Georeferencing Animal Specimen Datasets

Marieke van Erp*, Robert Hensel, and Davide Ceolin

The Network Institute, VU University Amsterdam, The Netherlands

Marian van der Meij

Naturalis Biodiversity Center, Leiden, The Netherlands

Keywords: *georeferencing, natural history, data integration, reasoning, knowledge engineering, confidence measure*

*Corresponding author. Tel: +31 20 598 5316 *E-mail address*: marieke.van.erp@vu.nl

Abstract

For biodiversity research, the field of study that is concerned with the biological diversity of our planet, it is of utmost importance that the location of an animal specimen find is known with high precision. Due to specimens often having been collected over the course of many years, their accompanying geographical data is often ambiguous or may be very imprecise. In this contribution, we detail an approach that utilizes reasoning and external sources to improve the geographical information of animal finds. We show that adding external domain knowledge improves the ability to georeference locations over traditional methods that focus solely on analyzing geographical information. Additionally, our system is able to output the confidence it has in its decisions through a confidence measure based on the difficulty of the instance and the steps undertaken to disambiguate it.

1 Introduction

Natural history museums such as Netherlands Centre for Biodiversity Naturalis (Naturalis)[1], harbour vast collections of biodiversity specimens collected from all around the world and their accompanying data. These collections encompass an enormous amount of information about the biodiversity of our planet, but institutions are still in the process of unlocking their potential through digitization of their collection metadata. An important aspect of biodiversity data is the location of the specimen find; without it, researchers cannot track species' geographical distribution over time, model the effects of environmental changes on species or try to predict how changes in the environment influence biodiversity in certain regions. The lack of precise geographical information presents a major problem in the efficient use of the collection metadata. Recently, Naturalis and the Computer Science department at VU University Amsterdam have teamed up to develop a solution to aid biologists in georeferencing their collections.

The majority of objects in the collection were collected a long time ago, often in countries that were previously colonies of the Netherlands. These specimens (or parts of them) are preserved in various ways. Invertebrates can be kept dried or preserved in alcohol, vertebrates are often kept dried and mounted or in drawers, and the entomology collection is often pinned or kept in bags. All specimens are kept in a secure climatized storage facility. This specimens are used by researchers from all over the world to investigate for example geographical spread and evolution of species and to classify and describe different species). As the oldest parts of the Naturalis collection date back to the 18th century, most geographical information in the collection records is captured in a textual description indicating location(s) and offset(s) such as "Anti-Atlas, 10-20 km S. Ait-Baha, Morocco" rather than precise geographical coordinates. The inves-

[1] `http://www.naturalis.nl`

tigated databases at Naturalis contain fields that describe the geographical information: country, province, town, location, coordinates and altitude. See [21] for an overview of the database fields and statistics on how many of the fields are filled. With semi-automatic approaches such as the MaPSteDI method [15], georeferencing a record reportedly takes approximately 5 minutes per record. As Naturalis harbours 37 million objects with each their own record, manually georeferencing each record would be time-consuming and costly. Crowdsourcing this task is not an option as most of the data is domain specific, delicate or cannot be made public. However, there is a large amount of biodiversity knowledge being made available online which, when integrated with geographical resources can be put to use in a knowledge-driven georeferencing approach.

To address this challenge, we have developed an automatic georeferencing approach that uses domain knowledge about species geographical distribution from the online Global Biodiversity Information Facility[2]. This approach has been realized in a prototype currently being tested at Naturalis. The contributions of our work are threefold:

1. The first knowledge driven approach for automated georeferencing in the biology domain.

2. A novel automatic confidence measure for georeferenced records to help curators pinpoint difficult records whilst handling, easier more certain records automatically.

3. A systematic evaluation that shows how a knowledge driven approach improves geographical information in a major biodiversity facility while limiting the need for input by domain experts.

The remainder of this paper is organized as follows. In Section 2, we describe the challenges in georeferencing biodiversity data, followed by previous work in

[2] `http://www.gbif.org`

Section 3. In Section 4, we describe the datasets we used. Our georeferencing approach is described in Section 5, followed by the results in Section 7. Our confidence measure is described in Section 8. The prototype demonstrator is described in Section 9. Conclusions and future work are discussed in Section 10.

2 Challenges

The problem of georeferencing natural history collections is not new: the different types of challenges have been categorised and described by Beamann and Conn [1]. In Table 1, we illustrate each of the challenges by an example from the Naturalis collection. It is not possible to georeference all types of localities with equal precision. Vague localities, such as "Southeast Michigan", simply contain too little information to pinpoint a spot within a small range (<5km) of the actual finding location, but for localities containing for example linear feature measurements such as "16 km N of Murtoa" this is feasible.

[Table 1 about here.]

3 Related Work

There is a fair body of research on georeferencing both outside and inside the domain of natural history. Within the natural language processing community georeferencing is treated as follow-up task to named entity recognition [11, 14], or possibly as complementary to it [4]. There are also several open source tools available such as OpenSextant[3] and CLAVIN[4]. However, these approaches assume full text, whereas the datasets in the natural history domain are part of structured database records, making them suboptimal for this domain.

[3] `http://opensextant.github.io`
[4] `http://clavin.bericotechnologies.com/`

Most approaches for structured data use some sort of gazetteers combined with some form of reasoning to disambiguate and ground location names [10, 12, 13]. These assume that the location names have been identified, skipping the step of recognizing the location name and possible extra locality information. Other approaches disambiguate toponyms from combining the analysis of events with gazetteer [17]. In our case study, we do not have event descriptions at our disposal. However, the procedure of information extraction, gazetteer lookup and geodisambiguation is very similar to the one described in [17].

