BISS Biodiversity Information Science and Standards

Biodiversity Information Science and Standards 3: e37345
doi: 10.3897/biss.3.37345

OPEN ACCESS

Conference Abstract

Comparison of Automated Georeferencing Tools Using Insect Collection Data

Leonor Venceslau[‡,§], Luis Filipe Lopes[|,¶]

‡ Faculdade de Ciências da Universidade de Lisboa, Lisboa, Portugal
§ Museu Nacional de História Natural e da Ciência, Universidade de Lisboa, Lisboa, Portugal
| Global Health and Tropical Medicine, GHTM, Instituto de Higiene e Medicina Tropical, IHMT, Universidade Nova de Lisboa, UNL, Rua da Junqueira 100, 1349-008 Lisboa, Portugal
¶ Centre for Ecology, Evolution and Environmental Changes (cE3c). Faculdade de Ciências da Universidade de Lisboa, Campo Grande 1749-016 Lisboa, Portugal

Corresponding author: Leonor Venceslau (fc50637@alunos.fc.ul.pt)

Received: 14 Jun 2019 | Published: 10 Jul 2019

Citation: Venceslau L, Lopes LF (2019) Comparison of Automated Georeferencing Tools Using Insect Collection Data. Biodiversity Information Science and Standards 3: e37345. https://doi.org/10.3897/biss.3.37345

Abstract

Major efforts are being made to digitize natural history collections to make these data available online for retrieval and analysis (Beaman and Cellinese 2012). Georeferencing, an important part of the digitization process, consists of obtaining geographic coordinates from a locality description. In many natural history collection specimens, the coordinates of the sampling location are not recorded, rather they contain a description of the site. Inaccurate georeferencing of sampling locations negatively impacts data quality and the accuracy of any geographic analysis on those data. In addition to latitude and longitude, it is important to define a degree of uncertainty of the coordinates, since in most cases it is impossible to pinpoint the exact location retrospectively. This is usually done by defining an uncertainty value represented as a radius around the center of the locality where the sampling took place.

Georeferencing is a time-consuming process requiring manual validation; as such, a significant part of all natural history collection data available online are not georeferenced. Of the 161 million records of preserved specimens currently available in the Global Biodiversity Information Facility (GBIF), only 86 million (53.4%) include coordinates. It is therefore important to develop and optimize automatic tools that allow a fast and accurate georeferencing.

The objective of this work was to test existing automatic georeferencing services and evaluate their potential to accelerate georeferencing of large collection datasets. For this end, several open-source georeferencing services are currently available, which provide an application programming interface (API) for batch georeferencing. We evaluated five programs: Google Maps, MapQuest, GeoNames, OpenStreetMap, and GEOLocate. A test dataset of 100 records (reference dataset), which had been previously individually georreferenced following Chapman and Wieczorek 2006, was randomly selected from the *Museu Nacional de História Natural e da Ciência*, *Universidade de Lisboa* insect collection catalogue (Lopes et al. 2016). An R (R Core Team 2018) script was used to georeference these records using the five services. In cases where multiple results were returned, only the first one was considered and compared with the manually obtained coordinates of the reference dataset. Two factors were considered in evaluating accuracy:

1. Total number of results obtained and
2. Distance to the original location in the reference dataset.

Of the five programs tested, Google Maps yielded the most results (99) and was the most accurate with 57 results < 1000 m from the reference location and 79 within the uncertainty radius. GEOLocate provided results for 87 locations, of which 47 were within 1000 m of the correct location, and 57 were within the uncertainty radius. The other 3 services tested all had less than 35 results within 1000 m from the reference location, and less than 50 results within the uncertainty radius. Google Maps and Open Street Map had the lowest average distance from the reference location, both around 5500 m. Google Maps has a usage limit of around 40000 free georeferencing requests per month, beyond which the service is paid, while GEOLocate is free with no usage limit. For large collections, this may be a factor to take into account.

In the future, we hope to optimize these methods and test them with larger datasets.

Keywords

georeferencing, data digitization, natural history collections

Presenting author

Luis Filipe Lopes

Presented at

Biodiversity_Next 2019

Funding program

FCT for funds to GHTM – UID/Multi/04413/2013

FCT for funds to CE3C - UID/BIA/00329/2013

Hosting institution

Museu Nacional de História Natural e da Ciência, Universidade de Lisboa, Lisboa, Portugal

References

- Beaman RS, Cellinese N (2012) Mass digitization of scientific collections: New opportunities to transform the use of biological specimens and underwrite biodiversity science. ZooKeys 209: 7-17. https://doi.org/10.3897/zookeys.209.3313
- Chapman A, Wieczorek J (2006) Guide to Best Practices for Georeferencing. Copenhagen: Global Biodiversity Information Facility URL: http://www.gbif.org/orc/?doc_id=1288
- Lopes LF, Correia AM, Almaça A, Verdasca MJ, Silva PB (2016) Insect Collection from the Museu Nacional de História Natural e da Ciência, Universidade de Lisboa, Portugal. https://www.gbif.org/dataset/79673413-746f-48f2-bd8a-7cf27807317e. Accessed on: 2019-4-05.
- R Core Team (2018) R: A language and environment for statistical computing. 3.5.0. URL: https://www.R-project.org/