.ed.ac.uk/. The second part of this tool is the Georesolver, that look ups these identified text strings in known gazetteers like the Gazetteer of British Place Names and the Fuzzy Gazetteer (Grover et al, 2008). The tool can process textual and numerical (coordinates) localities. Based on these references the tool plots the locality on a map an returns different types of georeference metadata, like latitude and longitude, country, ALTM, National Grid reference, region, habitat. The user can correct or make additions to this generated output in order to specify the exact location where the botanical sample was collected (Llewellyn et al, 2011).

The tool is particularly useful for georeferencing objects from legacy collection or that were collected on difficult geographic areas, such as coastlines. With the help of the different maps you can relocate the georeferenced point to the correct location. This tool only works with individual objects and cannot handle batches of object records.

d. BioGeomancer

The BioGeomancer tool was developed during a project by the University of California at Berkeley, aimed at the development of georeferencing tools and methods for biodiverse data managers. The tool uses textual location descriptions to find corresponding coordinates (Chapman, 2006). The project was completed in 2012, after which the tool is no longer available.

The tool knows different layers that include; satellite imagery, administrative boundaries, USGS topographic names and a topography layer.

[20] http://www.ltg.ed.ac.uk/clusters/Edinburgh_Geoparser/

Once a user uploaded records with locality information into the website, various natural language processing methods are attempted and parse the locality descriptions into element fields (place names, offset numbers, headings etc. that are required for georeferencing (Chapman, 2006).

The georeferencing is done, by searching a gazetteer for matching named places and combining this with the offsets and headings. The tool returns latitude and longitude data and estimates the uncertainty of the georeference data. This is based on the uncertainties of the original locality and the quality of the gazetteers. The uncertainty can be edited by the user by increasing or decreasing the polygon diameter that is placed around the plotted locality (Chapman, 2006). The tool used the MaNIS/HerpNET/ORNIS Guidelines and the Georeferencing Calculator to calculate the uncertainty (maximum error distance).

e. Georeferencing Calculator

The Georeferencing Calculator[21] was originally designed for MANIS, the Mammal Networked Information System as a java applet specifically suited for different locality descriptions that can occur in biodiverse collections. The tool makes georeference and/-or uncertainty calculations based on the MaNIS/HerpNET/ORNIS Georeferencing Guidelines[22]. There is a stand-alone application available as well as an online version of the tool. Both version of the tool, are based on Java 1.1. Since the most recent version of java Is 8, the tool does not function. This tool is therefore not really available, at this moment.

In the tool there are different types of calculations that you can select each requiring different metadata for the calculation (Wieczorek and Bloom, 2011).

- **Coordinates and error**: If you want to figure out coordinates and errors based on a named places, headings and offsets
- **Error only**: If you already have coordinates and only want to calculate the error.
- **Coordinates only**: If you need to determine the coordinates of a named place based on known reference coordinates.

Additional to the different types of calculations a user can also select different types of locations, based on the elements that can be determined in a locality description. These location types can also handle coordinates, headings, offsets, orthogonal directions (Wieczorek and Bloom, 2011). The tool is also multi linguistic and can process Spanish, Portuguese, French English and Dutch locality descriptions.

Thirdly a user can also select: the coordinate source, coordinate system/datum, coordinate precision, direction, offset distance and the extent of named place. This means the tool automatically takes into account different aspects of a textual locality that can affect the uncertainty of the georeferenced point. The output of the Georeferencing calculator consist of the following fields: Decimal Latitude, Decimal Longitude, Coordinate uncertainty in meters, Geodetic datum, verbatim coordinate system, extent, maximum error distance, distance units, distance precession, coordinate precision (Wieczorek and Bloom, 2011). The first five of these elements align with the terms defined by the Darwin Core standard (Wieczorek et al, 2015).

The Georeferencing calculator is a tool that can georeference textual locality description that contain many and various data types, like distances, headings etc. It is therefore very useful for individual objects that cannot be georeferenced in batches, because of fuzziness or bad data quality. It requires a relatively large amount of manual input (selecting al the option) and cannot be used for large batches. The Calculator can also just be used to calculate the extent, uncertainty etc. of existing coordinates in a collection.

f. MITCH

The prototype tool of the MITCH project, Geolmp 16, is a front end interface that can be used to georeference individual objects or batches of objects as one. The tool accepts csv- files for batch georeferencing. A user can enter different queries in one or more fields of the interface, to find locations and retrieve georeference data (van Erp et al, 2014). This is displayed visually on a map and in text form as coordinates. In addition to the coordinates the tool also calculates a confidence score for the retrieved

[21] http://manisnet.org/gci2.html
[22] http://manisnet.org/GeorefGuide.html

coordinates. This confidence score is based on a number of decisions like; if the country is known and if the location names can be found in the used gazetteers etc. (van Erp et al, 2014). To increase the confidence score a user can manually ad more geographic information to the original query.

To georeference textual locality descriptions the tool can use different database fields, namely “Town/City”, “Province/State”, “Country”, “Location” , “Altitude”, and “Coordinates”. For batch georeferencing the tool automatically selects these fields from the database file that is imported (van Erp et al, 2014). The tool can handle locality descriptions in Dutch, English, Spanish, Portuguese and German.

The georeferencing approach of MITCH is buildup of 5 automatic rule based modules that form a process through which all records have to go (van Erp et al, 2014). The final result has to be checked by a user (expert/researcher) before this can be added to the database.

1. **Record Retrieval module**: This module identifies and selects only those data fields from the database that are used by the system (“Town/City”, “Province/State”, “Country”, “Location”, “Altitude”, “Collection Date”, “Genus”, and “Species”)
2. **Text Parsing module**: The second module splits and tokenizes text strings that were selected in the first module. For all the tokenized terms different text analyses techniques are used, like place name recognition to identify offsets, place names and other words.
3. **Gazetteer Lookup module**: the terms that were identified and categorized as place names in module two are looked up in Geonames and Google Maps.
4. **Offset Calculation module**: in the text parsing module also offsets were identified and categorized. These numbers are combined with the coordinates that are retrieved from the gazetteers in module 3. For the calculation of the exact coordinates the Perl Geo::Calc15 module is used.
5. **Disambiguation Heuristics module**: in the last module domain specific knowledge, like spatial minimally, species occurrence data, is used to disambiguate location names. This is done with several disambiguation heuristics.

The MITCH prototype tool is no longer available since the project ended. Similar text analysis methods could however be very useful for the georeferencing process, because it tackles the process from a different perspective. Text analysis might help u find georeference data for object record that could not be georeferenced by the other tools and vice versa.

6.2 Guidelines/best practices

In this part existing best practices and guidelines, unconnected with tools, are described. These guides can be used for any georeferencing project and discus show to deal with different part of the georeferencing process, so that the data can be generated, accessed, stored and managed in a sustainable way.

a. MaNIS/HerpNET/ORNIS Georeferencing Guidelines

The MaNIS/HerpNET/ORNIS Georeferencing Guidelines[23] are developed by John Wieczorek to make the complicated process of georeferencing easier and to offer solutions to problems that user might encounter when georeferencing difficult and vague locality descriptions. The document explains the process of finding matching coordinates for textual localities and calculating uncertainty for different elements in geographic data (Wieczorek, 2001). It does not focus on tools, and can therefore be used for most projects regarding what tool that is used. The reason for designing these guidelines is to create a methodology that will lead to consistent results that can be reproduced over time, regarding the tool or source of data (biodiverse-, historical-, costumer- data etc.).

The guidelines uses 2 main data elements to determine georeference data of a locality; 1) the coordinate (point) of the center of a named place and 2) an uncertainty (“error”) around it to account for the extent of the

[23] http://manisnet.org/GeorefGuide.html

named place, uncertainty of direction and distance, datum used, etc. (Wieczorek, 2001). For the notation of georeferenced coordinates the decimal degrees are most useful, simply a locality can be described with only two elements - decimal latitude and decimal longitude. For assigning geographic coordinates are presented, with examples, in the first section, Determining Latitude & Longitude. In this section not only named places, but also offsets (distance and heading) and vagueness of locality descriptions is encountered for.

The Guidelines state there are several, 7 in total, sources that can be found in locality description that can cause uncertainties when georeferencing (Wieczorek, 2001). All seven sources of uncertainty are explained and a guideline is given of how to record these uncertainties in a single measurement; the maximum error distance. This is discussed in the second section, **Determining Maximum Error Distance from Uncertainties**.

1. **Extent of a locality:** The maximum error distance (uncertainty) is calculated by taking the distance between the coordinates and the furthest point within that named place.
2. **Uncertainty due to GPS accuracy:** Most GPS data that is recorded is missing the estimated uncertainty that was given by the GPS system. An uncertainty of 30 meters is reasonable for most GPS systems.
3. **Uncertainty due to an unknown datum**: For calculating the uncertainty of a record of which the datum of the original coordinates is unknown, the guidelines provide a map of the United states with a color gradient. For each color a reasonable maximum error distance is given.
4. **Uncertainty associated with distance precision**: To calculate the maximum error distance based on distance precision you have to split the distance precision in whole number and marginal remainders. Calculate the uncertainty for these distances based on the marginal remainder of the distance, using 1 divided by the denominator of the margin.
   Example: "**10.5 mi N of Bakersfield**" (the fraction is 0.5, uncertainty should be 0.5 mi)
5. **Uncertainty associated with directional precision**: heading like “North” and “South” usually cause uncertainties because, it is not possible to know what the recorder of this data meant by it, when it was recorded. The uncertainty caused by direction in reasonable 45 degrees in either direction from the given direction. When the direction is somewhat clearer, "NE", this could indicate any direction between “ENE” and “NNE”, which is twice as precise as only “north”. The directional uncertainty in this case is 22.5 degrees in either direction from the given direction. Directions that are even more precise like “ENE”, the directional uncertainty is divided in half again and therefore 11.25 degrees.
6. **Uncertainty associated with coordinate precision:** When recording coordinates, it’s best to always record as many digits as possible. Rounded coordinates with can result in uncertainties. To calculate the maximum error distance can be calculated by a formula that is given in the Guidelines (uncertainty=sqrt(lat_error^2 + $long\_error^2$)
7. **Uncertainty due to map scale**: For this source of uncertainty the Guidelines provide a table showing the uncertainty due to scale for USGS maps.

The MaNIS/HerpNET/ORNIS georeferencing guidelines explain how to georeference records and what theory/methodology is behind it. Besides these guidelines there are two additional documents that can be used; The "Georeferencing for Dummies" a simple table that summarizes the Guidelines and suggests the best georeferencing methodology for each defined locality type, and the Georeferencing Quick Reference Guide, a practical guide that should be used in addition to the guidelines. This guide uses the Point Radius method, the georeferencing calculator and several other resources like gazetteers and maps to create maximally useful georeference information, following the Darwin Core data standard (Wieczorek, 2001).

b. Chapman’s *Guide to best practices for georeferencing*

The Guide to Best Practices for Georeferencing van Chapman[24], is one of results of the BioGeomancer project and specifically focusses on georeferencing primary biodiverse records data. For creating a comprehensive best practice the authors used experiences from different projects (BioGeomancer Classic, MaNIS, MapSTeDI, INRAM, GEOLocate, ERIN) that also attempted to create best practices. This final

[24] http://www.HerpNET.org/HerpNET/documents/biogeomancerguide.pdf

document attempts to bring all those earlier experiences together to create one standard methodology for the biodiverse field (Chapman, 2006).

The Guide provides guidelines for the entire georeferencing process, from the recording of geographic data in the field during the collection event, preparing the data and databases for georeferencing, georeferencing legacy data, improving and maintaining data quality. Besides this the document contains many examples of locality types and the best way the georeference these.

The best practice calls itself a basis for georeferencing that could and should lead to internal documents that incorporate these guidelines and the organizations own policy, environment and goals (Chapman, 2006). Even though the guide is a result of the BioGeomancer project it could be used for georeferencing with any tool or resource. The methodology follows existing procedures and standards, like Darwin core.

One of leading guides in the document are the 'principles of Best Practice'. These are 8 guidelines for improving the data quality and maintaining this quality when georeferencing the primary biodiverse data:

1. **Accuracy:** this measure how well the original data is represented in the georeference value, in the measure of percentage, a polygon or an uncertainty in meters
2. **Effectiveness:** is measured by the number records for which usable latitude and longitude coordinates are found with an acceptable accuracy.
3. **Efficiency:** the amount of effort, for example expressed I man hours, that is needed to generate a desired/ acceptable output. This can be measured in the man hours needed to generate this output or the amount off original geographic that is needed to generate the desired results.
4. **Reliability**: this is what in most scientific research is called repeatability and is measured by the consistency of the obtained results.
5. **Accessibility:** how easy the users, public etc. can access the georeference data for a particular locality.
6. **Transparency**: this is measured by the quality of the metadata, methodology, tools, procedures etc. of georeference data is generated and documented.
7. **Timeliness:** this is related to how often the gazetteers, maps and guidelines that were used for the georeferencing process are updated and how much time is between the georeferencing and the release of the data to the users/public.
8. **Relevance**: This largely depends on the user of the georeference data, the "fitness for use. This can mean anything form the format, standards, accuracy notation etc.

The use of all named procedures for generating coordinates and calculating the extent and maximum uncertainty distance for locations, should eventually lead to data that conforms to these principles (Chapman, 2006).

Parts of the Guide to Best Practices for Georeferencing of Chapman complies with the MaNIS/HerpNET/ORNIS Georeferencing Guidelines. These two documents could be used side by side for creating an internal georeferencing protocol for institutions.

6.3 Data sources

The data sources that are described below are free and online available databases and resources containing worldwide geographic information. There are many more geographic data sources available, but these are specifically useful for georeferencing, because the know place names and coordinates, and some can handle the type of geographic data that can be found in biodiverse collections (i.e. large batches, messy data).

a. Geonames

Geonames is a geographic database that is available via a range of web services and as a daily generated database export and can be used without costs under the Creative Commons Attribution license (CC-BY). The database contains about 10 million geographical names, of which 2.8 million populated places and 5.5 lion alternate names. All geographic names are categorized in the following codes groups: country/

state/region (A), stream/ lake (H), parks/area (L), city/ village (P), spot/ building/ farm (S), mountain/hill/rock (T), undersea (U), forest/heath (V) (Geonames, n.d.).

Geonames also knows additional information of all 10 million geographic named places such as coordinates, alternative names, elevation levels and population numbers. The coordinates of named places are in WGS84 (World Geodetic System 1984) (van Erp et al, 2014). The database is specifically useful for geographic named places in populated areas. Geonames combines data from over 70 data resources, from different country's (in different languages) and different types of data, like postal and coast country-coordinate lists (Geonames, n.d.).

b. The Getty Thesaurus of Geographic Names

The Getty Thesaurus of Geographic Names (TGN) is one of the Getty structured vocabularies that can be used to improve access to information. De Getty contains a vocabulary of geographic names with hierarchical, equivalence, and associative relationships, but it can also be used to ad georeferencen data (The J. Getty Trust, 2015).