Another emerging relevant research thread regards the use of crowdsourcing to acquire useful information for georeferencing [3, 9]. This kind of approach assumes the availability of a crowdsourcing platform and of a population of information contributors which we do not have in our case. However, these approaches are complementary to the procedure described in our contribution, since the crowdsourced information might help in improving the disambiguation heuristics. This is out of the scope of this contribution, but a very interesting avenue of research for future work.

Within the natural history domain, several attempts have been undertaken to automatically assign coordinates to textual descriptions of locations in specimen datasets with BioGeoMancer as its most well-known application[5] [1, 7].

BioGeoMancer provides an application for text processing, interpreting, gazetteer querying (using a variety of sources[6]), intersecting spatial descriptions and as a result returning a standardized geographical reference including uncertainty levels. The initial version of BioGeoMancer supported interpretation of localities in English, Spanish and Portuguese. However, the latest available version of BioGeoMancer supports English queries only.

Also developed for georeferencing natural history data is GeoLocate[7] [16].

[5] `http://bg.berkeley.edu/` & `http://sourceforge.net/projects/BioGeoMancer/`
[6] `http://www.biogeomancer.org/metadata.html`
[7] `http://www.museum.tulane.edu/geolocate/`

It uses similar gazetteer data as the BioGeoMancer project[8]. GeoLocate uses different georeferencing heuristics as well as additional linear features to its gazetteer such as rivers, road, legal land descriptions and river miles. These additional information sources can lead to more accurate results, but are only available for the United States, Canada, and Mexico. A comparison of automated georeferencing tools found that, at the time, GeoLocate was the best software tool to efficiently georeference large datasets [15].

Although both focused on the biology domain, neither BioGeoMancer nor GeoLocate makes use of domain-specific knowledge, such as species occurrence data. Also neither can deal with non-English data.

4 Data

In this section, we describe our primary dataset as well as the resources used for georeferencing and the development of the gold standard.

4.1 Reptiles and Amphibians Database

Several large datasets of animal specimen datasets are maintained at Naturalis. The information in these datasets comes from the field logs and registers in which biologists that made these finds recorded them manually, usually during expeditions. Part of the information from these sources has been converted to electronic datasets over the course of time by many different biologists working with the specimens. Creating these databases was not a top-down organized undertaking, but rather taken up by the researchers themselves to improve access to the data for themselves. As such the structure of these databases differs, and currently initiatives are underway to standardize the electronic data

[8]Since 2006, GeoLocate is part of the BioGeoMancer Workbench, but the current status of the integration of the projects is unclear

recording process.

For our case study, we used the reptiles and amphibians database containing 29,752 records, each referring to a specific animal's find. Among the information in these records, one typically finds locality information indicating where these specimens were found, its species, the name of the collector, information about when it was entered into the database and by whom it was entered into the database. The location find information is divided over several different database fields, namely "Town/City", "Province/State", "Country", "Location"[9], "Altitude", and "Coordinates" (only filled in 3.4% of the records). In this contribution, we shall mostly focus on the information from the Town/City, Province/State, Country and Location fields. The records also contain some additional notes describing the circumstances under which the specimen was found or any other unusual information about the specimen. The data is mostly in Dutch and English, but also Spanish, Portuguese and German are present. Foreign languages are found in particular in local abbreviations or terms. Some examples of this are the use of "Municipio" (Spanish for municipality) instead of municipality or "Est." as abbreviation for "Estado" (Spanish for state). For the work presented in this contribution we focus on Dutch data records as this is the main language of the database.

4.2 Gold Standard

To test our system, we created a gold standard consisting of 200 records. 50 records were used to develop and tune our system on, for example to check if the offsets calculation performed as expected. 150 records were kept separate for final testing. Records were selected randomly but with two aspects in mind: common challenges and internal representativeness. The first aspect ensures

[9] This sometimes contains the town or city value, but more often it is used to describe offsets or particularities of the find, such as that the specimen was found under a branch or in a puddle.

that the different types of locality information present in the database are represented in the gold standard. The second aspect balances for the fact that some types of locality descriptions are more frequent than others.

Due to the limited resources for annotating the gold standard dataset, we decided to focus on those categories from the initial nine categories in Table 1 (presented in Section 2) that contained vagueness or linear feature measurements. We collapsed the categories that are closely related by means of challenges posed to our system, as indicated by the letters in Table 1. As there are no high quality resources available currently that delineate political borders over time, or provide accurate information about historical place names, we discarded categories E1 and E2 for this contribution. In Table 2, we show our categories, as well as an example, the distribution of records pertaining to this category in the gold standard and in the entire database[10].

[Table 2 about here.]

Manually adding geographical coordinates to the 200 records for our gold standard was done by two annotators using pre-agreed on guidelines which were based on the MaNIS/HerpNet/ORNIS Georeferencing Guidelines[11]. These guidelines outline a cascaded approach in which annotators will first try to look up place names in a gazetteer, after which they address offsets or other complex features of the locality to be georeferenced. Furthermore, annotators were asked to indicate records that were difficult to georeference with an extra tag "PRECISION=LOW". Both annotators georeferenced 120 records, of which 40 were compared to determine inter-annotator agreement. It took both annotators about 2 afternoons to complete this task. Despite the guidelines, there was a fairly large disagreement indicating the difficulty of the georeferencing task. Of the 40 records that were annotated twice, 60% were an exact match or did

[10] Estimate based on an automated categorisation script
[11] `http://manisnet.org/GeorefGuide.html`

not differ more than 1 km, 65% were correct within 10km, and 85% were correct within 100km. Records differed due to differences in interpretation of the offset, the use of different information sources (often resulting in minor differences) or selecting different places from the gazetteers.

