Fun and Easy Georeferencing with a New Online Tool from South Africa

Ian Engelbrecht ‡

‡ Natural Science Collections Facility, Pretoria, South Africa

Corresponding author: Ian Engelbrecht (ianicus.za@gmail.com)

Abstract

A new online tool for georeferencing specimen records has been developed through the Natural Science Collections Facility (NSCF) in South Africa to address the need for rapid, high quality georeferencing of specimen collections in the region (Fig. 1). A dataset of specimen records with Darwin Core fields dwc:scientificName, dwc:country, optional dwc:stateProvince, dwc:locality or dwc:verbatimLocality, optional dwc:recordedBy, and a record identifier such as dwc:occurrenceID (see dwc.tdwg.org/terms for definitions) is first uploaded into the tool and a team of georeferencers then work to georeference the dataset. Fuzzy string matching is used to group similar locality strings and to search for potential matching georeferences from a georeference database. The tool aims to improve efficiency by storing georeferenced localities so that they can be reused when the same locality is encountered again in other datasets. Thus, a locality only needs to be georeferenced once, and that georeference is reused for any other permutations of that locality string. A georeference includes the most important metadata from the Darwin Core standard: a measure of uncertainty, dwc:georeferenceDate, dwc:georeferencedBy, dwc:georeferenceProtocol, dwc:georeferenceSources, and the all too often neglected dwc:geodeticDatum. dwc:georeferencedByID is included for recording the ORCID iD of the georeferencer to facilitate attribution further down the data publication pipeline. In theory, the process of georeferencing should become more efficient with time as the georeference database grows. The georeferencing process is gamified by showing each georeferencer their own numbers of georeferenced records as they work, and they can see activity of fellow georeferencers as the dataset statistics update in real time. Dataset owners can also see overall progress with the dataset and numbers of records georeferenced by each team member, which may be useful for management purposes. Once a dataset is completed, it is downloaded with the new georeferences so these can be incorporated back into the original source database.

Within the landscape of currently available georeferencing tools the system presented here is specifically placed to facilitate the management of the georeferencing process for a dataset by a team of georeferencers. The georeferencing workflow still requires a full suite of tools for finding coordinates for localities, such as a GIS, gazetteers and online

resources, as well as a specific georeferencing protocol. It essentially replaces the use of spreadsheets for doing georeferencing, or doing georeferences directly in a collection database, which can be inefficient. Related to this, it includes a quality assurance process whereby georeferences are checked for correctness and adherence to the protocol being used, and for identifying geographic and environmental outliers for each species within the dataset. In this way the tool supports current workflows and best practices for georeferencing (e.g. Chapman and Wieczorek (2020), Zermoglio et al. (2020)). The technology stack includes Firebase as the primary database, ElasticSearch for fuzzy string matching, and the user interface is built with the modern Javascript framework Svelte. The tool has been in use by the NSCF since April 2021 after being populated with approximately 300 000 existing georeferences for southern Africa from various sources, including the South African National Biodiversity Institute (SANBI) Gazetteer and several collections databases. While initial emphasis in developing the tool has focussed on southern Africa, the tool can be extended to other regions easily. Please contact data@nscf.org.za for further information.

Keywords

fuzzy string matching, gamification, efficiency, collaboration

Presenting author

Ian Engelbrecht

Presented at

TDWG 2021

Acknowledgements

Testing was carried out by the Natural Science Collections Facility team including Bronwynne Petersen, Ayanda Mnikathi, Fezile Mathenjwa, Given Leballo, Ketelo Dinala, Mahlatse Kgatla, and Maxine Manickum. Hester Steyn provided invaluable inputs during development and testing. Gail Kampmeier and Maxim Shashkov provided comments that greatly improved the first draft of this abstract.

Hosting institution

South African National Biodiversity Institute

Author contributions

The presenting author developed and implemented the system that is being presented.

Conflicts of interest

There is no conflict of interest

References

- Chapman A, Wieczorek J (2020) Georeferencing Best Practices. GBIF Secretariat, Copenhagen, 107 pp. https://doi.org/10.15468/doc-gg7h-s853
- Zermoglio PF, Chapman AD, Wieczorek JR, Luna MC, Bloom DA (2020) Georeferencing Quick Reference Guide. GBIF Secretariat, Copenhagen, 67 pp. https://doi.org/10.35035/e09p-h128

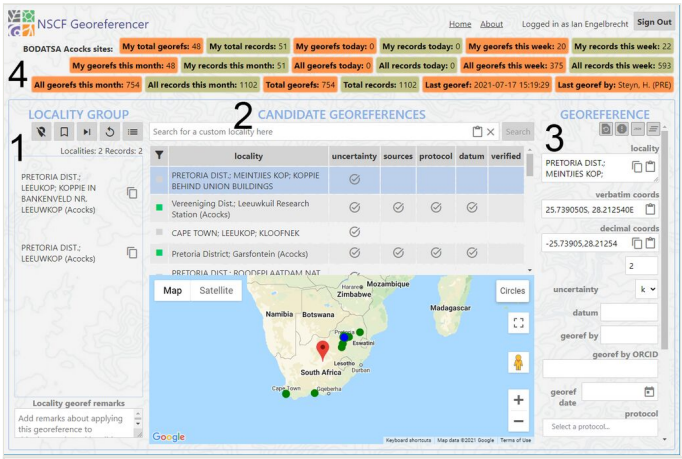

Figure 1.

The user interface for the georeferencing tool, showing:

1. a group of similar localities to be georeferenced,
2. possible matching georeferences based on fuzzy string matching,
3. locality, coordinates, and metadata fields for a selected georeference or for creating a new georeference to use for the localities in 1, and
4. realtime statistics for the current dataset.