The structured vocabulary knows geographic place names and for many of the records also geographic coordinates. These are expressed in degrees/minutes and decimal fractions of degrees using the WGS 84. The coordinates mainly correspond to the center or point near the center of a named (populated) place, physical feature or political entity. The Getty Thesaurus knows around 1 million named places worldwide. With the different relations, different languages and historical names, like roman names, are matched to each term in the vocabulary. The TGN is not a GIS and the coordinates should only be used for cataloging and referencing (finding) to localities (The J. Paul Getty Trust, 2015). The TGN can be used in two ways:

1. The TGN has a web application[25] that can be used for georeferencing. The web application does not offer the possibility to georeferencen records in batches. Here you simply type in your geographic query and extensive georeference data is returned (The J. Getty Trust, 2015).
   In the online version u can use Boolean operators (AND OR) and wildcards (*) to find named places of which the spelling is questionable. For exact named places u can also use quote (" ") marks to find the correct named place (The J. Paul Getty Trust, 2015).
2. The TGN can also be incorporated in a collection management system as a reference list/thesaurus it could be used without exporting datasets from the collection management system. The regular place names are usually indicate with the element 'thesaurus (term)', and the coordinates for that place are indicated with the element "thesaurus detail" (The J. Getty Trust, 2015). These elements are linked to one another. So if a locality description of a record is linked to a thesaurus term, this is automatically linked to the thesaurus details, aka the coordinates. It however depend on the system if this is visible (Arkel et al, 2014).

In theory there are two options of using the Getty Thesaurus for georeferencing objects, but it depends on the functionalities of a collection management system if both options are applicable. If this is possible it could create very uniform geographic data for a collection, that all originates from the same (good quality) thesaurus (Arkel et al, 2014).

c. Google Maps

Google Maps[26] is one of the most used geographic resources for the public, but it can also be used as a geographic resource for obtaining named places and coordinates. It can be used for looking up urban areas, making extents and for matching addresses to coordinates (Wikipedia, 2015b). Because it has a built-in ranking mechanism, like all google services, it return important places like postal offices, train stations and courthouses first. So for finding the center of named places this service is particularly useful (van Erp, et al, 2014). This does however; make it more difficult to find accurate georeference data for unpopulated areas...

[25] http://www.getty.edu/research/tools/vocabularies/tgn/
[26] https://www.google.nl/maps

Besides the Google maps online application, there are several web services that use Google Maps to create searchable gazetteers based on Googles data, like Maplandia.com[27]. This secondary web services of Google maps, knows more than two million continents, countries, cities and administrative regions.

Google maps is free for commercial use and unlike other google services does not contain ads. If you want to incorporate it into your own site it is provided that the site on which it is used is publically accessible with charges or access requirements (like user accounts) (Wikipedia, 2015b).

The google web maps web service is mainly useful for populated places and finding geographic centers, but because of the available high resolution satellite imagery, street maps and 360° panoramic views, it is possible to zoom in and find places in less populated areas as well (Google Developers, 2015).

d. The Fuzzy Gazetteer

The Fuzzy gazetteer is a specialized online geographic search service of the Joint Research Centre of the European Commission. The Fuzzy G contains over 7 million place names, mainly originating from the Geonet Names Server (Christian Kohlschütter, 2003). It returns found place names and corresponding coordinates in Degrees, Minutes, and Seconds. The datum of the returned data is WGS84 and the precision is to the nearest minute.

The special feature of the Fuzzy gazetteer is that it can find any place names, regarding if you know the correct spelling, because it focusses more on the vowels. You can simply type in the place name you have of think is correct and the gazetteer returns a list of possible matches with similar names. You can select the degree of fuzziness ("Very Fuzzy," "Quite Fuzzy," "Default Fuzziness," "Not So Fuzzy" and "Only Exact Matches"), based on how correct you think the original data is and similar the returned names can be (HerpNET, 2007).

e. Histograph

The Histograph[28] is a historical geocoder designed in Holland, for searching and standardizing historical place names in a dataset. The geocoder technique collects and links place names of the same locality over time and standardizes and georeferences these names (Histograph, n.d.). A user can type in a place name like Amsterdam and Histograph finds al different names that were used in history for that place name, including the place boarders, numerical time definitions, and the original source of the data (Erfgoed & Locatie, 2015).

At this moment the Histograph tool uses data sources with; birth places of Dutch East India Company crew members, monastery records and historical census data for the historical place names, and the Geonames and TGN for the standardized modern place names (Erfgoed & Locatie, 2015). The tool no focusses only on Dutch data, but the techniques are open source and could be used for the same purposes for different areas or countries, by adding different historical geographic data.

The Histograph tool is at this moment still under development, but a demo version can be used trough the Histograph API. In this API u can search based on place names and source URI, like the Geonames and TGN databases. With this last option you can search for place names in a specific source (Erfgoed & Locatie, 2015).

6.4 Data cleaning and validation

Datacleaning regards the improvement of the original geographic locality description from a collection to increase the percentage of georeference results. This can be done in advance of using a georeferencing tool, for example to filter unique localities, removing vague elements in locality descriptions or transforming them into the elements required for a georeferencing method/tool. Datacleaning could also be done after the first run of a batch through a georeference tool to manually adjust the records that did not return the desired

[27] http://www.maplandia.com/
[28] http://histograph.io/viewer/

results. Datacleaning can be done manually, by an expert or with the help of a tool, but in most cases it requires a reasonable amount of man hours. It does however increase the quality of the results.

a. Manually visual check with Google Maps

After the first run of data through the Google Geocoding API, some records will not have returned useful responses. This is mostly due to some elements in locality description, that Google does not recognize, like “1756 meter”, “NS”, “by road” etc. A good, but vey labor intensive, method to plot and adjust this data is with the visual datacleaning based on Google Maps (Arkel et al, 2014). This correction method is required when using the Google Geocoding API, for the records that did not return with a full match, but it can be used in addition to other tools.

For this method, you will need to import all the data into excel. Excel can transform the returned strings form the first test run, in links (URL’s) with a formula. For example, the following formula can be used for this:

=HYPERLINK(TEKST.SAMENVOEGEN("https://www.google.nl/maps/search/";[Cel *met zoekstring]*))

This link can visually show what the result of the returned georeference code is. By manually adjusting the search string in Google Maps, you can create a string that delivers better acceptable results (Arkel et al, 2014). This scan be done for records that have returned multiple results or records that didn’t deliver any results at al.

Because a search string is record specific this way of datacleaning can only be done one record at a time, making it time consuming. It does however deliver very precise results (Arkel et al, 2014). Since the underlying map is Google Maps, you can view high resolution satellite images, street maps and 360° panoramic views, u can zoom in and adjust the string to the precise location of the object (Google Developers, 2015).

The new string, with the good results, can then be used again, to georeference the records like the first run that was done with the initial georeference tool.

b. GBIF

The Global Biodiversity Information Facility is not only an open data portal for biodiverse material, but it can also used to check several data elements in a dataset. To guarantee high quality data GBIF has a backbone of different databases and thesaurus regarding taxonomy, like the Catalogue of Life[29], and geography, like the Getty thesaurus[30] (personal communication, Cees Hof). Every dataset that gets published on the data portal of GBIF is checked against this backbone for possible errors, called mismatches.

Mismatches in this case can be defined as differences between the dataset and the backbone, that could indicates errors or mistakes. For example: GBIF has a list of all countries and their coordinates in the backbone. When an object in a dataset is uploaded with ‘Netherlands’ as country but coordinates that fall outside the country- coordinates for the Netherlands from the GBIF backbone, this is marked as a mismatch.

GBIF defines many more mismatches

- Country coordinate mismatch: when a country does not match the matched coordinated.
- Geodetic datum assumed: GBIF uses the WGS84. When this is not recorded for datasets, GBIF assumes based on the recorded coordinates, if this datum is sued.
- Zero coordinate: objects that were given 0.00-0.00 coordinates are usually are wrong. There is only a very small percentage of species that occur on this location.
- Reversed georeferencing: GBIF does add country names from their database to, if the coordinates of the species occurrence is known, and does not result in a mismatch.

[29] http://www.catalogueoflife.org/

[30] http://www.getty.edu/research/tools/vocabularies/tgn/index.html

One part of the GBIF check can be viewed online. For each dataset GBIF provides a table that shows al objects categorized by kingdom, if it is a preserved specimen or only an observation, a fossil or a living specimen. For each of these categories the table shows the total number of records and the number of records that are georeferenced.

GBIF does not change the mismatches itself. If a dataset has coordinates, the backbone links the matching locality descriptions to the objects, but this does not work the other way around (matching coordinates tot locality descriptions). This is a task that the data-owning institution has to do. It is however possible to export the checked datasets, including an extensive report on all the mismatches. This way publishing data on GBIF could be used as an extra data quality check for institutions with georeferenced data in their collection. Recently the web services launched a Google Chrome extension that makes it possible to better visualize this type of export data.

c. SpeciesLink

The goal of the *SpeciesLink* network is to integrate specie and specimen data available in natural history museums, herbaria and culture collections, making it openly and freely available on the Internet. The web services provides a number of tools that can aid in the georeferencing process. The database mainly includes geographic data from brazil and it knows about 750 thousand names of Brazilian localities.

The most useful tool for improving the georeference data quality is the Datacleaning tool[31] (link). This tool can be used to identify errors and standardizing data. The tool only detects errors so they can be checked by each data owner. In some cases the tool offers a correct possibility for the suspect data element, but it does not modify the dataset itself.

Secondly the spOutlier tool[32] can be used to detect outliers in latitude, longitude and altitude according to the techniques from Chapman, 1999 (link). This tool accepts only batches of data in excel file formats. The output will highlight all suspect records in red and all good records in green. The output is shown in text as well as visually on a map. For this map the user can choose from a selection of different map layers. Lastly the SpeciesLink website also has a Converter tool, for converting different types of geographic coordinate systems and datum's.

Unfortunately the SpeciesLink website is not accessible for every institution to use. It 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.

d. SYNTHESYS NA-D 3.7 'Itinerary' project

The SYNTHESYS Itinerary project (BioCASE, GBIF, and RMCA) developed an algorithm that compares geolocations available in literature on expeditions like: field notebooks, hand-drawn maps, and rough terrain sketches, field number lists, with georeferenced primary specimen data to discover errors or inconsistencies within the geographic data.

The algorithm assesses whether or not a record belongs to an expedition by making a Boolean (yes- or no) decision if the data is consistent with the itinerary form the additional literature (Meganck et al, 2006). It analyzes what the most likely pathways and routes where during the expedition, and if there are any objects in the dataset that fall outside of these pathways.

The algorithm specifically focusses on the conformity of objects in a dataset and is therefore not useful for checking individual objects. The algorithm uses a conformity rule set to make the yes- or no decision. The data in the dataset is sorted by date and al objects are compared to the previous point (Meganck et al, 2006). Other records that fall beyond this are considered to be conforming if:

[31] http://splink.cria.org.br/dc/?criaLANG=en
[32] http://splink.cria.org.br/outlier?criaLANG=en

1. Time and date are sufficiently precise: with an initial value of 1 day, which could cover the most accurate points, and probably most of camp-making events.
2. The speed since the previous point doesn't exceed the fixed limit for possible distance per day. If his does occur, several actions can be taken, like warning a user to check the record point, considering some larger day distance could indicate the expedition team using faster transport resources (e.g. train, boat,...), or starting a new itinerary or rejecting the point.
3. The total spatial uncertainty of the end- and beginning point is less than the distance travelled

Apart from the conformity yes or no decision, a score calculated to indicate what the conformity of the points are together. For calculating this score the following parameters are used *(Meganck et al, 2006)*:

1. **Total distance travelled**: if the total distance traveled is to long it may indicate improbable constructions, or inconsistent data points.
2. **The number of points** : if there are more point describing the same distance increase the accuracy
3. **The total uncertainty of the points**: the total spatial uncertainty. It counts the uncertainty values of the middle point double (except for the first and the last point), since they work in two directions.
4. **The total "slack**": "Slack" indicates the measure of the maximum possible anomaly of the straight path between two points considering the theoretical distance that could be covered considering the time path.

The conformity score value can be calculated for any number of points, with a lower score indicating a better 'fit' of the points (Meganck et al, 2006). The itinerary project uses ABCD as a data standard. Some concepts from the existing standard that seemed relevant were used, and additional fields were determined to comply with the standard.

e. Plaatsnamen Standaardiseren

Plaatsnamen Standaardiseren[33] is an online tool that can be used to standardize Dutch place names in a dataset. Standardizing is very important for the searchability of collections and the possibility to exchange data between different collections (Erfgoed & Locatie, 2015). The tool, which is only a demo version and only available in Dutch, is designed for collection managers to easily import data, standardize the geographic elements in the dataset, check the results and export the standardized data. Below you can see the steps required to standardize a dataset (Den Engelse, n.d.):

1. Step 1: create a csv. file with the place names that need to be standardized. The csv. File should contain a column with the terms that you want standardized. This could be one or more columns.
2. Step 2: upload the csv. file and select the field that contains the place names and what type of locality it contains like, cities, states etc.
3. Step 3: let the tool do the standardizing
4. Step 4: review the results and edit the results if necessary (delete wrong matches, choose the right match if an object returned multiple results etc.

The output of the tool contains 4 tabs; 1) the standardized localities that returned in one full match, 2) localities that returned multiple matches. These results can be displayed on a map and can be edited by the user. 3) the third tab contains localities that did not deliver any results, this could indicate a misspelling or a place name that the tool does not recognize in any of the standard datasets, but these can also be adjusted manually by the user and 4) this tab is intended for localities that cannot be standardized by the tool nor with manual input of the user (den Engelse, n.d.).

The tool at this moment only works for Dutch geographic information, and does not know any coordinates. However the tool uses the TGN and the Geonames databases to generate the standardized terms (Erfgoed & Locatie, 2015). So if the filter that only focusses on Dutch data would be removed, this could be used for worldwide geographic data. This tool could be very useful for the data cleaning process before the actual georeferencing is done to generate a higher percentage of full matches, during the first run.

[33] http://standaardiseren.erfgeo.nl/

f. Open Refine

Open Refine, as mentioned in paragraph 6.1, is an open source and online available tool, based on the functionalities and usability of a relational database. It look very similar to excel and knows various formulas and features to edit and manipulate data and datasets (Enipedia TU Delft, 2015). With these features it can be used to improve your data quality of a datasets by cleaning up the data and transforming it into other formats (GitHub 2015). For example, with Open Refine you can:

- Improve messy data and create structured data.
- Transform data into other formats
- Parse data from websites, by using the URL fetch feature
- Add data to the dataset from web services, for example georeferencing textual locality descriptions to coordinate

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 locality descriptions, because google can see these as a part of an address and for deleting double location names. It can also serve as a transformation tool for creating different formats, parsing returned georeferencing codes into multiple database fields and creating strings for the visual check with Google maps.

g. Crowdsourcing

Another, possibly useful, topic that can be used for datacleaning in the georeferencing process is crowdsourcing. Crowdsourcing is not only useful for datacleaning, but can also be deployed for data preparation, validation, georeferencing individual objects, etc. In the field of science and scientific research, and therefor biodiversity, crowdsourcing is generally referred to as citizen science. The technique uses the crowd (the public) for carrying out tasks that normally requires allot of man-hours. This can be any test an institution traditionally preforms itself, but does not have the time or money to complete.

One important fact about to crowd is the 90-9-1 percent rule. This states that 90 percent of the whole crowd won't do anything, 9% only contributes once and 1% of the crowd is the group of people who will do most of the work (Ridge, 2007). However that 1 contributing percent of the crowd has the most knowledge and drive to contribute to the project. Especially in biodiversity and other scientific disciplines, the one percent can be very knowable and have more expertise than expected. These groups usually contain so called Pro-Am's, professional amateurs, who might not be directly employed by the organization to which they are contributing, but might have worked in the field for years, or be a member of an amateur organization (Oosterman et al, 2014).

Crowdsourcing could be very useful in the georeferencing process for the labor intensive datacleaning tasks. Labor intensive datacleaning requires allot of time and expertise, but is proved to be very effective. It can improve the percentage of georeferenced records and can create better quality georeference data. If a biodiverse institution would have to perform these tasks themselves the results would not out way the time and expertise spent (Blaauboer, 2014). Even the datacleaning methods, using software, like the visual check by Google, or adjusting records that were highlighted suspect by web services like GBIF and *SpeciesLink* could adjusted with crowdsourcing.