Common sources of disagreement were the distinction between administrative areas (such as provinces and states) and actual populated places, and choosing different paths when calculating offsets (such as "18 mls. E. of Kumasi"). If there is more than one linear feature (linear features include roads, streams, railways etc.) going east out of a town, it is up to the annotator to make an educated guess which of these has to be followed.

4.3 Gazetteers and biodiversity resources

Two geographical gazetteers were used to look up place names: **GeoNames**[12] and **Google Maps**[13]. GeoNames contains about 10 million place names and information about those places, such as coordinates, alternative names, elevation levels and population numbers. Its scope is global, although it contains more information about highly populated areas. In the context of this research, this may be an obstacle as much of the specimen finds are outside populated areas.

We prefer to use GeoNames because of its rich structured information about its location, but in cases where we cannot find a location in the GeoNames database, we fall back on Google Maps. Because of its built-in ranking mechanism, it will return more important places first which is information that can be used to confirm or deny confidence in results retrieved from the GeoNames.

For biodiversity background data, we use **The Global Biodiversity Information Facility (GBIF)**[14]. GBIF is the largest online portal for biodiversity data. As of November 2011, the portal contains 312 million records, of which

[12] `http://www.geonames.org`
[13] `http://maps.google.com`
[14] `http://www.gbif.org`

271 million also contain coordinates. These records come from the combination of many individual datasets provided by institutions from around the world. A study on the accuracy of geographical data in GBIF records [23] showed that the majority of the records were annotated with correct coordinates (83%), but the relatively large amount of incorrectly georeferenced records is something that has to be taken into account when using this data. In our approach we therefore do not use GBIF as a gold standard to derive exact coordinates from, but we use it to filter out outliers (see subsection 6.3).

5 Knowledge-driven Georeferencing Approach

Our georeferencing approach consists of 5, rule-based modules that form the pipeline through which each of the 150 evaluation records from the gold standard is processed. All modules are automatic. However, the final result of the georeferencing approach is presented to a researcher who needs to check the systems result before it is included in the database.

1. **Record Retrieval** This module filters the database record to include only those database fields used by the system ("Town/City", "Province/State", "Country", "Location", "Altitude", "Collection Date", "Genus", and "Species")

2. **Text Parsing** In the parsing module, sentences are split and tokenized. Then tokens are matched against patterns and keywords to recognize indicators for offsets (such as cardinal directions and units of measurement), place names and common words in Dutch and English.

3. **Gazetteer Lookup** Identified location name candidates from the text parsing module are looked up in GeoNames and Google Maps.

4. **Offset Calculation** If an offset, such as "112 km S El Dorado" is encountered, coordinates retrieved from the gazetteer for the place of reference

("El Dorado") need to be combined with the offset ("112 km South") to calculate the final coordinates. For the calculations of the coordinates we use the Perl Geo::Calc[15] module. An example is shown in Figure 1.

5. **Disambiguation Heuristics** As many place names share the same name ("Amsterdam, Netherlands" vs. "Amsterdam, MO, US") or similar names ("York, UK" vs. "New York, US") several disambiguation heuristics were selected to disambiguate location names.

[Figure 1 about here.]

6 Disambiguation Heuristics

In this section, we will detail each of our disambiguation heuristics.

6.1 Spatial Minimality

The spatial minimality heuristic is a fairly standard statistic in georeferencing and relies on co-occurrence of geographic entities within the same discourse. This heuristic assumes that, in a text which mentions more than one location, the cluster of physical locations in the world that are most closely related by distance are the most likely candidates to be actually referred to. For example, if "Amsterdam" and "Utrecht" are mentioned in the same text it is assumed that Amsterdam refers to "Amsterdam, NH, Netherlands" whereas if "Amsterdam" is mentioned together with "Albany", it is more likely to refer to "Amsterdam, NY, United States". Li et al. [13] use a maximum weight spanning tree (MST) to determine the best candidate based on its closest mentioned neighbours. Leidner et al.[12] use an approach based on finding the smallest polygon that binds a set of candidates to achieve a similar result. We follow the more

[15] `http://search.cpan.org/~asp/Geo-Calc-0.11/lib/Geo/Calc.pm`

common approach from Leidner et al. [12] and compute polygons that span each combination of potential candidates for a location name, and select the smallest polygon. We start with a list of potential candidates for each place name and their corresponding coordinates and match each candidate to every possible combination of candidates from the other place names. For each of these combinations the system creates a polygon that encloses these candidates. The system selects the smallest polygon, and the set of candidates used to create that polygon are seen as the most likely candidates.

6.2 Expedition Clusters

The spatial minimality heuristic uses only information from within individual records. However, specimen database records are not independent. The Expedition Clusters heuristic assumes that information from similar records can be used to aid georeferencing. Work on this same dataset by Van Erp [20] shows that it is possible to use information available in the dataset to rediscover expeditions from a dataset. Information about which expedition a record belongs to is only explicitly available in a small number of records, but it is "re-discovered" by using data such as collection date and country. Enriching the data in such a way enables comparison between records which would otherwise not be possible. For example, it is very unlikely that two records from the same expedition are in entirely different locations. An example expedition plot is shown in Figure 2. Thus, if such an anomaly was to be detected it would be a clear signal that one of the records is incorrectly georeferenced. Furthermore, the information can be used for disambiguation of place names as also suggested in the work of Guo et al. [6], and increase confidence in the outcome of the georeferencing process.

