Improving geographical data in animal specimen datasets

Research plan

By: *R. Hensel*
Supervisor: *Marieke van Erp*
Second reader: *Davide Ceolin*

VU UNIVERSITY AMSTERDAM

Context

NCB Naturalis, the Netherlands Centre for Biodiversity[1] is an organization whose mission it is to keep an open archive of life's biodiversity. The organization keeps several large databases of animal specimen finds. The information in these databases comes from biologists that made these finds and recorded them manually in field logs that were eventually converted to electronic databases. This information is now being used for biodiversity research, for example to track a species geographical distribution over time.

Among the information in these records (each referring to a specific find) is typically locality information about the places where these specimens were found, a biological specification, the name of the researcher that made the discovery, and additional notes. The database is structured; all pieces of information are divided into several distinct fields. Within the fields, the data is sometimes (semi) structured, but little integrity constraints are enforced. Most of the fields are not bound to a specific format and some of them are unstructured text fields such as the location field and additional notes, which contains natural language descriptions.

Problem statement

The geographical information on animal specimen finds is, especially in older records, not recorded in a structured way and these records are often incomplete (e.g., missing fields) or imprecise (e.g., ambiguous geographical references). The original records from logbooks were eventually converted to electronic databases. During this manual data-entry process some misspellings and other inconsistencies can have been introduced as well.

For biologists that use these records to track movements of species, having a structured and more accurate description of geographical data would benefit their work.

Therefore, the main objective of this research is to specify more accurately the location of specimen finds to aid biodiversity research. Where possible, this is done by specifying coordinates of the locations in the databases. In order to do so, the locality information needs to be extracted and disambiguated by reasoning over this information within the databases and linking this to external public information sources and animal taxonomies. The public resources to be used include Geographical Information Systems (GIS), which provide geological data on objects (for example coordinates for cities) and animal taxonomies, which can provide geo-referenced information about biological specifications.

Following the above problem statement, the main question that drives this research is formulated as follows:

> *How can current incomplete and ambiguous geological data in animal specimen databases be improved by combining it with information from publicly available GIS and animal taxonomies?*

In order to answer the problem statement three research questions are investigated.

[1] http://www.NCB Naturalisnaturalis.nl/

The geographical references in the databases consist of descriptions and place names that in many cases can refer to more than one location since place names are highly ambiguous. In order to disambiguate these references contextual information in the records should be used in conjunction with external GIS and animal taxonomies. The basic problem here is to find out which fields and external information sources add most evidence to disambiguate between locations. A balance needs to be found between usefulness, reliability and difficulty in extracting the information.

1. *What are the optimal heuristics for disambiguating and improving geographical data within the NCB Naturalis databases?*

When a structured data format is used, it can be automatically interpreted by machines. There are however several text fields in the databases which contain natural language descriptions of geographic locations. In order to reason over this data and link it to other data sources, the information of interest within these fields needs to be extracted. This is commonly referred to as information extraction (IE). The issue that needs to be resolved is:

2. *How can place names and spatial offsets be properly recognized in unstructured text fields?*

An automated disambiguation process such as described above will inherently have a certain amount of uncertainty to it. In addition to inconsistency and incompleteness in the databases errors can be made when linking the data to external sources. Therefore, part of the end result needs be a confidence score that gives an indication of the reliability of the outcome.

3. *Can a useful confidence score for the outcome of the approach taken be determined?*

Method

The research will follow a mainly empirical approach. For each research question a literature review will be carried out in order to investigate current best practices and standards. Based on the specific context of the NCB Naturalis databases an appropriate method or technique will then be selected. The implementation of these methods and techniques will likely involve constructive research as well, as it is expected that the there is room for improvements in this specific context. After implementation and possible adjustment of the selected methods and techniques the precision and recall (F-measure) of this approach can be measured. This process can be repeated several times with different methods and algorithms until an acceptable result is achieved.

In order to improve the current geographical data, appropriate data from the NCB Naturalis database(s) will be linked with GIS and animal taxonomies to form an end result consisting of a footprint of the location described in the database and a confidence measure. Assigning coordinates to place indications is commonly referred to as geoparsing [2]. Georeferencing

The idea to automate the geoparsing of specimen databases is not new. A significant contribution in this field was made with the development of BioGeomancer [3, 4]. This project provides a set of tools (based on best practices) which provide functionality for natural language processing, interpreting and gazetteer queries, intersecting spatial descriptions and finally returning a standardized geological reference including uncertainty levels. The project does not seem to be under active development anymore and (as a result) it does not have support for multiple languages or querying additional resources such as animal taxonomies. In the specific context of this research, including these taxonomies can improve the disambiguation process as these can contain geo-references that can delimit the possible locations. The BioGeomancer project can thus be used as a stepping stone for this research, as many of the principles used apply to the steps in this research.

Global schedule

March 1 – Start thesis

May 1 – Research plan/proposition finished

May 10 – Global literature review complete

May 20 – Finished converting databases to workable format (PostgreSQL[2])

June 10 – Finish investigating data (heuristics) and created initial program to parse input

July 1 – Finish program to consult external sources and calculate confidence score

August 1 –First version of report on comparison between different data and algorithms

September 1 – Finish report and possibly expand results to other databases

[2] http://www.postgresql.org/

References

[1] Huifeng, L., Srihari, R. K., Niu, C. Li, W (2002, August). Location Normalization for Information Extraction. Proc. 19th COLING, Taipei, Taiwan.

[2] Larson, R. R. (1996). "Geographic Information Retrieval and Spatial Browsing". In Smith and Gluck (eds.): “Geographic Information Systems and Libraries: Patrons, Maps, and Spatial Information” (Graduate School of Library and Information Science, University of Illinois at Urbana-Champaign).

[3] Beaman R., Conn, B (2003). Automated geoparsing and georeferencing of Malesian collection locality data. Telopea 10.

[4] Guralnick, R. P., Wieczorek, J., Beaman, R., Hijmans, R. J., the BioGeomancer Working Group (2006). BioGeomancer: Automated Georeferencing to Map the World’s Biodiversity Data. Public Libr. Sci. Biol. 4.

[5] Murphey, P.C., Guralnick, R. P., Glaubitz, R., Neufeld, D., Ryan, J.A. (2004). 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.