The crowd could for example be used to create unique localities, by matching place names that refer to the same locality, or by linking place names to thesaurus terms. This way the first run of records will already deliver higher results. The crowd could also be used, to check and adjust records that need to be checked individually, for example with the Visual check of the Google API. It can even preform manual georeferencing with tools like Specimap, or improving the accuracy of a georeference code with SpeciMap, the visual check with google or the Georeferencing Calculator.

This option does however; assume that an institution has access or the possibility of creating a crowdsourcing platform and a population of citizen scientist (van Erp et al, 2014). This does require time and money to create, but generally the overall conclusion is that a crowdsourcing project delivers more than just

the data the institution gathers. It creates a higher involvement of the crowd in biodiversity; it can bring new visitors and even generate publicity with the project (Blaauboer, 2014).

The second 'but' allot of institutions have is the quality of the data that the crowd produces, because you can never know exactly who the crowd is (Fleurbaay, Eveleigh, 2012). Besides this, even the biggest experts in the crowd mainly focus on urban and semi-urban regions and more popular topic fields, like fish and birds (Pulla, 2013). But if the crowd were to help with the more easy tasks and more popular topics, this would leave the experts to have more time to spend on the difficult parts of the collection.

The quality of the data that the crowd produces can in increased with a number of measures (Fleurbaay, Eveleigh, 2012):

- **Guidelines:** many crowdsourcing platforms provide users with guidelines that indicate how the platform should be used and what data is wanted.
- **Formats:** to prevent that al crowd sourced data differs in form you could use a notation format, for example for the desired notation of dates, the maximum length of elements etc.
- **Pick lists:** is another way to ensure that all data is comparable. An example of this is a thesaurus. It is also possible to combine the formats and the pick lists, so that user can enter their own interpretation and match this to a thesaurus.
- **Double entry:** this is a technique were one record is crowd sourced by multiple users. Only the corresponding data of both entries is accepted. The number of entry's that are needed depends on the data and the data owning institution. Another option is to have two or more users enter the data and one user (or expert from the museum) check this and approve or edit this.
- **Nichesourcig:** is a measure that can be used, when the crowdsourcing tasks requires a very specific expertise. With nichesourcing the data owning institution can pick the crowd itself, for example students from a certain school, members of an amateur organization etc. and only grant access to this group to the crowdsourcing platform (Blaauboer 2, 2014).
- **Marking:** the crowd sourced data when recording this in the collection management system. This it will always be clear which data is created by the crowd and the end-users of the data can decide for their own, if this is still useful for their purposes.

When starting a crowdsourcing project for georeferencing objects or datacleaning activities it is important to give the crowd the opportunity to really contribute to the project, and the institution. This has a very positive effect on their motivation and is more useful for the institution (Oosterman et al, 2014). If you decide to spent time, expertise and money in a crowdsourcing project it is important that it delivers real results that can really be used by the institution, for research or collection management (personal communication. Otherwise the costs won't out way the results (Blaauboer, 2014).

7. Tool/ methods selection

In the tool selection all information collected in the chapters 2 -6 is combined to come to a selection of tools that can best be used and or combined for georeferencing biodiverse primary data. To asses which tools, methods, data sources and datacleaning methods are most useful, and could complement each other, these are placed in a table with al important features based on the user needs, data element requirements '*Principles of Best Practice'*, of Chapman (Chapman, 2006). The effectiveness of the tools is not included in this tool selection process, since no test were preformed to measure this.

The goal is not to come to one tool that works best. The expectation and aim of the research is that there could be more tools useful and fit for georeferencing, and that when these are combined the lager part of the whole process could be automated as much as possible.

Tools

After deleting tools that are not or no longer available; 4 tools for georeferencing remain: Geolocate, The Google geocoding API, the Georeferencing calculator and the Getty Thesaurus of Geographic names as a thesaurus integrated in a CMS. These tools can be used to transform textual locality descriptions into geographic reference points. Preferably this is done in larger batches for the objects that have relatively simple geographic metadata, and individually (partially manual) for objects with more challenging geographic information, like historic collections.

	Geolocate	**Google Geocoding API**	**Georeferencing Calculator**	**SpeciMap**
Accessibility				
Online	yes	no	Yes, if updated	Yes, once released
Web service	yes	Yes	Yes, if updated	unknown
Offline	yes	no	Yes, if updated	Unknown
Free	yes	2500 free requests a day (for educational institutions 100,000 requests a day)	yes, if updated	Yes, once released
File formats				
Accepted import formats	XML, CSV, or TXT files	HTTPS request via Open Refine.	not applicable, since it's not possible to upload batches	Individual objects go directly from collection database
Output format	delimited text files , CSV and specially formatted XML file	JSON or XML	not applicable	Unknown
Standard compliance	MaNIS/HerpNET/ORNIS Guidelines	no	5 elements align with Darwin Core	No
Batches	Yes, but very difficult.	yes	no	No
Individual	yes	yes	yes	Yes
Results				
Accuracy	Uncertainty in meters and Polygon error	Based on the outer southwest and northeast corners of the smallest found geographic unit (given in coordinates)	Maximum error distance based on the MaNIS/HerpNET/ORNIS Guidelines	Unknown
Datum	WGS84	WGS 84	WGS84	unknown
Precision	exact to the nearest second	6 decimals	multiple possibilities: nearest second, nearest degree, etc.	exact to the nearest second

Data elements	Latitude, Longitude, Uncertainty in meters and Polygon error	Address components, Formatted address (textual and coordinates), Viewport (accuracy)	Decimal Latitude, Decimal Longitude, Coordinate uncertainty in meters, Geodetic datum, verbatim coordinate system, extent, maximum error distance, distance units, distance precession, coordinate precision	latitude and longitude, country, ALTM, National Grid reference, region, habitat
Manually adjustable	Yes	No	No	Yes
Geographic focus area	North America, Canada, Mexico	Worldwide	Worldwide	Worldwide, but old maps and grid reference, only for England.
Efficiency	requires allot of manual input	Only requires input to import/adjust data in Open refine. Tool itself takes 1 hour per 1000 objects	Requires allot of manual input, since it only works for individual objects	Requires allot of manual input, objects are replaced individually and manually
Reliability	yes	yes	yes	yes
Transparency code/algoritm	Sample source code and documentation: http://www.museum.tulane.edu/geolocate/developers/default.html	Documentation: https://developers.google.com/maps/documentation/geocoding/intro	no	Will be released open source.
Data quality				
Accepted languages	English	54 languages	Spanish, Portuguese, French English, Dutch	English, the rest is unknown
Historic names	no	Unknown	no	Yes, manually
Miss spellings	no	Built in ranking mechanism that finds well known and populated places based on original request	no	Yes, manually
Other elements	no	No	yes	Yes manually

There are only a two tools available that can georeference objects in records in batch volumes, namely The Google Geocoding API and Geolocate.

Geolocate seems like a very useful tool, considering the additional metadata elements it generates, the compliance with standards and the possibility to use it online as well as offline. Based on experiences from Naturalis however, the batch georeferencing option of Geolocate is not fully functioning and difficult to use, and therefore requires allot of man-hours. Secondly Geolocate only accepts English data, and works best for localities in North America, Canada and Mexico. This makes it unfit for large parts of biodiverse collection, that contain different languages (like Naturalis) and contain objects originating from different parts of the world.

So, **The Google Geocoding API** is the only available tool that can georeference objects in batches. With the help of Open Refine, also free, open source and available online, users also don't need that much technical expertise to use it. Open Refine works as a database, but looks like a Excel sheet to the users and there many manuals available online. Google also accepts over 50 languages and has geographic information covering the entire world and offers high resolution satellite imagery, street maps and 360° panoramic views. It has a build in ranking mechanism that finds matching well known and populated places based on the original request, so it could work for data with minor miss spellings. The output of Googles Geocoding API does not comply with standards, but does contain most of the elements required to fill these standard elements.

Specimap, is a tool that can be used to georeference objects individually and requires a reasonable amount of manual input. However the high level of accuracy it can guarantee and the possibility to use historic data, like old maps, makes it suitable for objects with challenging textual localities. As the Royal Botanical Garden of Edinburg uses it no, users can search directly in the collection database, select an object and plot it on a map, to gather georeference data (lat- long coordinates) or correct already documented coordinates. This makes it usable for georeferencing projects as well as for updating a collection that is already georeferenced. It is at this moment not available for the public, but it will be released open source (personal communication, Elspeth Haston, October 6th 2015). Lastly the Tool is very useful because it shows all data at the same time, textual generated georeference data, the original geographic information, additional data like the collectors name and a plotted point on a map. This makes the entire context of an object visible.

The remaining Tool, **the Georeferencing Calculator**, does not function correctly, because it is based on a severely outdated java version (1.1). Both the web version and the standalone application do not have the feature to georeference batches of data. It is however the only tool that can handle additional aspects of geographic location descriptions like headings and offsets. The tool can also be adjusted to fit all the wishes a user has; the calculation type, the coordinate precision, the locality type, coordinate source and system, datum, etc. Secondly, the tool complies with the Darwin Core standard. All this combined makes the tool very useful for multiple stages of the georeferencing process like;

1. To calculate additional metadata for objects that have already been georeferenced/contain coordinates
2. After the batch georeferencing for the parts of the collection with challenging geographic information
3. After the first run with an automatic tool, to georeference objects that resulted in partial/multiple matches, or no matches.

In conclusion to georeference large batches of data the Google Geocoding API is the only useful tool for the first run of georeferencing. When looking back on the user needs, being able to georeference large batches at once, was important. However the accuracy and manual adjustment are also considered imported to keep the data quality high.

So if you want to work more accurate in the following step, to georeferencing objects that did not return the desired results, or require more manual adjustment the Georeferencing Calculator and SpeciMap could be used best. This is under the condition that these are updated and released. SpeciMap in this case is preferred, for objects because it combines all data one object in one screen. The Georeferencing calculator can be used for objects with many different aspects like headings, offsets etc. in the original locality.

Methods

To assist data mangers during the process, there are methods/ guidelines available that can lead to a uniform format for all records in a collection or dataset. Specifically for biodiverse collections and primary species data there are two guidelines available. These guidelines do not refer to manuals for the specific tools.

When it comes to the discussed georeferencing methods, it was very clear that there is not one that can be called ‘better’ for the focus of this research. Both guidelines discus very important (and similar) parts of the georeferencing process. The Guide to Best Practice for Georeferencing, by Chapman, focusses on the whole project, from acquiring GPS data, to the maintenance of georeferenced object data. This would therefore be very useful for institutions that do not have a georeferencing policy or project plan for this yet.’

The MANIS/ORNIS/HerpNET Georeferencing Guidelines are very useful for institutions that are only looking for practical information about how to deal with various elements in textual locality description. It provides clear examples and could be used directly during a georeferencing project. Both guidelines could be used side by side, where the Chapman guide would serve the beginning stage, of the project startup, and the MANIS guide could be used a reference for data mangers during the georeferencing process.

Resources

The data sources that are described are all online available databases and resources containing geographic information that is useful for georeferencing, like coordinates and place names. Considering the data in the collections, it is important that the resources contain worldwide geographic information, since specimen occurrences are also spread worldwide. Since most tools are online tools, it is most useful if the resources are also available online. An additional feature that would be an advantage is if the resource could handle fuzzy data (i.e. Miss spellings, historic names etc.).

	Geonames	**TGN**	**Google Maps**	**The Fuzzy G**	**Histograph**
Accessibility					
Online website	yes	yes	yes	yes	yes
Web service	yes	yes	yes	no	yes
Available as download	yes	Yes *	no	unknown	no
Free	yes	yes	yes	yes	yes
Geographic focus area	Worldwide (better with populated places	Worldwide	Worldwide	Worldwide	Netherlands
Results					
Data type results	Textual	Text with associative relationships, hierarchical and equivalence terms	Textual and visual	Textual	Textual and visual
Coordinates	yes	yes	yes	yes	no
Fuzzy data	Doesn't handle misspellings or historic data	Knows terms in different languages and historic names	With the ranking mechanism it can handle minor misspellings	Very good	Knows historic names from different periods of time

** Full data in JSON, RDF, N3/Turtle, N-Triples under (ODC-By) Open Data Commons Attribution License. An institution must acquire a license and pay the required fees in order to access the data, for annual releases of XML and relational tables. With the license, the data is also available via Web Services.*

When evaluating the table above it becomes clear that all 5 data resources can be used for similar purposes but they differ greatly in their original structure and features.

The **Geonames** database is a web service that contains the names of over 7 million places, like cities, countries, areas, important buildings, parks etc. and additional data like coordinates and hierarchal places. It is one of the largest databases available and because of the coordinates very useful for georeferencing. It does however have problems with misspelling and historic names. Geonames only finds matches for places that have the exact same spelling as the one recorded in their database. It can therefore only be used for georeferencing object records with standardized locality terms.

The Getty Thesaurus of Geographic names can be used online as well as offline (when integrated in a CMS). It is not just a database, but a thesaurus and knows associative relationships, hierarchical and equivalence terms. So it does not only find the named place from your query, but also the higher hierarchal terms. Lastly the TGN also knows historic place names, from different periods of times. So when you look for Amsterdam, it will also return associative names from for example the Roman or Celtic time period. This resource is only one in this list that has that last feature for worldwide data.

Google Maps is one of the most used geographic resources and it knows place names, rivers, bus stops and many more units with a geographic location. The most important advantage of Google Maps it that it fully matches with the georeferencing tool The Google Geocoding API. The user actually does not have to undertake any additional steps to use this resource, because the API already uses the information from this resource. Google maps can show results visually on high resolution satellite imagery, street maps and 360° panoramic views and return the results textually. Because of its built in ranking mechanism, it returns more important places first. This means it can handle some minor misspellings, for example: [Amtedam will still lead to Amsterdam].

The Fuzzy G is the only gazetteer who can still find matches for textual localities with very “fuzzy data”. The tool knows different degrees of “fuzziness” ("Very Fuzzy," "Quite Fuzzy," "Default Fuzziness," "Not So Fuzzy" and "Only Exact Matches”) which users can select based on the estimated quality of the original data. The fuzzier the data the more it focusses on vowels. However the Fuzzy G can only be used for 1 object at the time. This makes it not suitable for georeferencing batches.

Histograph is a tool specifically for legacy collections with historic place names. It combines different archival datasets and combines these with modern geographic resources like the TGN and Geonames. Unfortunately it now only works for DUTCH data, but the techniques that are used for this tool, would be useful for different countries. Due to its limited geographic focus area it is not useful for most biodiverse collections and institutions.

Besides the Histograph tool and the Fuzzy G, all resources could be useful for georeferencing, but it fully depends on the quality of the original data which one is best fit. 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. Due to the ranking mechanism it still handles minor deviations. In conclusion: an institution can decide to standardize the datasets before georeferencing and use Geonames, or the TGN when the dataset contains standardized historic names or the institution can use Google Maps does not use standardized datasets.

Datacleaning methods

Datacleaning regards the improvement of the original geographic locality description from a collection to increase the percentage of georeference results. Datacleaning can be done manually, by an expert or with the help of a tool, but in most cases it requires a reasonable amount of man hours. There are several types of datacleaning;

Light labor-intensive datacleaning regards the correction activities that can be done in little time and with few man-hours, but that can increase the amount of good results for the georeferencing first run. Examples are: adjusting or removing troubling characters, deleting double names, filtering unique localities. Heavy labor-intensive datacleaning is the improvement of aspects of the original geographic location descriptions, that disrupted the georeferencing tool and therefore resulted in no matches or partial/multiple matches. Examples are: headings, offsets, postal codes. With Very heavy labor- intensive datacleaning each object record is evaluated and georeferenced manually. This requires expertise and allot of man-hours. Therefore this should only be done for objects with data that are still not recognized by georeference tools after the first two datacleaning methods have been tried.