Van Erp [20] found that grouping records by collection dates alone was very efficient already (.83 F-Measure). We also added country information to maxi-

mize precision. Before implementation, the date field in the dataset was standardized to the format "YYYY-MM-DD" and the retrieved records were subsequently ordered by date and country. Records with incorrect or incomplete collection dates (e.g. "30-02-1960" and "1960"), as well as records without country information were not processed in this heuristic. A candidate for a place name that is close to the previous georeferenced location record (when that record belongs to the same expedition) will be assigned a higher confidence measure.

[Figure 2 about here.]

6.3 Species Occurrence Data

Occurrence data from existing specimen finds can be used to check if new data fits the currently known locations for species. In the current implementation, this data is retrieved solely from GBIF as, at the time of writing, this is the only openly available resource containing such information. This information is used to disambiguate location descriptions and validate results in much the same way as the expedition heuristic. By querying GBIF data, coordinates are retrieved for all currently known finds of the species in the record. Each coordinate for a previously found specimen find is then compared to each place candidate, and based on the closest specimen find to a candidate a confidence measure is assigned to the candidate; the smaller the distance to a candidate the higher the confidence. The confidence measure is detailed in Section 8. A visual representation of the GBIF disambiguation heuristic is shown in Figure 3.

[Figure 3 about here.]

7 Results and Discussion

[Table 3 about here.]

[Table 4 about here.]

All presented results are measured by applying the heuristics in our knowledge-driven georeferencing approach to the 150 evaluation records that were manually georeferenced for the gold standard (see Subsection 4.2). We computed a baseline score to compare our approach to a simple look-up approach by retrieving the coordinates of the first location name found in the record, looking up this name in the GeoNames gazetteer, filtering by country and province and returning coordinates of the first candidate. Table 3 presents the accuracy results of the different modules on the test set. Table 4 presents the precision (percentage of correctly georeferenced records), recall (percentage of records for which the system suggested coordinates) and F-measure (the harmonic mean of precision and recall) of the best system at 25km. Application of the t-test shows that all modules provide significant improvement over the baseline at $p < 0.005$

The spatial minimality heuristic improves results for records that contain more than one place name (which is the case for around 50% of the records in our gold standard), but with some caveats. The first implementation included each location found in the record (Place, Location and Province/State). However, because the location field is a free text field, it contains long sentences in a number of records, negatively affecting the rule-based system to recognise location names. However, in many other cases, the location field does contain useful information, so it was decided to not parse any location fields with a length exceeding 60 characters. The spatial minimality heuristic performs better if the Province/State field is not included in this heuristic. Provinces and states generally cover larger areas but the gazetteer will return only one single

point that does not represent this fact. As such, these points do not add much information on a smaller scale and pollute the created polygons. Since the country is almost always known, this already dramatically decreases the area that has to be searched. As a result, the heuristic mainly improves results that were not too far off to begin with.

As our data is in Dutch, we could not run our data in BioGeoMancer and GeoLocate. For GeoLocate it is also the case that only georeferencing in the US is supported. We could also not get hold of the data they tested their systems with, therefore an exact comparison of our system to BioGeoMancer and GeoLocate is not possible, but we have strived to set up our experiments in similar fashion. We therefore assume that our results for the spatial heuristics are in the same ballpark as those reported in the work of Murphey et al. [15].

Although the results in Tables 3 and 4 seem to indicate that the expedition heuristic does not improve the results, manual inspection of the records showed that the heuristic does add valuable information. For now this information mostly affects the confidence score (see Section 8), and we attribute the lack of improved scores to the configuration of our gold standard dataset. As our gold standard contains a sample of random records from across the entire dataset, the number of records belonging to the same expedition in this sample is small, and as such these small clusters add little evidence to support the disambiguation process.

The use of GBIF Species Occurrence Data is especially useful in situations for disambiguation of location names in a large geographical area (notice that is why the mean distance off improves more than the percentages of correctly georeferenced localities). If a specimen find is only annotated with the place name "Sibil", a list of 20 possible candidates would be retrieved from the GeoNames gazetteer in different continents. By cross-referencing these candidates with

existing finds of the species ("Sphenomorphus schultzei"), only two likely candidates remain: "Ok Sibil, Papua, ID" and "Sibil, Papua New Guinea", greatly decreasing the search space.

Care needs to be taken however that on a smaller scale, the heuristic should not be used too rigorously since it will only favour locations that fit within the existing data model and many species occurrences are spread out across an area. Furthermore, a significant part (16%) of the geographical data in GBIF records was found to contain errors, as demonstrated by Yesson et al. [23]. Species occurrence records for "Sphenomorphus schultzei" show that the species was found on multiple locations across the island "New Guinea", in an area of almost 600,000 km. In this case, the occurrence data should not be used for disambiguation of the two remaining candidates on this island.

[Table 5 about here.]

The results for different categories presented in Table 5 show that records that are annotated with one single location name and an offset (category B) are georeferenced with a much higher accuracy than other categories. Obviously, the textual complexity of these records is limited but there are two other points of interest. In each of these cases, there is no problem with the distinction between administrative areas (provinces, states) and populated places (cities, villages) since it is obvious that an offset will always be from a populated place and not from a province. Secondly, the offsets usually appear to be from a well-known (or important) place. A major difficulty in geo-referencing biological collections is the use of place names that are only locally known. The location of place names such as "Meyers' farm" or "Base Bivouac" might be very well known during expeditions and to local inhabitants. However, it is nearly impossible to use this information on its own without the use of very specific information sources such as the field logs and maps created for specific expeditions. In

specific implementations, one could consider manually creating an additional gazetteer for such places.

As can be seen in the third column, the other categories (A, C and D) have an almost similar score for correct matches within 5 km. However, results for single location names (category A) show that the number of additional places found within 25 or 100 KM is limited, whereas records with more than one location show improvements. Records that are annotated with more than one location benefit the georeferencing process by adding contextual information. For example, when encountering a description such as "Lake Jaroe, Kampong Gariau, Indonesia", "Lake Jaroe" does not occur in any generic gazetteer. However, the record can still be georeferenced using the more generic location "Kampong Gariau". However, this means records are georeferenced to locations several kilometres away from the correct location, decreasing accuracy.

8 Measuring Confidence

There is a large number of potential uncertainties in the georeferencing process. These stem from the data itself, external data-sources used, and the process of linking data to these external sources. It is important that these sources of uncertainty are identified and recorded to be able to calculate a confidence score (CS) for the resulting georeferenced locality. Although Graham et al. [5] found that "species distribution modeling approaches in general are fairly robust to locational error", not having information about the uncertainty of georeferenced localities makes it impossible to know if this geospatial data is suitable for a specific purpose and it may thus be of little use as also suggested in the work of Wieczorek et al. [22] and of Guo et al. [6].

Inspired by a basic manual confidence value system used in the MaPSteDI method [15], a scale from -12 to 12 is used to *automatically* indicate the confi-

dence in a georeferenced locality (12 indicating the highest degree of confidence, -12 lowest). This automatic measure represents the confidence that the returned coordinates for a georeferenced location are accurate. The confidence measure is based on several different indicators presented in Table 6. Each indicator contributes to the final confidence score with a different weight. These weights have been determined from manual adjustment based on the development dataset. We will investigate this issue in depth in the future.

The most important component of the confidence score is the amount of information available in a record. A single place name with structured additional information about the province and country such as "Santa Bárbara, Amazonas, Brazil" can usually be retrieved with a higher confidence than a single description such as "Forest between 20-10 km from Ambohaobe". Therefore the latter record receives a lower CS based on absence of country and province information. Secondly, it is based on the consistency and type of input data from gazetteers and biodiversity resource. For example, if no direct match in a gazetteer is found but a result is found using fuzzy matching, that result will still used but it decreases the confidence. If a georeferenced location is consistent with existing occurrence data from GBIF, this will increase the confidence.

Each heuristic can increase or decrease the confidence. For example, based on the spatial minimality heuristic, the confidence will be increased if the polygon describing the area of co-occurring place names is very small or decreased if very large. If a record belonging to the same expedition is georeferenced to a location that is close to other specimen finds from that same expedition, the confidence is also increased. For instance, in our dataset we have a record for which: the country is known (+2), the province is unknown (0), the location description contains unknown words (-1), the place description is found in GBIF (+1), but a fuzzy search in the gazzetteers does not return a positive result (-3) and the

place description is found only in Google Maps (-3), gets a confidence score of -4. The fact that the distance of the georeferenced location of this entry from its actual location is approximatively 1,204.72 meters, confirms the indication given by the low confidence score.

The extent to which certain variables influence the accuracy cannot always be determined and as such make the method not infallible. In some cases, there is simply not enough information to determine an indicative confidence measure. To estimate the reliability of our confidence measure, we treat it as an estimated observation about the correctness of the corresponding georeferenced entry. Similarly to the works of Jøsang [8] and Ceolin et al. [2], this estimated evidence is used to build a Beta probability distribution that describes the probability of each confidence score in the interval $[0 \dots 1]$ to represent the trustworthiness of the entry.

More in detail, each heuristic is utilized as evidence for the correctness of the estimate: when the heuristic provides a positive value (e.g. the country of a record is known), this counts as positive evidence; when the heuristic provides a negative score (e.g. in case of unknown parts in the record), this counts as negative evidence. In fact, each heuristic can be seen as an indication of the possibility to correctly georeference the record. The more heuristics positively indicate the possibility to correctly geolocate, the more confident we will be about the geolocation. We aggregate all heuristics (per record) by building a Beta probability distribution. This distribution is shaped by two parameters, corresponding to the amount of positive and negative evidence (each plus one). For a comprehensive list of positive and negative values of the heuristics, we refer the reader to Table 6. The weights contributed by each heuristic are aggregated in positive and negative evidence counts as follows: if an heuristic ranges between -2 and 2 and the value of that heuristic is -1, we count 1 positive

piece of evidence and 3 negative ones. So, if max is the upper bound of an heuristic (e.g. 2) and min its lower bound (e.g. -2), given the value h of the heuristic, we compute the corresponding positive and negative evidence p and n as follows:

$$p = h - min$$

$$n = max - h$$

. We aggregate all the evidence for all the heuristics, and then we compute the expected value of the resulting Beta distribution. The Beta distribution is a probability distribution that ranges between 0 and 1 and we use it to estimate the probability of each value in the [0,1] interval to represent the probability for the georeferencing to be correct. If the heuristics provided a lot of negative evidence, the expected value of the Beta distribution will be close to zero, and vice-versa. Also, the variance of the distribution measures the uncertainty in this estimate, and it is therefore smaller when we are more certain about our estimate, i.e. when more heuristics are available.

A Shapiro-Wilk normality test at 95% confidence level shows that both the error in the georeferencing process and the expected values of the Beta distributions computed using the heuristics are not normally distributed, hence we use a Spearman's rank correlation test [18] at 95% confidence level to check the existence of a linear correlation between the two series of values. In particular, since the Spearman's test compares the rank between variables (without taking into account their differences), we standardize the distances and we round them (to 7 decimal digits) because we do not expect our confidence scores to be extremely precise, rather they should help us to distinguish between good and bad georeferences. The test results in a weak negative correlation (-0.22), as shown in 4. This suggests that the procedure is often able to compute a confidence

score that resembles the real trustworthiness of the result of the georeferencing process, although there is still big room for improvement. Also, another Spearman correlation test at 95% confidence level shows a weak positive correlation (0.21) between the variance of the Beta distribution based on the heuristics and the error of the georeferencing process, indicating that the more certain the score is, the lower the error.