	Visual Check Google API	**GBIF**	**Specieslink**	**Plaatsnamen standaardiseren**	**Open refine**
Accessibility					
Online	yes	yes	yes	yes	Yes
Offline	no	no	no	No	
Free	yes	yes	yes	yes	yes
File formats					
Accepted import formats	HTTP request with Open Refine or similar	DarwinCore Archive files, offers converter for MSExcel files	unknown	CSV.	TSV, CSV , Excel XML, RDF as XML, JSON, Google Spreadsheets, RDF N3 triples
Output formats	JSON or XML	No output possible	unknown	CSV.	Same as import
Batches	Creating strings in batches, but viewing individually	yes	yes	yes	Yes
Individual	yes	no	yes	no	Yes
Results					

Type of results	New (more accurate) coordinates	Highlighted mismatches in a dataset	Highlighted error	4 tabs; 1) standardized localities 2) localities with multiple matches. 3) Localities with no results, 4) intended for localities that can't be standardized by expert or tool.	Various, depends on the formulas and features used. examples: structured data, new formats, parsed data, unique localities, georeference http strings
Manual input required	yes, string has to be adjusted manually for a better results	yes, GBIF does not adjust, just highlight	yes, does not adjust, just highlights errors.	Yes, selecting he field that contains the place names and what type of locality it contains	Yes, but only for selecting/creating columns when importing data
Compliance with standards	no	yes, complies with Darwin Core	yes, Complies with Darwin Core	no	Depends on Google Geocoding API and column in dataset
Geographic focus area	Worldwide	Worldwide	Brazil	Netherlands	Not applicable. Depends on tool that is used.
Efficiency	Requires manual input to adjust each string	Low, does not correct errors, only highlights them	Low, does not correct errors, only highlights them	High, only requires manual input to create csv file.	High, requires little manual input
Reliability	Low, adjustment depends on manual input of a user	High, backbone of GBIF is consistent	unknown	High, data sources (Geonames and TGN) are consistent	Only uses imported data, so it depends on the tool that is used and the original data.
Transparency	High, strings are based on Google geocoding API out.	Low, documentation not available. It is unknown how the data check works exactly.	unknown	High, documentation http://erfgeo.nl/wat-hoe/standaardiseertool.html	High, is open source: https://github.com/OpenRefine/OpenRefine/wiki/Documentation-For-Users

The first conclusion that can be made, is the **SpeciesLink** and **GBIF** are very similar. They come from similar web services, international data portals with open source biodiversity data, and with similar datacleaning methods; highlighting possible errors. The biggest difference is that GBIF focuses on data worldwide and SpeciesLink is only intended for data from Brazil and with Brazilian locations. GBIF is therefore the most useful one of the two tools, since it can be used to check all collections, regarding the geographic occurrence area. GBIF can only be used to find errors, since it does not improve the results itself. The checked datasets can be exported and adjusted by the data owning institution. Most of the time a detected error, like a country coordinate mismatches, regards multiple records, with the same locality and same error. By adjusting one, you could adjust all, for example in Excel sheets. The error checks are done automatically by GBIF, when an institution publishes their data, so will only cost an institution time that is required for exporting and improving the data, not in finding the errors.

The **Visual Check of the Google API**, is a datacleaning method that could be used for the labor intensive datacleaning process. By using the returned strings from the Google geocoding API run, a user can check the results visually on the Google Maps application, and adjust the string manually for a more accurate result. It still requires some degree of manual input, but increases the amount of object records with good georeference data.

Plaatsnamen-Standaardiseren is a tool that can be used to standardize locality terms in a dataset or collection. This can be very useful before the first run of a georeferencing tool, because it improves the data quality and therefore the amount of object records that return with good georeference data the first time. It can also be used to decrease the amount of unique localities before the first run. The Plaatsnamen-Standaardiseren tool, only works for Dutch geographic information. This means it is not useful for most biodiverse collections.

Lastly **Open Refine** is very useful for corrections. This can be used to prepare datasets for the first (or second) run with a georeferencing tool. It works similar to Excel, which is also a good option, and rather easy to use. With the help of various features and formula user can:

- Create unique localities
- Delete troubling characters that could cause problems with tools, like comma's, dashes, headings, etc.
- Create separate fields for all parts of a locality description, like country, state
- Create search strings/ HTTP request for the Google Geocoding API

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.

For this reason in chapter 5 crowdsourcing is discussed. This not a tool to do datacleaning with software, but it is a way to reduce the amount of man-hours needed to do so. By using the crowd for adjusting the detected errors of the GBIF check, or to adjust the strings for better fitted results with the visual Google check, it would save the institution time and deliver more results.

8. Conclusion

Of the relatively long list of tools, methods, resources and datacleaning methods, only a small amount of useful possibilities for georeferencing primary species data remains after the tool selection process. This mainly has to do with de availability of tools for this particular topic. Georeferencing is a subject that has been researched and tested by many institutions and projects. However most projects or institutions start from the beginning, without building on experiences and data that already exists. As can be seen in the tool chapter, many institutions design their own georeferencing tool, but these are not available for the public and in some cases no longer available after the initial project ended. So, there is not that much available open source, free and online for all interested parties to use.

Based on the tool selection it can be stated that the Google Geocoding API is the most usable tool for georeferencing primary species data. It is currently the only tool that can georeference objects in batches, is easy to use the help of Open Refine, and perfectly matches the datacleaning method of the Visual Check with Google Maps. Even though the tool does not fully comply with international data standards, the generated metadata from the Google Geocoding API can be mapped to these required standard elements. This combined is the best match with the user needs defined in chapter 4.

Since Naturalis also used the Google geocoding API, for georeferencing their collection, the process that is described here will be very similar to the one they developed. The process showed below is based on the process Naturalis described in their final report, but as expected the Google Geocoding API can be complemented with other tools and additional resources to complete the georeferencing process. This combined, leads to the following process:

Step	**Tool**
Light labor intensive datacleaning to increase the efficiency	➢ Open Refine: With the use of features like facets, text filters, transformations etc. The amount of unique localities can be reduced and characters that might result in errors can be deleted or replaced. ➢ Microsoft Excel: with various formulas it is possible to filter unique localities, replaced and strange characters can be removed or replaced. **For Dutch collections** ➢ Plaatsnamen- Standaardiseren: can be used to standardize specific features of locality description in a dataset. This can be done in batches. ➢ Histograph: Specifically for standardizing historic names.
1e run of set unique locations	➢ The Google Geocoding API: for offering batches of HTTP requests via open refine to the Google geocoding API.
Heavy labor intensive datacleaning	➢ Visual check with Google Maps: to manually adjusting the retuned strings with Google maps. For this step crowdsourcing can be used, to reduce the amount of man-hours needed for the data owning institution.
2e run with remaining localities	➢ The Google Geocoding API for relatively simple textual descriptions of the remaining results after the 1st run and 1st datacleaning effort. **If available** ➢ The Georeferencing Calculator for relatively challenging textual descriptions that contain headings, offsets etc. ➢ SpeciMap: to manually replace the found georeference point to a more accurate point.
Calculating accuracy	➢ Viewport of the Geocoding API: by calculating the distance between the two given coordinates of the outer southwest and northeast corners of the smallest geographic unit of the address component the Geocoding API found (viewport). ➢ Georeferencing Calculator: Calculating uncertainty in meters for objects

	which originally had distances and heading in the locality description
Very heavy labor intensive datacleaning and checking the results	➢ Expert validation: experts within an institution can validate (or adjust) georeference data for objects that still have questionable georeference data, or do not have any georeference data. This has to be done manually, but can be done for objects with sensitive locality data, that cannot be made public for crowdsourcing **To reduce man-hours for this process** ➢ Crowdsourcing
Update or validation	➢ GBIF Validation: when publishing data on the data portal the web service checks and validates georeference data. This can be used to detect possible errors ➢ Crowdsourcing: to adjust the object that were highlighted with the GBIF check.

In the process described above, the SpeciMap tool and the Georeferencing Calculator are mentioned for a few steps. These tools show very promising, but is unfortunately are not yet or no longer available for public use.

The institution that created the SpeciMap tool did mention public release in the near future. When combined with the Google Geocoding API, this would form the perfect match and be used following on each other. Combination of these tools (combined in 1 tool or following each other) automates the georeferencing process more and leads to more accurate results.

The Google API could be used for bulk georeferencing and the SpeciMap tool for the more accurate adjustments of the retuned strings. They could however also be combined in one tool. If the SpeciMap would have the feature to georeference batches of data based on the text strings, and plot these on maps, user could then manually adjust the result with the tool. This way collections could be georeferenced with one tool and institutions would only need additional sources or datacleaningtools for preparing data before georeferencing, or to further adjust the really difficult objects that require expert input.

The georeferencing Calculator could also add a lot of value to the georeferencing process described above. It is the only tool that can handle additional aspects of geographic location descriptions like headings and offsets. The tool can also be adjusted to fit all the wishes a user has; the calculation type, the coordinate precision, the locality type, coordinate source and system, datum, etc. Additionally the tool generated metadata that fully matches the Darwin Core standard. If this tool were to be updated to a more current version of java (8), this would be a very promising option.

The most important conclusion that can be derived from all these experiences and projects, is that the best usable tools, the Google Geocoding API, SpeciMap and the Georeferencing Calculator, are already created, they just need to me be maintained and managed for a long period of time and for the entire public.

There is already enough knowledge and tooling created for georeferencing primary biodiversity data. The big problem is however, that most institutions only create these tools and produce this knowledge for themselves and after a project is finished, this disappears. There is no long term maintenance budget and the application the software becomes outdated. An international effort can contribute to preserve this knowledge, and tooling, so that not every institution needs to invent the wheel.

Various institutions of the CETAF have been working on projects and researches to find and create guidelines, tools and processes for improving the georeferencing of primary biodiversity data. This document attempts to bring some of these projects together and summarizes the available information on this topic in order to create a best practice that is internationally applicable and recognizable.

Best Practice

The best practice, derived from this document, describes the complete process with detailed information also regarding dataset preparation, import formats and usable features of all tools.

This can be used as a roadmap for biodiverse institutions planning a georeferencing project. This process is applicable to other interested institutions of the CETAF, which ultimately can form the bases for an international standard method for (semi) automatic georefencing of natural history collections. This includes guidelines for the entire process that is involved semi-automated georeferencen and suggestions for useable tools and resources. Application of the best practice could lead to an enrichment of natural history collections with reliable and comparable georeferenced data. Additionally such a standard method could increase the usability and quality of digital natural history collections.

Secondly the research document also describes various georeferencing methods and projects, regarding their use of publicly available tools. It could therefore also be used by institutions to derive ideas from and possibly form new collaborations. The process described is a standardized method, applicable for most institutions and projects, but expansion or additions based in institution related ideas, policy and data quality is possible.

9. Discussion

In this discussion part of the research document the last few subjects, that arose during some parts of the research are discussed, like prioritization, limitations of publishing georeference data, the truth levels of the results that georeferencing tools can generate and three datacleaning methods, that were not included in the tools selection process, due to availability. The subjects discussed were not been studies extensively but are imported to remember when starting a georeferencing project or could lead to further research, testing or development.

Prioritization

An aspect that is very important for georeferencing a collection for research purposes is prioritization. With all aspects of the digitization of a collection, and therefore also with georeferencing, there is always a conflict between quality and quantity, between which you have to find a balance. With collections that are as large as those of Naturalis, it is sometimes not possible to secure the highest level of quality for all objects (quantity). In georeferencing this means a decision will have to be made whether you want to georeference the bulk of the collection to an accuracy of 10 or 20 km georefereerd or if you want to focus on the highly accurate georeferencing of some parts of the collection.

At this moment there is a particular need for georeferenced data from large parts of the collection objects. However, the georeferencing of the bulk of the collection objects means that there is less time to evaluate every location description and georeference these with a very good accuracy. The smaller the accuracy generally means the more datacleaning and manual input is required, and would therefore cost more time.

Another evaluation you have to make is for example, if your first georeference areas that are considered data- poor (very few geographic information is known) or if you georeference areas that are very data rich (a lot of geographic information is known). Of the data poor areas, such as tropical rainforest's, the collection data usually contains localities with very few named places, and the ones that are named mostly concern historical or local names, which are not found in the standard gazetteers. For data- rich areas such as the Netherlands a lot of geographic data can be found. Based on the use case of Naturalis, and specifically the researchers here, the preference goes to georeferencing data poor areas. Here the georeference data provides a greater improvement in data quality (Personal communication, 8 October 2015). Georeferencing locality descriptions with historical or local names is however more challenging and would take more time.

When analyzed from aspects of findability and publication, the main preference goes to georeferencing as many specimen records as possible. This will probably lead to rather inaccurate results, but the object will be searchable based on countries. Fine-tuning for a better accuracy can be done later.

Which choices an institution makes depends greatly on the type of research questions it wants to support with their collection and whether or not they want to support the majority of questions, or focus on the more specific ones. Generally most research question could do with the advised accuracy of 25 to 10 km, but there are some research questions, that would require a smaller accuracy (Personal communication, 8 October 2015).

Publishing georeference data

This research document mostly discusses requirements and the difficulties that come with the documenting and recording georeference data in a collection, but publishing georeference data also comes with some requirements and difficulties.

This specifically applies to sensitive species data, meaning specimen data of animals and other species on the red list of the IUCN, or in preserved areas, etc. GBIF published an international best practice discussing the sensitivity of the exact localities of rare, endangered species or species of commercial value and how to deal with this data: *The Guide to Best Practice for Generalizing Sensitive Species Occurrence Data* (Chapman, Grafton, 2008). Complying to this standard it is possible that biodiverse institutions sometime have choose to publish a larger accuracy that that is recorded in their own collection. Naturalis does this for example when publishing specimen data of animals on the red list of the IUCN, on Wikipedia. With these

species, only the country and possibly the city or region is indicated (Personal communication Hans Muller, 2015).

The Bets practice states that if it does not concern sensitive specimen occurrence data, "*Biodiversity information should be made freely available to be shared globally to enable their use for not-for-profit decision-making, education, research and other public benefit purposes*" (Chapman, Grafton, 2008).

How to use and implement this best practice is difficult for various institutions. For example: When a collector collects different specimens, both endangered and not endangered, during one trip: people can still find exact localities when using the dates of the collection events. More tooling and collaborations on the implementation of these principles could be beneficial for an international standard.

Truth levels georeference results

The Google Geocoding API returns georeference data after a HTTP request with a confidence score based on:

- if a unique locality was found
- if this is a full match with the data from the original HTTP request.

Based on a test round Naturalis executed with set 190 object records, the Google API proved very effective since it returned all objects with full matches and correct georeference data. However, the assumption that all results with a good confidence score and a full match are correct, can be questionable. This goes especially for locality descriptions that contain historic names, or don't contain a country name.

For example; *A locality description contains the historic name of one of the islands of Indonesia, like Alkmaar, but there is no country recorded in the collection. Since [Alkmaar] is the only named place in the original data, most georeferencing tools will return coordinates for [Alkmaar, Netherlands] as a result, because this is the only current named place in most gazetteers that matches. It is however not the correct named place.*

This is due to the built in ranking mechanism that the Google Geocoding API has, that focusses on well known, and populated places. If you use Alkmaar, Indonesia in the HTTP request, Google will find the result you want.