[Table 6 about here.]

[Figure 4 about here.]

9 GeoImp Demonstrator

As a proof of concept for biologists who need to georeference their datasets, we built an online demo GeoImp[16]. This front-end can be used to georeference a single instance, or batch-reference multiple records using the CSV mode. A user can enter query terms in one or more fields in the interface after which the system will try to return a georeferenced location visualised on a map along with the confidence it has in its decision. Entering more information into the system decreases the ambiguity in the process. The results are shown on a map, and as coordinates together with the calculated confidence score. Currently, only Dutch and English are supported and only a limited number of records can be parsed at once, additional input methods and integration with the Naturalis collection registration system will be provided in the future. A screenshot can be found in Figure 5.

[Figure 5 about here.]

[16] `http://semanticweb.cs.vu.nl/geoimp`, GeoImp stands for Georeference Improver

10 Conclusions and Future Work

We have presented a method to automate georeferencing of records in animal specimen datasets that utilizes the spirit semantic web technology by integrating different types of information to form a richer knowledge base. Several heuristics for the disambiguation of location names that use domain knowledge from external resources and reasoning were implemented and tested. In addition, the method produces a confidence score to indicate how certain the system is of its decisions to help curators select records to inspect manually. Using our prototype, experiments using a manually created gold standard were carried out to test the impact of the heuristics on the georeferencing process. We have shown that domain-specific knowledge such as occurrence data from a biodiversity resource contributes to more accurate results.

The complexity of the georeferencing task is high. A substantial amount of specimen finds are not annotated with enough information to return accurate coordinates, and generic gazetteers are only partially suited for the natural history domain as they often lack information on location names mentioned in locality descriptions. However, our confidence measure proves useful in these cases, pointing experts at Naturalis to these records so they can focus their attention on those cases that require input from a human expert. The effectiveness of our confidence measure in correctly representing the precision of the georeferencing process with respect to each entry has been demonstrated by means of a statistical test.

We are of the opinion that our approach will translate well to domains other than the biology domain. As more and more structured data becomes available, for example through the Linked Open Data cloud, integration of data becomes more feasible. Furthermore, many institutions struggle with a lack of resources to manually georeference their data, for which an approach that can resolve

simpler cases automatically and pinpoints the cases where human knowledge is required would be useful. We have shown in previous work that such a semi-automatic system works well for other tasks too such as data cleaning [19].

In future research, we will investigate incorporating additional resources to cope with location names described in different languages such as "Midden Java" (Dutch for "Central Java") and places that would not occur in generic gazetteers (such as historic names or base camps for expeditions which are recorded at the institution). We will also look into incorporating linear feature types such as rivers and roads in our offset calculations. The geographical resources that we used did not specifically record such features, but other resources such as Open Street Map[17] could help resolve this. When we have access to linear features, the text parsing module can also be expanded as currently the system cannot accurately interpret "On the road between place X and Y". Furthermore, the expedition heuristic shows interesting possibilities for new research, as it utilizes the fact that individual records in a database are dependent on each other. To further test this feature, we intend to use a slightly different experimental setup with a less random data sample, to make sure we have enough instances belonging to an expedition to test our assumption that the interdependence of records may aid the disambiguation process. Also when the system is integrated in the workflow researchers will continuously update the database with georeferenced records which will grow the gold standard dataset, enabling expanding to different subdomains within biology and further finetuning of the system.

Our results presented here show that there is still much to be gained by combining domain specific knowledge for georeferencing.

[17] `http://www.openstreetmap.org/`

Acknowledgments

This work was funded by NWO in the CATCH programme, grant 640.004.801.

References

[1] R. Beaman and B. Conn. Automated geoparsing and georeferencing of malesian collection locality data. *Telopea*, 10(1):43–52, 2003.

[2] Davide Ceolin, Willem Robert van Hage, Wan Fokkink, and Guus Schreiber. Estimating uncertainty of categorical web data. In *Proceedings of the 7th International Workshop on Uncertainty Reasoning for the Semantic Web (URSW 2011)*, pages 15–26. CEUR-WS, 2011.

[3] Dong-Po Deng, Tyng-Ruey Chuang, Kwang-Tsao Shao, Guan-Shuo Mai, Te-En Lin, Rob Lemmens, Cheng-Hsin Hsu, Hsu-Hong Lin, and Menno-Jan Kraak. Using social media for collaborative species identification and occurrence: issues, methods, and tools. In *Proceedings of the 1st ACM SIGSPATIAL International Workshop on Crowdsourced and Volunteered Geographic Information*, GEOCROWD '12, pages 22–29, New York, NY, USA, 2012. ACM.

[4] Julio Godoy, John Atkinson, and Andrea Rodriguez. Geo-referencing with semi-automatic gazetteer expansion using lexico-syntactical patterns and co-reference analysis. *International Journal of Geographical Information Science*, 25(1):149–170, 2011.

[5] C. Graham, J. Elith, R. J. Heijmans, A. Guisan, A. Townsend Peterson, and B. Loiselle. The influence of spatial errors in species occurrence data used in distribution models. *Journal of Applied Ecology*, 45(1):239–247, Feb. 2008.

[6] Q. Guo, Y. Liu, and J. Wieczorek. Georeferencing locality descriptions and computing associated uncertainty using a probabilistic approach. *International Journal of Geographical Information Science*, 22(10):1067–1090, 2008.

[7] R. P. Guralnick, J. Wieczorek, R. Beaman, and R. J. Heijmans. BioGeomancer: Automated georeferencing to map the world's biodiversity data. *PLoS Biology*, 4(11):1908–1909, 2006.

[8] Audun Jøsang. A logic for uncertain probabilities. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, 9(3):279–212, 2001.

[9] Roula Karam and Michele Melchiori. Improving geo-spatial linked data with the wisdom of the crowds. In *Proceedings of the Joint EDBT/ICDT 2013 Workshops*, EDBT '13, pages 68–74, New York, NY, USA, 2013. ACM.

[10] Tomi Kauppinen, Riikka Henriksson, Reetta Sinkkilä, Robin Lindroos, Jari Väätäinen, and Eero Hyvönen. Ontology-based disambiguation of spatiotemporal locations. In *Proceedings of the 1st international workshop on Identity and Reference on the Semantic Web (IRSW2008), 5th European Semantic Web Conference*, 2008.

[11] Jochen L. Leidner and Michael D. Lieberman. Detecting geographical references in the form of place names and associated spatial natural language. *SIGSPATIAL Special*, 3(2):5–11, July 2011.

[12] Jochen L. Leidner, Gail Sinclair, and Bonnie Webber. Grounding spatial named entities for information extraction and question answering. In *Proceedings of the HLT-NAACL 2003 workshop on Analysis of geographic references (HLT-NAACL-GEOREF'03)*, pages 31–38, 2003.

[13] H. Li, R. K. Srihari, C. Niu, and W. Li. InfoXtract location normalization: a hybrid approach to geographic references in information extraction. In *Proceedings of the HLT-NAACL 2003 workshop on Analysis of geographic references (HLT-NAACL-GEOREF'03)*, pages 39–44, 2003.

[14] Vitor Loureiro, Ivo Anastácio, and Bruno Martins. Learning to resolve geographical and temporal references in text. In *Proceedings of the 19th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems*, GIS '11, pages 349–352, New York, NY, USA, 2011. ACM.

[15] P. C. Murphey, R. P. Guralnick, R. Glaubitz, D. Neufield, and J. A. Ryan. Georeferencing of museum collections: A review of the problems and automated tools, and the methodology developed by the mountain and plains spatial-temporal database-informatics initiative (MaPSTeDI). *Phyloinformatics*, 1:1–29, 2004.

[16] Nelson E. Rios and Jr. Henry L. Bart. Geolocate users's manual. `http://www.museum.tulane.edu/geolocate/standalone/manual_ver2_0.pdf`.

[17] Kirk Roberts, Cosmin Adrian Bejan, and Sanda M. Harabagiu. Toponym disambiguation using events. In Hans W. Guesgen and R. Charles Murray, editors, *FLAIRS Conference*. AAAI Press, 2010.

[18] C. Spearman. The proof and measurement of association between two things. *American Journal of Psychology*, 15:88–103, 1904.

[19] Antal Van den Bosch, Marieke Van Erp, and Caroline Sporleder. Making a clean sweep of cultural heritage. *IEEE Intelligent Systems*, 24(2):54–63, March/April 2009. Special Issue on Cultural Heritage.

[20] Marieke van Erp. Retrieving lost information from textual databases: Rediscovering expeditions from an animal specimen database. *Proceedings of the ACL 2007 Workshop on Language Technology for Cultural Heritage Data (LaTeCH 2007)*, 2007.

[21] Marieke van Erp. *Accessing Natural History: Discoveries in Data Cleaning, Structuring, and Retrieval*. PhD thesis, Tilburg University, Tilburg, the Netherlands, June 2010.

[22] J. Wieczorek, Q. Guo, and R. J. Heijmans. The point-radius method for georeferencing locality descriptions and calculating associated uncertainty. *International Journal of Geographical Information Science*, 18(8):745–768, 2004.

[23] C. Yesson, P. W. Brewer, T. Sutton, N. Caithness, J. S. Pahwa, M. Burgess, W. A. Gray, R. J. White, A. C. Jones, F. A Bisby, and A. Culham. How global is the global biodiversity information facility? *PLoS One*, 2(11), 2007.

List of Figures

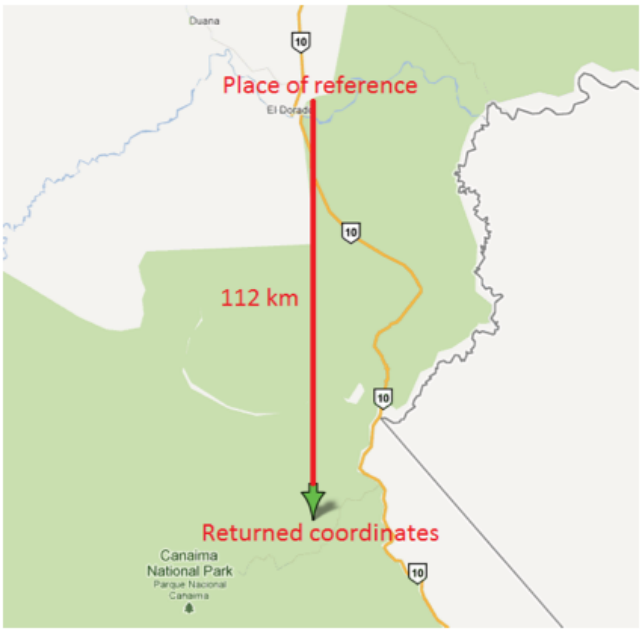

Figure 1: Visualization of an offset from a named place

Figure 2: Automatically georeferenced cluster of locations belonging to an expedition on the right

a

b

c

Figure 3: Visualization of the GBIF disambiguation heuristic. Figure **a** shows the GeoNames gazetteer candidates, Figure **b** shows the GBIF occurrence data. Figure **c** shows the search space (red box) after combining the GeoNames candidates with the GBIF occurrence data.