This 'problem' decreases the truth level of full matches. Which means it has to be considered if it is necessary to also have the full matches checked for these types of misinterpretations and validated by an expert? This can for example be done only for objects that miss a country name in the original locality description. Another possibility is to have objects checked and validated of which it is known that they were collected before a certain year or date (or buy a specific collector), in an area that has changed names since then, like former colonies.

Checking a dataset with objects for missing country names, or misinterpreted georeference results can be very difficult and time consuming, because it is hard to spot these types of anomalies in numerical coordinates in large batches. A way to make this easier is to this in an environment where al data, original geographic data, georeference data, and the visual location on a map, can be seen at once, for example, with the possibility to plot object directly on maps in a CMS.

This way an expert within a biodiverse institution can work collaboratively on checking and validating georeference data in a sort of knowledge based process. When doing this, it could be useful to add an extra metadata element in which the status of the validation process is recorded.

Crowdsourcing

A validation process as described above, can of course, also be executed with the help of the public, in a crowdsourcing environment. This could save the data owning institution allot of time. These options do however, assume that an institution has access or the possibility of creating a crowdsourcing platform and a population of citizen scientist (van Erp et al, 2014). This does require time and money to create, but generally a crowdsourcing project delivers more than just the data the institution gathers. It creates a higher

involvement of the crowd in biodiversity, it can bring new visitors and even generate publicity with the project (Blaauboer, 2014).

Crowdsourcing could also be used in various other stages of the georeferencing process. when going back to the table in the conclusion, crowdsourcing could be used to aid in the following steps:

- Filtering unique localities: this can be done with formulas in Excel or with various features in Open refine, but these methods only match unique locations with the exact spelling. But in most collections multiple ways of spelling or combinations of named places refer to the same locality. For example: [Box Hill, Box Hill; Surrey, Box Hill, Kent, Box Hill; near Dorking, Box Hill; Dorking] all refer [Box Hill; Surrey; UK; 51.254 N, - 0.308W]. Matching these together can decrease the amount of unique localities even more. The crowd could link locality description that refer to the same place or by linking place names to thesaurus terms.
- Labor intensive datacleaning tasks with Google Geocoding API: this requires allot of time but it can improve the percentage of georeferenced records and can create better quality georeference data. A crowdsourcing project could be deployed to manually adjust the individual objects strings for the Visual check by Google.
- Very heavy labor intensive datacleaning: with the help of the crowd objects that still don't have a georeference after all process steps and require complete manual input to get results can be done individually.

Visual Possibilities

In addition to the tools, that are described in the conclusion, there is a possibility to develop an interface as an extra layer over the tools. The Google geocoding API is, for example, quite technical and is not that easy to use for everyone. The National History Museum of London designed an interface for user to make the georeferencing process with this easier.

By creating a more user friendly tool/ method for georeferencing the group of users can be expanded and the workload could be spread out more. For a user friendly interface a user would not have to need the same knowledge or experience they would need for tools like open refine or the Google geocoding API, and more people would be able to help. Building you own interface also provides the ability to choose the fields and functionalities that are most important to your institution or project. This could also be useful for the expert validation step within institutions, or even for a crowdsourcing project.

Histograph, Plaatsnamen- Standaardiseren and MITCH

Histograph, Plaatsnamen- Standaardiseren and MITCH are three tool/ methods that were described in tool selection process. Unfortunately none of these is useful at this moment, because they are not available or only have a very small geographic focus area. The techniques that these tools are based on are however very interesting and pro missing for biodiverse institutions.

The Histograph[34] is a tool designed in and for Holland, for searching and standardizing historical place names in a dataset. At this moment the Histograph tool uses data sources with; birth places of Dutch East India Company crew members, monastery records and historical census data for the historical place names, and the Geonames and TGN for the standardized modern place names (Erfgoed & Locatie, 2015). These last to data sources, the TGN and Geonames, contain worldwide geographic data, but this tool focusses on the Dutch data. The techniques are open source and could be used for the same purposes for different areas or countries, by adding different historical geographic data.

Plaatsnamen Standaardiseren[35] is an online tool that can be used to standardize place names in a dataset. This tool could be very useful for the data cleaning process before the actual georeferencing is done to generate a higher percentage of full matches, during the first run.

[34] http://
.io/viewer/
[35] http://standaardiseren.erfgeo.nl/

The tool also only works for Dutch geographic information, and does not know any coordinates. However the tool uses the TGN and the Geonames databases to generate the standardized terms (Erfgoed & Locatie, 2015). Just like the Histograph a similar tool, with additional (country) specific geographic data sources could be interesting for institution in other countries that Europe.

The MITCH project used different, well known, text analysis and data mining techniques to georeference object records in a biodiverse collection. The prototype tool that was created during this project is no longer available, but the techniques can still be applied in other environments. There were several articles published on the techniques that were used, and how they were used.

The main advantage of these techniques is that it looks at the collection data and the georeferencing process from a completely different level, than most automatic tools do. Some of these techniques focus on coherence of clusters of data instead of separate data points. This could complement the process described in the conclusion, which only automatic uses tools.

References

Internet source

Biocase (2005). *Biological Collection Access Services*. Retrieved November 2, 2015 from http://www.biocase.org/

BioCase (2005). SYNTHESYS NA-D 3.7- Providing itinerary related datasets and tools (for integration, visualization and quality check). Retrieved November 2, 2015 from http://www.biocase.org/products/geo_services/itineraries/

Blaauboer, R. 2(2014). *Nichesourcing helpt het Rijksmuseum collecties in kaart brengen*. Geraadpleegd op donderdag 20 november 2014 http://www.frankwatching.com/archive/2014/03/20/nichesourcing-helpt-het-rijksmuseum-collecties-kaart-brengen/

Blaauboer, R. 1 (2014) *Digitaal erfgoed: zo past crowdsourcing in je bedrijfsmodel*. Retrieved on November 20th 2015, from http://www.frankwatching.com/archive/2014/11/03/digitaal-erfgoed-zo-past-crowdsourcing-in-je-bedrijfsmodel/

Engelse, M. den (n.d.) Handleiding Standaardiseertool. Retrieved on November 19th 2015 form, http://erfgeo.nl/wat-hoe/standaardiseertool.html

Enipedia TU Delft (2015). Open Refine Tutorial. Retrieved on November 17the 2015, from http://enipedia.tudelft.nl/wiki/OpenRefine_Tutorial

Erfgoed & Locatie (2015).Demo-versie Historische Geocoder online. Retrieved on November 18th 2015, from http://erfgoedenlocatie.nl/2015/04/demo-versie-historische-geocoder-online/

GBIF (2010). *Darwin Core Archives – How-to Guide, version 1.* Released on 1 March 2011, Copenhagen: Global Biodiversity Information Facility, 21 pp

Github (2015). Geocoding, Translate Street addresses to lat/lng coordinates. Retrieved on November 17the 2015, from https://github.com/OpenRefine/OpenRefine/wiki/Geocoding

Google Developers (2015). The Google Maps Geocoding API. Retrieved on November 17th 2015, from https://developers.google.com/maps/documentation/geocoding/intro#GeocodingRequests

HerpNET (2007). Georeferencing – Step by Step & Online Resources. PowerPoint Presentation available at: http://HerpNET.org/HerpNET/documents/StepbyStep.ppt.

Hine, A. (2015). ICollections, Mass Digitization of British & Irish Lepidoptera. Retrieved November 3, 2015, from http://slideplayer.com/slide/3541251/

Histograph (n.d.) Histograph: geocoding places of the past. Retrieved on November 18th 2015, from http://histograph.io/

Llewellyn, C.A. & Hasten, e. & Grover, C. (2011). *Georeferencing Botanical Data using Text Analysis Tools.* TDWG 2011 Conference, retrieved on October 15th 2015, from http://www.researchgate.net/publication/282701713_Georeferencing_botanical_data_using_text_analysis_tools
Llewellyn, C. & Grover, C. & Oberlander, J. & Haston, E. (2012). Enhancing the Curation of Botanical Data Using Text Analysis Tools. *TPDL 2012, LNCS* 7489, pp. 480–485

Natural History Museum London (2015). ICollections.Retrieved on November 3, 2015, from http://www.nhm.ac.uk/our-science/our-work/digital-museum/i-collections.html

Naturalis Biodiversity Center (2015a). *Wat wij doen: onderzoek*. Retrieved on September 4th 2015, via: http://www.naturalis.nl/nl/over-ons/wat-doen-wij/onderzoek/

Oosterman, J. & Nottamkandath, A. & Dijkshoorn, C. & Bozzon, A. & Houben, G.J. & Aroyo, L. (2014) Crowdsourcing knowledge-intensive tasks in cultural heritage. (Online publication) WebSci 2014: 267-268, https://sealincmedia.wordpress.com/publications/

Pulla. P (2013). Indian ecologists turn to crowdsourcing. Retrieved November 20th 2015, from http://www.natureasia.com/en/nindia/article/10.1038/nindia.2013.152

Ridge, M. (2007) *Sharing authorship and authority: user generated content and the cultural heritage sector*. (Online publication) 2007 Web Adept - UK Museums on the Web, Museums Computer Group conference, http://www.miaridge.com/projects/usergeneratedcontentinculturalheritagesector.html

Specimenlink (2008). SpeciesLink, Information about the Project. Retrieved on November 19th 2015, from http://splink.cria.org.br/project?criaLANG=en

TDWG (2010 A). ABCD - Access to Biological Collection Data. Consulted on September 14th 2015, http://wiki.tdwg.org/twiki/bin/view/ABCD/WebHome

TDWG Wiki (2010 B). *Geospatial Extension Concept List*. Consulted on September 14th 2015, http://wiki.tdwg.org/twiki/bin/view/DarwinCore/GeospatialExtension

TDWG (2007). *Mapping between Darwin Core 1.4 concepts (DwC) and ABCD 2.06b*. Consulted op 14th of September 2015, http://www.bgbm.org/TDWG/CODATA/Schema/Mappings/DwCAndExtensions.htm

The J. Paul Getty Trust **(2015).** Getty Thesaurus of Geographic Names Online**.** Retrieved on November 18th 2015, from http://www.getty.edu/research/tools/vocabularies/tgn/about.html

Tilburg University. (2005). MITCH, Mining for Information in Texts from Cultural Heritage. Consulted on November 9th 2015, from http://ilk.uvt.nl/mitch/

Wieczorek, J., D. Bloom. 2011. Georeferencing Calculator Manual v2. Retrieved on November 18th 2015, from http://manisnet.org/GeoreferencingCalculatorManualv2.html

Wikipedia (2015a). R, programming language. Retrieved on November 17th 2015, from https://en.wikipedia.org/wiki/R_(programming_language)

Wikipedia (2015b). Google Maps. Retrieved on November 18th 2015, from https://en.wikipedia.org/wiki/Google_Maps

Willemsen, J. (2011). *Erfgoed en Crowdsourcing*. Geraadpleegd op vrijdag 3 oktober 2014, http://www.frankwatching.com/archive/2011/02/28/erfgoed-en-crowdsourcing/

Interviews

Personal communication, 23 September 2015, Luc Willemse, Collection manager

Personal communication, 25 September 2015, Ruud Altenburg, developer BioPortal

Personal communication, September 29th 2015, Agnes Kirchhoff, Berlin-Dahlem Botanical Garden and Botanical Museum

Personal communication, September 30th 2015, Cees Hoff, Dutch Node Global Biodiversity Information facility

Personal communication, October 6th 2015, Elspeth Haston, The Royal Botanical Garden of Edinburg

Personal communication, 6 October 2015, Hans Muller, Wikipedian in Residence

Personal communication, 8 October 2015, Niels Raes, in- house researcher

Personal communication, October 15th 2015, Patricia Mergen, The Royal Museum of Central Africa Tervuren

Personal Communication, October 29th 2015, Malcolm Penn, the Natural History Museum of London

Internal Documents

Arkel, B. van & Creuwels, J. & Leusen, J. van & Schnörr, S. (2014). *Eindrapport FCD- Pilot Georeferencing: productiematig georeferencen*. Internal document Naturalis, Leiden

Hine, A. & Penn, M. (2014).Georeferencing at the NHM. Presentation. Naturalis Biodiversity center, Leiden

Naturalis Biodiversity center (n.d.a). *Botanische gedachten BRAHMS vs BOLD*. Intern document, consulted on September 17th 2015

Naturalis Biodiversity center (2012). *Collectieplan Naturalis Biodiversity center 2013-2016*. Intern document, consulted on September 15th 2015

Naturalis Biodiversity center (2014a). *Follow-up Actions Meeting Automated Georeferencing.* Internal document, consulted on September 2nd 2015

Naturalis Biodiversity center (n.d.b). *Help Datamodel CRS*. Intern document, consulted on September 15th 2015

Naturalis Biodiversity center (2013). *Registratie*. Intern document, consulted on September 16th 2015

Naturalis Biodiversity Center (2013) Beleidslijn Open Content. Intern document, consulted on November 13th 2015

Naturalis Biodiversity center (2014b). *Technical Summery, Automated Georeference meeting December*. Intern document, consulted on September 2nd 2015

Schnörr, S. (2014). *Georeferencing Methode Niels Raes*. Summery. Intern document, Consuleted on November 7th 2015

Published articles and books

Chapman, A.D. & Wieczorek, J. (2006). *Guide to best practices for georeferencing*. Copenhagen: Global Biodiversity Information Facility

Chapman, A.D. and Grofton, O. (2008). Guide to Best Practice for Generalizing Primary Species-Occurrence Data. Copenhagen: Global Biodiversity Information Facility

Chrisman, N.R. (1983). The role of quality Information in the long-term functioning of a GIS. *Proceedings of AUTOCART06*, 2: 303-321. Falls Church, VA:ASPRS

Erp, M. van & Hensel, R. & Ceolin. Davide. Meij, M. van der(2014). Georeferencing Animal Specimen Datasets. *Transactions in GIS,* 12/2014: 19(4)

Fleurbaay, E. & Eveleigh, A. (2012) *Crowdsourcing: prone to error?* (Online publication) International Council on Archives Conference 2012, http://ica2012.ica.org/files/pdf/Full%20papers%20upload/ica12Final00271.pdf

Grover, C., Givon, S., Tobin, R., Ball, J. (2008). Named Entity Recognition for Digitized Historical Texts. In. Proceedings of the Sixth International Conference on Language Resources and Evaluation

Meganck, B. & Meirte, D. & Mergen, P. & Theeten, F. (2006). *Milestone report for SYNTHESYS Network Activity: NA-D 3.7 Providing itinerary related datasets and tools for integration, visualization and quality check.* Retrieved November 2, 2015 from http://www.biocase.org/products/geo_services/itineraries/files/SYNTHESYS_milestone_report_may31.pdf

Murphey, P. & Guralnick, R. & Glaubitz, R. & Neufeld, D. & Allen Ryan, J. (2004). *Georeferencing of museum collections: a review of problems and automated tools, and the methodology developed by the mountain and Plains Spatio-Temporal Database-Informatics Initiative.* PhyliInformatics, volume 3, 1-29

Naturalis Biodiversity Center (2015b). *Uit het Depot, op het web: Twee eeuwen nationaal natuurhistorisch erfgoed in het digitale domein.* Leiden: Naturalis Biodiversity center

Rios, N.E. & Bart, H.L. (n.d.).User Manual, GEOLocate Georeferencing Software. University of Tulane, Bell Chase. Retrieved on November 17yh 2015, from https://www.museum.tulane.edu/geolocate/standalone/manual_ver2_0.pdf