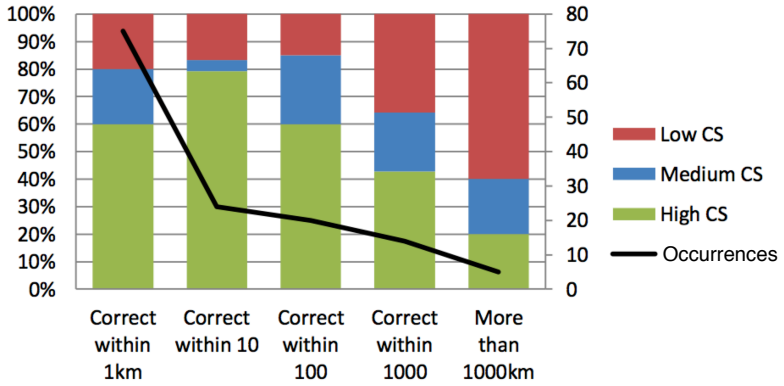

Figure 4: Distribution of confidence scores for georeferenced locations of the gold standard. The black line displays the number of occurrences in each category (right axis scale)

GeoImp

Results for 1 record(s) can be found below

Switch to Batch Input Mode

Find Single Record

Find coordinates for place name descriptions

Place e.g. Amsterdam	5km north of Amsterdam
Province e.g. North Holland	
Country e.g. Netherlands	
Additional location e.g. in the Vondelpark	
Collection Date Format: DD-MM-YYYY	
Bionomial Name e.g. Leposoma guianense	Helophorus brevipalpis
Elevation e.g. 20 m	

Geocode Record

Found: 52.37403 / 4.88969 Confidence Score: 4

Figure 5: Screenshot of the GeoImp Demonstrator

List of Tables

Challenge posed	Example textual locality
A1. Two or more locations that share the same name	Amsterdam
A2. Vague localities	Southeast Michigan
B1. Linear feature measurement	16 km (by road) N of Murtoa
C1. Two or more location descriptors	Wakarusa, 24 mi WSW of Lawrence
C2. Topological nesting	Moccasin Creek on Hog Island
C3. Linear ambiguity	On the road between Sydney and Bathurst
D1. Complex interpretative description	Bupo [?Buso] River, 15 miles [24 km] E of Lae
E1. Political borders change over time	Yugoslavia
E2. Historical place names	British North Borneo

Table 1: Georeferencing challenges and examples from the Naturalis collection

Category	Example	#in gold standard	# in full set
A. Single Place	"Maastricht"	90 (45%)	10,750 (42,7%)
B. Single Place with offset	"18 mls. E. of Kumasi"	20 (10%)	2,363 (9.4%)
C. Two or more places	"Sibil, Sterrengebergte"	62 (31%)	9,150 (36.4%)
D. Two or more places with offset	"Alachua Co., 10 mi S. Gainesville on Wachahoota rd."	28 (14%)	2,856 (11.3%)
Total	-	200	25,119

Table 2: Categories for different types of textual descriptions

	Correct @5km	Correct @25km	Correct @100km	Mean distance off	Not Found
Baseline	38.9	47.0	58.4	251.1km	26.2
+ Google Maps & Fuzzy match	53.0	65.1	74.5	244.1km	8.7
+ Spatial Heuristics	59.1	71.8	77.2	171.1km	7.4
+ Expeditions	59.1	71.8	77.2	171.1km	7.4
+GBIF	61.7	74.5	79.9	114.5km	7.4

Table 3: Accuracy of the georeferencing heuristics within 5km, 25km and 100km of the gold standard coordinates, compared to baseline in percentages. The table also shows the mean distance the different heuristics were off, as well as the percentage of cases for which no coordinates were found by the system

	Precision	Recall	F_1
baseline	0.64	0.47	0.54
+ Google Maps & Fuzzy match	0.71	0.65	0.68
+ Spatial Heuristics	0.78	0.72	0.75
+Expeditions	0.78	0.72	0.75
+GBIF	0.80	0.74	0.77

Table 4: Precision, recall and F-measure of the different heuristics at 25km from the coordinates in the gold standard.

Category	5km	25km	100km	Mean distance Off	No Result
A: Single Location (67)	58.2	64.7	68.7	140.1	16.4
B: Single Location + offset (15)	86.7	100	100	1.7	0
C: Multiple Locations (46)	60.9	73.9	82.6	146.4	0
D: Multiple Locations + offset(s) (21)	57.1	85.7	95.2	54.9	0

Table 5: Results split out per category based on best results from Table 3 (GeoNames + Google Maps + fuzzy search + spatial heuristics + expeditions + GBIF). The numbers behind the categories indicate the number of records in that category.

Level	Indicator	Points
Record	Country known	+2
Record	Province known	+1
Record	Unknown parts in description	-1
GeoNames Result	Place not part of province	-1
GeoNames Result	Fuzzy string search	-3
Candidate	(SM) Close together	+x,x $\in$ [0...2]
Candidate	(OD) Close to GBIF	+x,x $\in$ [$-$1...2]
Candidate	(EXP) Close to previous find	+x,x $\in$ [$-$2...2]
Candidate	GeoNames Candidate very close to Google Maps	+x,x $\in$ [0...2]
Candidate	Only found on Google Maps	-3
Candidate	GeoNames first candidate	+1
Candidate	Administrative area	-1

Table 6: Calculation of the Confidence Score. SM denotes the spatial minimality heuristic, EXP denotes the expedition heuristic, and OD indicates the use of species occurrence data. In some cases, a range of values can be deducted or added.