Spencer, C. & Yamamoto, K. & Fang, J. & Constable, H. & Koo, M. & Wieczorek, J. (2005) Georeferencing for Dummies: An Elaboration of the MaNIS/HerpNET/ORNIS Guidelines. Berkeley: University of California. Available online at www.HerpNET.org/HerpNET/documents/georeffordummy.xls

Verhoeven, N. (2014). *Wat is onderzoek? Praktijkboek voor methoden en technieken.* Amsterdam: Boom Uitgevers

Wieczorek, J. (2001)"*MaNIS/HerpNET/ORNIS Georeferencing Guidelines*. Retrieved November 2, 2015 from http://manisnet.org/GeorefGuide.html

Wieczorek, J. & Döring, M. & De Giovanni, R. & Robertson, T. & Vieglais, D. (2015). Darwin Core Terms: A quick reference guide. Biodiversity Information Standards – TDWG: Brussels

Tools and applications

Christian Kohlschütter (2003). The JRC Fuzzy Gazetteer. Retrieved on November 18th 2015, from http://isodp.hof-university.de/fuzzyg/query/

Geonames (n.d.). Genoames, data resources. Retrieved on November 19th 2015, from http://www.geonames.org/data-sources.html

GitHub (2011). *StanDAP-Herb WebGenesis Web Services*. Retrieved 2 Nov. 2015 from https://github.com/gentisaliu/StanDAP_Herb_WebGenesis_WebServices

Google (2015). Geocoding service - Google Developers Retrieved November 2, 2015, from https://developers.google.com/maps/documentation/javascript/examples/geocoding-simple

Tulane University (2015). "GEOLocate - Software for Georeferencing Natural History Data. Retrieved November 2 2015, from http://www.museum.tulane.edu/geolocate/

Best Practice

An International best practice for georeferencing primary specimendata in a (semi-) automated process

Josine Blom

Josine Blom
Naturalis Biodiversity Center
Marian van der Meij

Naturalis Biodiversity Center

Darwinweg 2
Postbus 9517
2300 RA Leiden

T 071 751 91 02
josine.blom@naturalis.nl
www.naturalis.nl

Table of Contents

Introduction

Precise geographic information is essential for harvesting collections for research on biodiversity. The taxonomic data (such as species names), and the coordinates of objects are the main types of information needed by a researcher to research the distribution and occurrence of species. This is also the most widely used data of these collections. But also collection management and accessibility of the collection could benefit from georeference data. Linking the corresponding coordinates to the verbatim location descriptions can increase the usability and accessibility of the collection. With coordinates you can directly plot species on maps, research the changes of the occurrence of the species and combine it with other data such as climatic or environmental data to examine the impact of this species.

Many institutions and many experienced georeferencers have been working on tools, guidelines and standard methods for georeferencing primary biodiversity data. Each developing their own preferences for the order in which they georeference.

In order to share knowledge on these initiatives and experiences and to prevent that every institution creates its own standard method or tool, Naturalis Biodiversity Center, and other partners of the CETAF, concluded that an international best practice should be created. This can contribute to an international standard method for (semi) automatic georeferencing and an infrastructure for all natural history collections within the EU for the future

This best practice is based on a research on a large amount of these projects tools, guidelines and standard methods. Application of the best practice should lead to an enrichment of natural history databases with reliable and comparable georeferenced data.

The document provides guidelines to a best practice for semi- automated georeferencing. This best practice will gives an overview of currently available tooling and possibilities to reduce the amount of man hours needed for georeferencing. It will not answer all questions, as many are institution specific, but it can be used as a roadmap, inspirational document or to create new guidelines, tools and collaboration for the questions, that are not fully answered yet.

Documents like '*Guide to best practices for georeferencing*' (Chapman, 2006), '*MaNIS/HerpNet/ORNIS Georeferencing Guidelines'* (Wieczorek, 2001) and '*The point-radius method*' (Wieczorek, 2004) can be used in addition to this best practice for information and best practices on how to georeference a range of different location types, and on how to determine the extent and maximum uncertainty distance for locations based on the information provided.

1. Pre- georeferencing process

There are a number of issues that will need to be addresses by any institution before starting a georeferencing project. These issues regard various aspects of the institution, collection and data that are involved with the georeferencing process. In this chapter all these issues will be addressed:

- What is the condition of the collection and what collection specific aspects are important to examine before georeferencing?
- What requirements are there for the institute's data model to enclose georeference data?
- For what purpose will the newly gathered georeferenced data be used, and by whom?
- What information regarding the institution is needed to create a georeferencing protocol?

1.1 Collection (data) specific

Georeferencing is a part of maintaining a biodiversity collection that leads to a more usable and durable natural history collection. However, converting textual locality description recorded in these collections into matching geographic reference point, like coordinates, is not that simple.

The most difficult issue in georeferencing primary species occurrence data is the amount of legacy data held in biodiversity institutes. These collections often contain one or more of the following data problems:

- **Missing information**: For some species, only the country of the collecting event is known
- **Misspellings:** because the labels are transcribed verbatim during digitizing project, the spelling mistakes that were on the labels are also adopted in the digital collection.
- **Historical names:** Some collections are up to 200 years old and contain names of sites that no longer exist in modern gazetteers.
- **Changes in the landscape:** because of changes of river banks, the growth of cities and the disappearance of settlements, it can be difficult to find some locality descriptions on modern maps.
- **Unclear descriptions:** some descriptions consist of only a direction, distance, or a landmark, making it difficult to find a precise location.
- **Location descriptions of data poor areas:** where few name-bearing landmarks exist, such as in tropical areas.

But even verily new collections that already contain GPS data know various data problems that could affect georeferencing results:

- The **plus and minus** of the coordinates are reversed, making the location description and coordinate data not compatible
- Differences in the number of **coordinate decimals** that get recorded, making the accuracy of the specimen collection differ from each other
- **Accuracy of the GPS** device can differ from the decimals recorded
- **"0.00 / 0.00"** is entered when the coordinates are unknown. Resulting in wrongly matched locations and coordinates.
- The **coordinate system or datum** sometimes not recorded or the wrong datum or system is recorded. Making it harder to use with other spatial data, of which the datum or system is different.

These types of data often lead to conflicts with the use of a georeference tool, or could lead to less accurate results. Some data problems can even lead to completely wrong georeference data.

Example:
A locality description contains the historic name of one of the islands of Indonesia, like Alkmaar, but there is no country recorded in the collection. Since [Alkmaar] is the only named place in the original data, most georeferencing tools will return coordinates for [Alkmaar, Netherlands] as a result, because this is the only current named place in most gazetteers that matches. It is however not the correct named place.

To avoid these problems it is important to inventory what types of data problems occur in the collection, if they can be corrected without too much effort (see chapter 3, step 2) or if they can be left out of the georeferencing process.

1.2 Database specific

The second issue that will need to be addressed is what requirements there are for a data model and a database management system to enclose georeference data.

This issue can be divided into two questions:

1. What are the fields do you need in your database to store georeferencing information correctly?
2. What measures have to be made to maintain the data quality?

Meta data Fields

Geographic information regarding species occurrence data can be divided into two sets of elements. The first set of geographic data elements regards the original locality description, as written on the label, in the field notebook or in the CMS. This set usually contains fields like [Country, State, Provinces, Island, locality, Station number, Full locality etc.]. These metadata elements are already present in most biodiversity collections, and are therefore not part of this best practice.

NOTE: It is however important to mention: that georeference data must be added to existing records, so that the existing data (the data as written on the label or in the field notebook) is not over written. This is important information and has historic value and should therefore never be overwritten by standardized data.

The second set of geographic data elements are those actually describing the georeference data and georeferencing process. This concerns the actual geocode, the latitude and longitude, but also some associated data. This associated information on methods used to determine the georeference, and on the extent and uncertainty of the geocode, are very important pieces of information for the end user. Additionally, these are very important pieces of information for managing and improving the quality of your collection data.

Standards

One of the main purposes of the georeferencing collections in a standardized way is the possibility to exchange and combine data from different collections and institutions. To make this possible, it is important that the details of the georeference data can be accessed in the same way throughout different collections and institutions. International standards such as Darin Core and ABCD can be used to facilitate this exchange.

- The Darwin Core standard is a relatively simple standard with a total of some 180 terms, divided into different categories. Darwin Core is particularly useful for the core data of specimen collection events but knows several extensions where additional data can be linked to the 'core data', like the 'Geospatial Element Definitions Extension to Darwin Core'.
  More information: http://rs.tdwg.org/dwc/
- Access to Biological Collections Data (ABCD) is a highly advanced data standard that makes it possible to record a lot of information standardized. The standard includes more than 1000 terms and can be scaled for specific information needs. It is not possible or even necessary to use all terms that the ABCD standard provides, since there are so many.
  More information: http://wiki.tdwg.org/twiki/bin/view/ABCD/WebHome

ABCD and Darwin Core can be combined through the use of mappings, to match the elements from both standards so that they complement each other. Projects like the BioGeomancer project use these mappings. When you analyze these mappings, it becomes clear that there are many fields that are occur in both

standards and are therefore probably highly valuable. For this best practice these are considered the **minimum required (primary) metadata** field for the recording georeference data in a biodiverse collection.

In addition to the primary fields, there are some fields, which occur in both standards, but differ in form a little such as a different definition or a different name. These fields can add value to the georeference data and contribute to a standard method for enclosing it, but the precise definition, is institution specific. these are considered as **Secondary metadata elements** and also contain some fields that describe the georeferencing process. These can be included in the collection management system, but will primarily contribute to the documentation of the development of the collection and to the reliability of the georeferencing process.

The following table shows which of the metadata fields from the standards are considered as primary and secondary fields. The fieldnames and description that appear in table are derived from the ABCD standard element Shema. For institutions that prefer to use the Darwin Core standard, the mapping found here: http://rs.tdwg.org/dwc/terms/history/dwctoabcd/, can be used.

Primary metadata elements	
Fieldname per standard	**Description/definition**
LatitudeDecimal	The latitude of the geographic center of the locality expressed in decimal degrees.
LongitudeDecimal	The longitude of the geographic center of the locality expressed in decimal degrees.
SpatialDatum	The geodetic datum to which the latitude and longitude refer.
CoordinatesText or Coordinates UTM/UTMText	A text representation of the coordinates. This can be one element for latitude and longitude, but can also be separate elements
CoordinatesGrid/GridCellSystem	The name of the system in which the verbatim geographic coordinates were recorded.
CoordinateMethod	A reference to the methods used for determining the coordinates and uncertainties.
coordinateErrorDistanceInMeters or AccuracyStatement	A measure of the area in which the described locality must lie. Usually expressed by the distance from the coordinates to the upper limit/outer corners of the polygon/radius, in meters. A free text statement of the degree of accuracy of the latitude and longitude coordinates.
Secondary metadata elements	
Fieldname	**Description/definition**
FootprintSpatialFit	The overlap between the actual locality and the georeference polygon/ point
GeoreferenceSources	A list of maps, gazetteers or other resources used to georeference the locality.
GeoreferenceVerificationStatus	The status of the georeference data; “What validation steps have been conducted?”
GeoreferenceRemarks	Comments about the recorded georeference data
Georeferenced By	The person or organization making the coordinate and uncertainty determination.
Georeferenced Date	The date on which the determination was made.

Maintaining data quality

Recording georeference data in standardized metadata fields is one part of the maintaining of this newly gathered collection data. You will also need to make sure that the actual georeference data is ‘clean and comparable’. The most effective way to ensure this is to apply constraints to certain metadata fields, so that data can only be recorded in the desired form. This way data cannot be recorded in the wrong fields, or in a deviating form.

Example:
In the fields [LongitudeDecimal] and [LatitudeDecimal], you could only allow values between +90 and −90. These pre-determined values ensure the latitude and longitude of all objects are comparable. Pick lists are also very common constraints to control the metadata that is recorded. This can be applied to fields like: 'SpatialDatum', 'CoordinateMethod', 'Georeferenced By' etc.

These types of constraints and guidelines need to be documented before any georeferencing process. This can help lead to consistency throughout the collection and minimize the occurrence of errors. Some examples of guidelines that can be documented:

- Units of measure.
- Methods and formats for determining and recording uncertainty and extent.
- Degree of accuracy in determining points where known.
- Required fields
- Format for recording coordinates (i.e., for lat/long, degrees/minutes/seconds, degrees/decimal minutes, or decimal degrees).
- Coordinate precision
- How to deal with the occurring data problems (see 2.1) in the existing collection

There are some general guidelines available that can be used to develop these institution specific documents or to derive certain methodologies from. The guidelines often contain best practices for certain types of localities and collection aspects. They do not cover usable tools, and are therefore applicable to many different situations. The following guidelines are widely accepted:

- Wieczorek, J. (2001)"*MaNIS/HerpNet/ORNIS Georeferencing Guidelines*. Retrieved November 2, 2015 from http://manisnet.org/GeorefGuide.html
- Chapman, A.D. & Wieczorek, J. (2006). *Guide to best practices for georeferencing*. Copenhagen: Global Biodiversity Information Facility
- Wieczorek, J., Q. Guo, and R. Hijmans. 2004. The point-radius method for georeferencing locality descriptions and calculating associated uncertainty. International Journal of Geographical Information Science. 18:745-767.

1.3 User specific

The quality of data is a concept that has many definitions, and often depends on the purpose and the origin of the data. In the geographical world one definition is widely accepted: fitness for use (or potential use) (Chrisman 1991). Which states that data quality is strongly related to use and thereby also to the users. The Fitness for Use cannot be determined without analyzing the users and their goals (Chapman, 2005). For biodiversity collections this means that the value of geographic data is centered on improving / enriching a collection so that it is useful for research, collection management and accessibility of the collection.

> *Example:*
> *If you are conducting a research on the spreading of a bird specimen in the northern part of Europe, the precision of a georeference code can be quite large, like 20 km. But if you want to conduct a research on the spreading of that specimen in the northern part of Norway, you will want a georeference code that has a smaller precision, like 5 km. If the available dataset has an accuracy of 10 km, it is fit for use for the first research, but not fit for use for the second research.*

Within biodiversity collections there are several application fields that benefit from the enrichment of collections with coordinates. But for geographic data from species the user group is much larger than the user group of taxonomical data, such as educational institutions, amateur associations, hobbyists, and in many cases even the entire public. Where the collection is often arranged systematically (taxonomically) for internal purposes, it is more accessible for the public if objects can be viewed on a map, or can be searched by area. This makes biodiverse data accessible to users without knowledge of taxonomy.

To make sure the georeference data is 'fit for use' it is important to determine wishes and requirements from key users of the future georeference data. By taking these wishes and requirements into account, geogereference codes better reflect the activities of the users and can lead to more efficient data use.

Important aspects to keep in mind when identifying user requirements:

- What **purposes** do the users use georeference data for?
- **Additional meta data:** what associated meta data is important for the goals/use of users.
- **Accuracy levels of the coordinates:** what accurate is most useful for them, and how should this be recorded.
- **Accessibility:** do all users need to have access to all the georeference data.
  NOTE: here it is important to keep in mind the sensitivity of occurrence data. See 'Guide to Best Practice for Generalizing Sensitive Species Occurrence Data', by Chapman &Grafton, 2008.
- **Prioritization:** what collections should be georeferenced first?
- **Validation:** do the georeference cods need to validated, by the users of one specific application field.

The most important aspect of the user specific wishes and requirements of georeferencing data and process is the accuracy of the generated coordinates. Accuracies and Inaccuracies highlight the limitations of the application of the data. These limitations determine the ultimate quality of the data, and thus the quality of the results, and how a researcher or collection manager should interpret them. So by recording this in the collection it can be made clear for what purposes the data is fit for use.

The most important thing is that the accuracy of the coordinates has to be indicated clearly and without the possibility of misinterpretation. This could be done, for example, with a field such as [uncertainty in meters], [the maximum error distance] or [maximum uncertainty]. Based on this researchers or other users can determine whether or not the coordinates are useful for their goal. Defining the degree of accuracy is best done in in measurable units, like meters. This is understandable and unambiguous for each user so that there no danger for misinterpretation or false precision.

1.4 Institution specific

This document provides guidelines to a best practice for semi- automates georeferencing, but it is of importance that institutions also take into account aspects of their individual situation. Many aspects of georeferencing are institution specific and cannot be standardized for al biodiverse institutions. Some aspects should be taken into account:

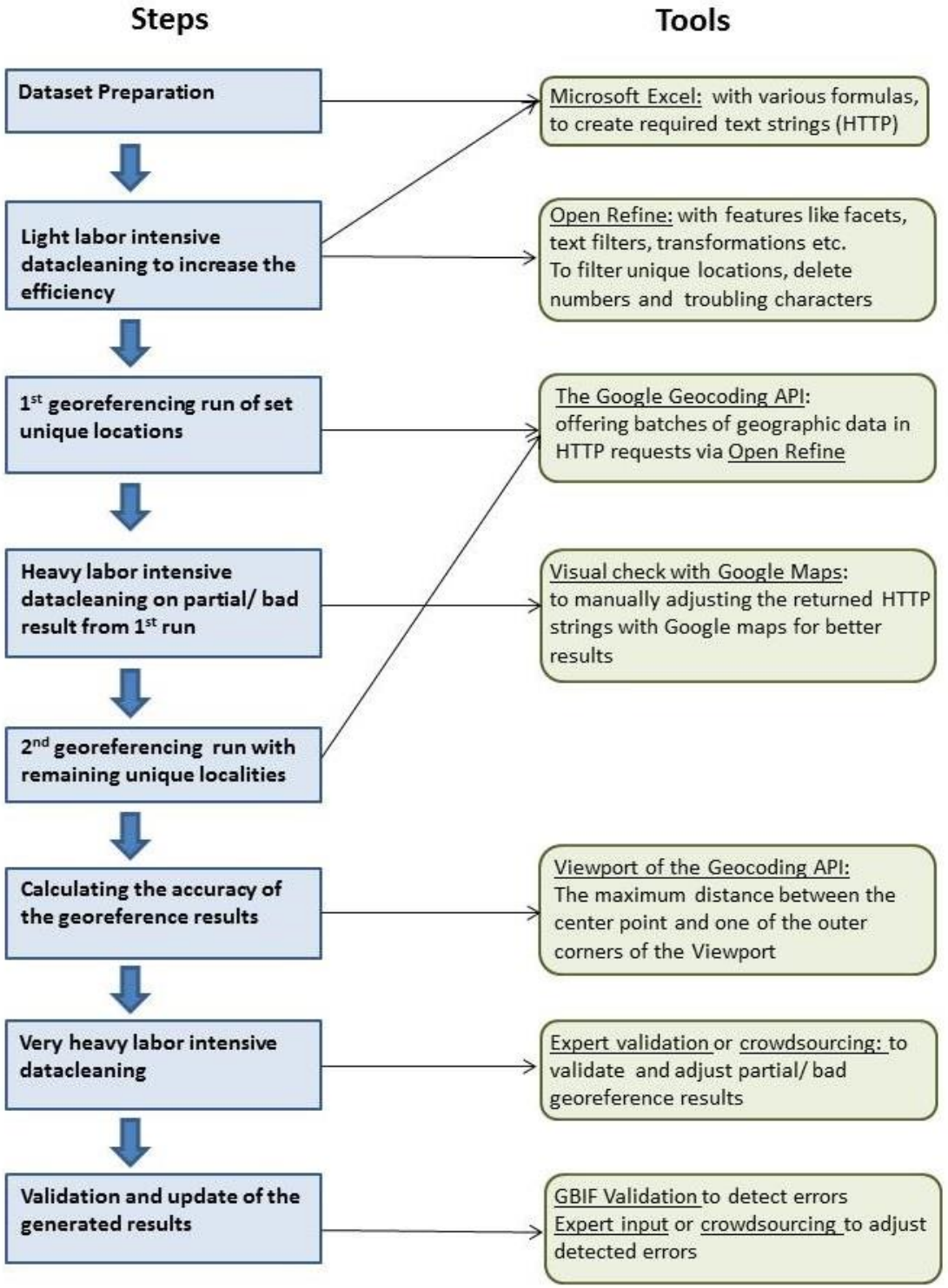
- Vision and mission of the institution
- Strategic plans, such as policies and procedures
- A standardized writing method (see guidelines)
- Possibilities regarding the collection management system
- Possibilities regarding employees and internal knowledge
- Quality of the original collection data

Based on all these aspects institutions can create their own internal document that incorporates best practices, general methods and agreements outlined with the working environment.

3. The georeferencing process

Many institutions and many experienced georeferencers have been working on tools, guidelines and standard methods for georeferencing primary biodiversity data. Each developing their own preferences for the order in which they georeference. This best practice is based on a research on a large amount of these projects tools, guidelines and standard methods. In an attempt to bring all these initiatives and experiences together an international standardized process was created.

In the figure below you can see the process visualized, including a short explanation for each step and the most usable tools, to conduct this step. In the following chapters, each step is explained with more details.



The figure above shows the process as a visualized roadmap for a quick guide through the georeferencing process. In the upcoming part of this chapter each step is discussed with more detail and links to the mentioned tools. In chapter 3 steps 8 and 9, regarding the validation and documentation of gathered georeference data is discussed.

For some tools, named in the process there some additional details, like actions, formula etc. that needs to be taken, in order to georeference batches of primary biodiverse data. This can be found in the appendix.

Step 1: Preparing dataset

To georeference a dataset with primary biodiversity data, this set must first be extracted from the collection management system. For this process Microsoft Excel can be best used to store this extracted data. The following data of the object can contribute to the georeferencing of the object location:

1. Registration number
2. All locality data (Country, State provinces, Island, locality, Station number, Full locality etc.)
3. Collection date
4. Collector
5. Altitude

The main tool that will be used in this process is the Google Geocoding API. Which georeferences via Google Maps by providing a single HTTP request / string for every record with all locality data to Google Maps. So, in order to use this tool, all locality data of one object also needs to be merged to one single string.

- **Microsoft Excel**

Microsoft Office Excel is a program in the Office Package of Microsoft, and can be sued to create spreadsheets. Excel can be used to create a new column the extracted dataset were, all separate location aspects (country, city, full locality, etc.) are merged to one string.

Step 2: Light labor intensive datacleaning to increase the efficiency

With light labor-intensive data cleaning a higher efficiency can be achieved in this method. Data cleaning tasks for this step can take place in the same Excel file that was created in the previous step or in Open Refine Light labor intensive datacleaning tasks:

- Adjusting characters that where copied incorrectly in the previous step, for example [IndonesiÃ «→ Indonesia].
- Removing numbers in the dataset. These numbers often refer to distances from a given point, but Google Maps can see these numbers as a house number or postcode, causing an incorrect coordinate to be returned.
- Removing duplicate names. Because of this, Google Maps can catch the wrong locations, or even not find locations when things are double in a string. This also reduces the amount of unique locations.
- Merging objects with the same location description to one unique locality. This reduces the amount of search strings for the Google Geocoding API.

- **Microsoft Excel:** https://products.office.com/EN/excel?omkt=en

With various formulas and feature, like [REPLACE] and [CONCENATE] it is possible to filter unique localities and remove or replace strange characters. Using Excel means, that u can perform this step in the same file that was created in the previous step. This tool requires users to create their own formulas for this datacleaning.

- **Open Refine:** http://openrefine.org/

Open Refine is an open source and online available tool, based on the functionalities and usability of a relational database. It looks very similar to excel and knows various formulas and features to edit and manipulate data and datasets**.**

With the use of features like facets, text filters, transformations etc. the amount of unique localities can be reduced and characters that might result in errors can be deleted or replaced. Open Refine works very similar to Excel, but has the big advantage that various standard feature can be used. U will not have to create your own formulas. Secondly for the next step open refine is also used for offering data to the Google Geocoding API. So eventually the Excel file from step on will have to be imported into Open refine.

For collections with Dutch localities.

- **Plaatsnamen- Standaardiseren:** http://standaardiseren.erfgeo.nl/

Plaatsnamen Standaardiseren[36] is an online tool that can be used to standardize Dutch place names in a dataset. The tool, which is only a demo version, is designed for collection managers to easily import data, standardize the geographic elements in the dataset, check the results and export the standardized data. The output of the tool contains 4 tabs;

1) The standardized localities that returned in one full match,
2) Localities that returned multiple matches. These results can be displayed on a map and can be edited by the user.
3) A tab that contains localities that did not deliver any results, this could indicate a misspelling or a place name that the tool does not recognize in any of the standard datasets, but these can also be adjusted manually by the user and
4) A Tab Intended for localities that cannot be standardized by the tool or with manual input of the user.

NOTE: The tool only works for Dutch geographic information at this moment. It uses the TGN and the Geonames databases to generate the standardized terms

- **Histograph:** http://histograph.io/viewer/

The Histograph[37] is a historical geocoder designed in Holland, for searching and standardizing historical place names in a dataset. The geocoder technique collects and links place names of the same locality over time and standardizes and georeferences these names. Histograph can be used to replace the historic names in textual locality description with modern names that the georeferencing tools would recognize.

A user can type in a place name like Amsterdam and Histograph finds al different names that were used in history for that place name, including the place boarders, numerical time definitions, and the original source of the data (Erfgoed & Locatie, 2015).

At this moment the Histograph tool uses data sources with; birth places of Dutch East India Company crew members, monastery records and historical census data for the historical place names, and the Geonames and TGN for the standardized modern place names (Erfgoed & Locatie, 2015). The Histograph tool is at this moment still under development, but a demo version can be used trough the Histograph API.

NOTE: The tool no focusses only on Dutch data, but the techniques are open source and could be used for the same purposes for different areas or countries, by adding different historical geographic data.

[36] http://standaardiseren.erfgeo.nl/
[37] http://histograph.io/viewer/

Step 3: georeferencing: 1e run of set unique locations

- **The Google geocoding API:** https://developers.google.com/maps/documentation/geocoding/intro

The Google Geocoding API[38] is an online tool that can be used to find latitude and longitude coordinates of a location. The Google Geocoding API is the only available tools that can georeference objects in batches. Google accepts over 50 languages and has geographic information covering the entire world. It has a build in ranking mechanism that finds matching well known and populated places based on the original request, so it could work for data with minor miss spellings.

With the help of Open Refine, the localities can be offered to the Google Geocoding API. The set unique localities that where created in Excel or Open Refine during the previous step, first need to be imported in Open refine and transformed to HTTP requests/strings. This creates a HTTPs request/string for the Google Geocoding API, for all objects in the dataset, with the following structure:
http://maps.google.com/maps/api/geocode/json?sensor=false&address= *[location data]*.

> *Example:*
> *If the locality description is Darwinweg 2, Leiden, the http request/string is:*
> *http://maps.google.com/maps/api/geocode/json?sensor=false&address=Darwinweg 2,Leiden*

The Output the Google Geocoding API returns the following data:

4. **Address components:** this shows which components are used to find the address; like land, political borders, postal code, house number etc. Not all parts of the original locality are used in all strings.
5. **Formatted address:** this contains the textual address that was found and the latitude and longitude of the found address
6. **Viewport:** shows the accuracy of the found results, in coordinates. It indicates an area, shaped like a rectangle, in which the results are located. This is based on the outer southwest and northeast corners of the smallest geographic unit of the address component the Geocoding API found like

Step 4: Heavy labor intensive datacleaning

After the first run of georeferencing with the Google Geocoding API, a part of the results will not meet the requirements to be recorded back in the collection management system. With heavy labor intensive datacleaning this set unique localities (with partial, or bad results) can be cleaned up for a second run.

The **Visual Check of the Google API** is a datacleaning method that could be used for this process. By using the returned strings from the Google geocoding API run (from the previous step), a user can check the results visually on the Google Maps application, and adjust the string manually for a more accurate result. It still requires some degree of manual input, but increases the amount of object records with good georeference data.

Step 5: 2e run with remaining localities

After the heavy labor intensive data cleaning this set of new search strings is used to do a second run of georeferencing with the Google geocoding API, according to the procedure explained in step 3.

[38] https://developers.google.com/maps/documentation/geocoding/intro

Step 6: Calculating accuracy

Calculating uncertainties in georeferenced data is a key aspect in determining the data's fitness for use and thus their quality. The Google Geocoding API output consists of 3 elements, as can be seen in step 3. Of these 3 elements [the Viewport] houses the data that can be used to calculate the accuracy of the retuned results.

- **Viewport of the Geocoding API:** https://developers.google.com/maps/documentation/geocoding/intro
  The Viewport (or Bounding Box) indicates the accuracy of the found results. It indicates an area (polygon), in the shape of a rectangle, in which the results are located. Generally this area covers the smallest geographic unit of the address component the Geocoding API found. The viewport is based on coordinates of the outer southwest and northeast corners of the Viewport.

  A good measure of the accuracy of the original location, is the largest distance between one of these two corners to the central point of the viewport area. This distance can be seen as a radius that defines the area within which the location must be located.

Step 7: Very heavy labor intensive datacleaning and data check

The last step of the actual georeferencing of object records, handles the remaining objects that even after datacleaning and 2 runs with the Google Geocoding API, still do not return reliable and usable results. This mainly concerns object record that contain localities that are too difficult, or vague to georeference automatically with tools. they either contain too much noisy data, 10 miles of shore, next to bridge, etc. or have too much missing geographic aspect, for example only a street name, but no house number, no city, no country. These objects require manual input in order to collect georeference data. For this there are two options:

- **Expert validation**

In some cases the georeferencing data and benefit from knowledge of the original object and species, and expert experience with the collection. Experts within an institution can validate (or adjust) georeference data for objects that still have questionable georeference data, or do not have any georeference data. A way to make this easier is to this in an environment where al data, original geographic data, georeference data, and the visual location on a map, can be seen at once, for example, with the possibility to plot object directly on maps in a CMS.

These way experts within a biodiverse institution can work collaboratively on checking and validating georeference data in a sort of knowledge based process. When doing this, it could be useful to add an extra metadata element in which the status of the validation process is recorded.

- **Crowdsourcing**

A process as described above, can of course, also be executed with the help of the public, in a crowdsourcing environment. This could save the data owning institution allot of time. These options do however; assume that an institution has access or the possibility of creating a crowdsourcing platform and a population of citizen scientist. This does require time and money to create, but generally a crowdsourcing project delivers more than just the data the institution gathers. It creates a higher involvement of the crowd in biodiversity; it can bring new visitors and even generate publicity with the project.

Crowdsourcing could also be used in various other stages of the georeferencing process. When going back to the table in the conclusion, crowdsourcing could be used to aid in the following steps:

- **Filtering unique localities**: this can be done with formulas in Excel or with various features in Open refine, but these methods only match unique locations with the exact spelling. But in most collections multiple ways of spelling or combinations of named places refer to the same locality. For example: 'Box Hill', 'Box Hill; Surrey', 'Box Hill, Kent', 'Box Hill; near Dorking', 'Box Hill; Dorking' all refer 'Box Hill; Surrey; UK; 51.254 N, - 0.308W'. Matching these together can decrease the amount of unique

localities even more. The crowd could link locality description that refers to the same place or by linking place names to thesaurus terms.

- **Labor intensive datacleaning tasks with Google Geocoding API**: this requires allot of time but it can improve the percentage of georeferenced records and can create better quality georeference data. A crowdsourcing project could be deployed to manually adjust the individual objects strings for the Visual check by Google.
- **Very heavy labor intensive datacleaning**: with the help of the crowd objects that still don't have a georeference after all process steps and require complete manual input to get results can be done individually.

4. Post- Georeferencing process

After the actual georeferencing is completed, during step 1 to 7, the gathered results will need to be validated and documented in the collection management system. It is an important part of the georeferencing process, as these are the steps that make the gathered data usable for the future. It is imported that information on the progress of these steps is also tracked. For example with the fields [GeoreferenceVerificationStatus], [GeoreferenceRemarks], [Georeferenced By] and [Georeferenced Date].

Step 8: Update or validation

The Google Geocoding API returns georeference data after a HTTP request with a confidence score based on if a unique locality was found and if this is a full match with the data from the original HTTP request. However, the assumption that all results with a good confidence score and a full match are correct can be questionable. Especially with original locality descriptions that contain historic names, or don't contain a country name.

Example:
A locality description contains the historic name of one of the islands of Indonesia, like Alkmaar, but there is no country recorded in the collection. Since this Alkmaar is the only named place in the original data, the Google Geocoding API will return coordinates for 'Alkmaar, Netherlands' as a result. This is a full match according to Google, since this Alkmaar is the only current named place in most gazetteers, even though it is not correct. This is due to the built in ranking mechanism that the Google Geocoding API has, that focusses on well known and populated places. If you use Alkmaar, Indonesia in the HTTP request, Google will find the result you want.

This 'problem' decreases the truth level of full matches. Which means it has to be considered if it is necessary to also have the full matches checked for these misinterpretations and validated by an expert? This can for example be done only for objects that miss a country name in the original locality description. Another possibility is to have objects checked and validated of which it is known that they were collected before a certain year or date (or buy a specific collector), in an area that has changed names since then, like former colonies. The most simple solution is to make a distinction in your collection between object that where georeferenced manually an automatically. This way the objects that possibly have a questionable truth level are marked.

Checking a dataset with objects for missing country names, or misinterpreted georeference results can be very difficult and time consuming, because it is hard to spot these types of anomalies in numerical coordinates in large batches. There are various option for detecting anomalies and possible errors with the help of additional techniques, web services and data resources:

- Using additional literature and data resources such as collector's itineries, gazetteers, field note books, etc.
- Checking against other fields in the collection (to make sure the georeference matches the collection date/time)
- Checking if the georeference results fall within the correct state, country, region, etc. This can be done with the GBIF validation check. When publishing data on the data portal of the GBIF web services, various checks are preformed and all correct data is validated. This can be used to detect possible errors.
- Using statistical methods or text mining techniques such as cluster analysis to identify outliers in latitude or longitude.

Step 9: Documentation of the data

The documentation of the gathered georeference data after is very institution specific, because it largely depend of the type and structure of the collection management system.

If the document containing al the georeference results, is created and stored in Open Refine it can be exported in any of the following file formats.

- TSV
- CSV
- Excel
- XML
- RDF as XML
- JSON
- Google Spreadsheets
- RDF N3 triples

If the document containing al the georeference results, is created and stored in Excel, it might be necessary to convert this to a more suited file format, depending on the requirements of the collection management system. Converting file to different formats can be done in Open refine.

Regarding the recording of the generated georeference data, it is important to note that the output of the Google Geocoding API does not fully match the ABCD or Darwin Core standards. But in the tabel below a mapping is created to match the output of the Google Geocoding API as best as possible to the advised primary standard elements from chapter 2.2.

Primary Standard elements	**Google Geocoding API output elements**
LatitudeDecimal	Returned Latitude
LongitudeDecimal	Returned Longitude
SpatialDatum	WGS84
Verbatim coordinate system	World Geodetic System 1984
Coordinate precision	Decimal degrees
CoordinateMethod	The Google Geocoding API + the address components that were used to find the coordinates
coordinateErrorDistanceInMeters or AccuracyStatement	The Viewport indicates a polygon in which the results are located, based on the outer southwest and northeast corners of the smallest geographic unit in the results. The distance between these points can be the uncertainty in meters, or the coordinates of this viewport can be used to record the polygon.

5. Discussion

In the process described above, only the Google Geocoding API is mentioned as a usable tool. Derived from the research this best practice is based one, there are two additional tools that could contribute to the described process. The SpeciMap and the Georeferencing Calculator show very promising, but is unfortunately are not yet or no longer available for public use.

The Institution that created the SpeciMap tool did mention public release in the near future. When combined with the Google Geocoding API, this would form the perfect match and be used following on each other. Combination of these tools (combined in 1 tool or following each other) automates the georeferencing process more and leads to more accurate results. The Google API could be used for bulk georeferencing and the SpeciMap tool for the more accurate adjustments of the retuned strings.

The Georeferencing Calculator could also add allot of value to the georeferencing process described above. It is the only tool that can handle additional aspects of geographic location descriptions like headings and offsets. The tool can also be adjusted to fit all the wishes a user has; the calculation type, the coordinate precision, the locality type, coordinate source and system, datum, etc. Additionally the tool generates metadata that fully matches the Darwin Core standard. If this tool were to be updated to a more current version of java (8), this would be a very promising option.

There are several steps in the georeferencing process, described in this document that can/need to be extended if these tools would come available or be released for the public. Both tools only work with individual object records, not batches. In this chapter you can find a description of the steps where these tools could be deployed.

Step 3: georeferencing: 1e run of set unique locations

SpeciMap or the Georeferencing Calculator: for georeferencing objects that contain textual description with headings, offsets etc. This also goes for step 5, the second run with the remaining localities.

- **SpeciMap**: is semi- automated georeferencing tool that takes certain fields from the collection database and identifies text strings as places with the help of the Geotagger and part of the Geoparser tool. These strings are plotted on a combination of different maps, including Google maps and various historical maps, survey maps and the National Grid Reference. The user can move through the different layers of maps and correct or make additions to this generated output in order to specify the exact location where the object was collected (Llewellyn et al, 2011).

- **Georeferencing Calculator**: Is a tool that can be used for different types of calculations:
  - **Coordinates and error**: If you want to figure out coordinates and errors based on a named places, headings and offsets
  - **Error only**: If you already have coordinates and only want to calculate the error.
  - **Coordinates only**: If you need to determine the coordinates of a named place based on known reference coordinates.

  Additional to the different types of calculations a user can also select different types of locations, based on the elements that can be determined in the original textual locality description. These location types can also handle coordinates, headings, offsets and orthogonal directions. Thirdly a user can also select: the coordinate source, coordinate system/datum, coordinate precision, direction, offset distance and the extent of named place. This means the tool automatically takes into account different aspects of a textual locality that can affect the uncertainty of the georeferenced point.

Step 6: Calculating accuracy

SpeciMap: can be used for improving the accuracy of objects that require a very small accuracy.

By manually replacing the found georeference point to a more accurate point, the 'maximum errors distance' of an object can be decreased. This is especially useful for objects with a historic component or with special research demands.
The tool allows users to add georeference codes based on textual localities or complement already existing coordinates. This means it can be used for two types of calculations:

1. Complement existing coordinates: to manually adjust objects that received a good georeference code in step 3 and 5, but requires a more accurate 'error'.
2. Georeferencing textual localities: to georeference objects that did not return good results in step 3 and 5 and require manual input to georeference.

Georeferencing Calculator: is very useful for calculating the uncertainty in meters for objects which originally had distances and heading in the locality description.

During the light labor intensive datacleaning step, difficult charters, numbers and heading were removed from the dataset, because most tools do not recognize this type of data. The Georeferencing Calculator is the only tool that does recognize this data.

1. The calculation option 'Error only', lets users calculate the maximum error distances for already known coordinates. Here the original headings, offsets etc. can be added to make the georeference code more accurate.

The calculation option 'Coordinates and error': If you want to figure out coordinates and errors based on a named places, headings and offsets for objects that did not return good results in step 3 and 5.

References

Chapman, A.D. & Wieczorek, J. (2006). *Guide to best practices for georeferencing*. Copenhagen: Global Biodiversity Information Facility

Chrisman, N.R. (1983). The role of quality Information in the long-term functioning of a GIS. *Proceedings of AUTOCART06*, 2: 303-321. Falls Church, VA:ASPRS

Erfgoed & Locatie (2015).Demo-versie Historische Geocoder online. Retrieved on November 18th 2015, from http://erfgoedenlocatie.nl/2015/04/demo-versie-historische-geocoder-online/

Github (2015). Geocoding, Translate street addresses to lat/lng coordinates. Retrieved on November 17the 2015, from https://github.com/OpenRefine/OpenRefine/wiki/Geocoding

Google (2015). Geocoding service - Google Developers Retrieved November 2, 2015, from https://developers.google.com/maps/documentation/javascript/examples/geocoding-simple

Language Technology Group, School of Informatics, University of Edinburgh (n.d.). *The Edinburgh Geoparser*. https://www.ltg.ed.ac.uk/software/geoparser/

Llewellyn, C.A. & Hasten, e. & Grover, C. (2011). Georeferencing Botanical Data using Text Analysis Tools. TDWG 2011 Conference, retrieved on October 15th 2015, from http://www.researchgate.net/publication/282701713_Georeferencing_botanical_data_using_text_analysis_tools

TDWG (2007). *Mapping between Darwin Core 1.4 concepts (DwC) and ABCD 2.06b*. Consulted op 14th of September 2015, http://www.bgbm.org/TDWG/CODATA/Schema/Mappings/DwCAndExtensions.htm

TDWG (2010 A). ABCD - Access to Biological Collection Data. Consulted on September 14th 2015, http://wiki.tdwg.org/twiki/bin/view/ABCD/WebHome

Wieczorek, J. (2001)"*MaNIS/HerpNet/ORNIS Georeferencing Guidelines*. Retrieved November 2, 2015 from http://manisnet.org/GeorefGuide.html

Wieczorek, J., Q. Guo, and R. Hijmans. 2004. The point-radius method for georeferencing locality descriptions and calculating associated uncertainty. International Journal of Geographical Information Science. 18:745-767

Wieczorek, J., D. Bloom. 2011. Georeferencing Calculator Manual v2. Retrieved on November 18th 2015, from http://manisnet.org/GeoreferencingCalculatorManualv2.html

List of tools

The Google Geocoding API
Tool: https://developers.google.com/maps/documentation/geocoding/intro
More information/manual: https://developers.google.com/maps/documentation/geocoding/intro

Microsoft Excel
Tool: https://products.office.com/EN/excel?omkt=en (download required)
More information/manual: http://www.microsofttraining.net/download/manuals/Excel-2010-Advanced-Best-STL-Training-Manual.pdf (PDF)

Open Refine
Tool: http://openrefine.org/
More information/manual: http://openrefine.org/documentation.html

Visual Check Google Maps
Tool: Excel
More information/manual: *not available*

HistoGraph
Tool: http://histograph.io/viewer/
More information/manual: http://histograph.io/ and https://github.com/histograph/installation

Plaatsnamen- Standaardiseren
Tool: http://standaardiseren.erfgeo.nl/
More information/manual: http://erfgeo.nl/wat-hoe/standaardiseertool.html

SpeciMap
Tool and: More information/manual *Not yet available, but based on the Edinburgh Geoparser:* https://www.ltg.ed.ac.uk/software/geoparser/

Georeferencing Calculator
Tool: http://manisnet.org/gci2.html
More information/manual: http://manisnet.org/GeoreferencingCalculatorManualv2.html

Appendix: formulas and functionalities: a short manual for various tools

Step 1: Preparing dataset

The merging of all locality data is with the help of Excel goes as follows:

1. Create an excel file with in each row a Registration number and in each column one of the locality data elements from the registration system. Example:

Registr. number	Country	Provinces	City	Locality
12082619	Netherlands	Zuid-Holland	Leiden	Darwinweg 2

2. Then create a new column in the excel file a place the following formula in the first cell of this column:

```
=TEXT.S CONCATENATE (
IF(ISTEXT[1st cell with locationdata];[1st cell with locationdata] & ", ");
IF(ISTEXT[2nd  cell with locationdata];[2nd cell with locationdata] & ", ");
etc.
)
```

3. Draw the cell through the entire column

Step 2: Light labor intensive datacleaning to increase the efficiency

Filtering unique locations with Excel:

1. Create a new column next to the new column from the previous step.
2. Place the following formula in this column:

```
= VLOOKUP (
[cell from previous column];
[cellreach (columns, location names and location numbers) in file with all the unique location data];
[column number with location number];
FALSE)
```

   and draw the cell through the entire column.
3. Select the just created column with the wanted location numbers.
4. Sort this from low to high.
5. Select the location data from the previous column and choose: *Data > Sort and filter > Advanced*
6. Select the column with the pastes values in a new tab for the option 'list range'
7. Check the 'only unique records' box.
8. Press OK
9. Now select the location cells in this column that don't have a location number. This can be recognized by *'#N/B'*
10. Copy these cells
11. Paste these cells below the location cells in the file with the previous found unique localities.
12. Give these just pasted cells a unique location number.

- **Plaatsnamen- Standaardiseren (For Dutch collections)**

Standardizing geographic data in a dataset with Plaatsnamen- Standaardiseren:

5. Create a csv. file with the place names that need to be standardized. The csv. File should contain a column with the terms that you want standardized. This could be one or more columns.
6. Upload the csv. file and select the field that contains the place names and what type of locality it contains like, cities, states etc.
7. Let the tool do the standardizing
8. Review the results and edit the results if necessary (delete wrong matches; choose the right match if an object returned multiple results etc.

NOTE: The tool only works for Dutch geographic information at this moment. It uses the TGN and the Geonames databases to generate the standardized terms

Step 3: georeferencing: 1e run of set unique locations

The set unique localities that where created in Excel or Open Refine during the previous step, need to be imported in Open refine and transformed to HTTP requests/strings. This is done as follows:

1. Store the unique localities in a file, like Excel
2. Open Open refine
3. Open the file with the unique localities
4. Click on the dropdown menu of the column with the localities
5. Edit Column > Add column by fetching URLs
6. Type in the following:

```
"http://maps.google.com/maps/api/geocode/json?sensor=false&address=" + escape(value, "url")
```

7. Give the column a name, for example ‘geocodingResponse’

This creates a HTTPs request/string for the Google Geocoding API, for all objects in the dataset, with the following structure: http://maps.google.com/maps/api/geocode/json?sensor=false&address= *[location data]*.

To extract the latitude and longitude with the help Open Refine, from the Google Geocoding API output, the following actions need to be taken:

1. Click on the dropdown menu of the column
2. Edit Column > Add Column based on This Column…
3. Fill in the following value to place the lat and long in a new column:

```
with(value.parseJson().results[0].geometry.location, pair, pair.lat +", " + pair.lng)
```

4. Give the new column a name. For example ’GeocodingResponse lat/long’

Step 4: Heavy labor intensive datacleaning

This datacleaning can be done by creating new text strings in Excel and plotting them on Google Maps. By viewing the results visually, you can manually adjust the text string for a better fitted or more precise result. This can be done in the following way:

2. Place all search strings that did not return a good result in a new column
3. Create an additional column next to this column where all the search strings were placed that did not return a good result, with the following formula:

=HYPERLINK(CONCATENATE ("https://www.google.nl/maps/search/";[Cel *met searchstring*]))

4. Next you can click on these hyperlinks and the locality is shown on Google maps accompanied with the search string
5. In this search string you can now make small adjustments, that directly visible on the map. This way you can keep adjusting until the search string delivers one unique locality and which continues to meet the original search string
6. The adjusted string can be placed in a separate column in excel. This way the link between the original string and the new string is maintained.