Mobilizing and Enhancing Legacy Biodiversity Data: The case of Karl Wilhelm Verhoeff's correspondence

Carlos A. Martínez-Muñoz[‡,§], Dorothee Huff[|], Marie Meister[¶,#], Christine Driller[‡]

‡ Senckenberg – Leibniz Institution for Biodiversity and Earth System Research, Frankfurt am Main, Germany
§ University of Turku, Turku, Finland
| Universitätsbibliothek Tübingen, Tübingen, Germany
¶ UMR 7044 ArcHiMèdE CNRS, Strasbourg, France
# Musée Zoologique de Strasbourg, Strasbourg, France

Corresponding author: Carlos A. Martínez-Muñoz (archilegt@gmail.com)

Abstract

A considerable amount of biological data is preserved as physical documents, the legacy of former explorers, collectors, researchers, and others. Mobilizing data from handwritten documents has been considered particularly challenging, with well-known cases such as the manual transcription of specimen labels and herbarium sheets by museum staff, or crowdsourced transcription of data card collections through online platforms.

Here we present a pipeline of open-source software that can be used to

1. automatically transcribe handwritten text,
2. make it publicly available,
3. annotate it with e.g., scientific names,
4. extract names in Darwin Core Archive (DwC-A) for third-party reuse, and
5. automatically recognize named entities in the machine-readable text.

We based our use case on the correspondence of the German zoologist Karl Wilhelm Verhoeff, related to the Myriapoda collection held at the Musée Zoologique de Strasbourg.

The documents were processed with Transkribus (Muehlberger et al. 2019), a mostly open-source virtual research environment (OS VRE), which allows text in images to be converted into machine-readable text amenable to semantic enrichment. We achieved a character error rate as low as 5%, a remarkable result for handwritten material, as an accuracy higher than 95% for printed material is acceptable (Deutsche Forschungsgemeinschaft 2016). We then used Myriatrix (Martínez-Muñoz 2019), an instance of the Scratchpads OS VRE (Smith et al. 2011), to create bibliographic references, publish the full text, and annotate the correspondence with scientific names of myriapods. During the process, we added new scientific name spellings and combinations

to the taxonomic backbone of Myriatrix and exported the full taxon classification in DwC-A via the Global Biodiversity Information Facility (GBIF) for reuse by the Global Names Architecture and its open-source tools (Patterson et al. 2016, Mozzherin et al. 2017).

As a next step we are planning to subject the corrected text from Transkribus to a specific text-preprocessing workflow combining natural language processing (NLP) and machine learning (ML) techniques (Lücking et al. 2021). This includes, inter alia, a multiple annotation approach for general and bioscientific term classification in order to detect the respective entities automatically. The workflow has been developed in the framework of the Specialized Information Service Biodiversity Research (Koch et al. 2017) to make biodiversity information available via a customized and (bio-)ontology-based semantic search engine (Pachzelt et al. 2021).

We recommend our comprehensive approach to natural history institutions seeking to efficiently digitize and mobilize the rich biological data present in their archival documents.

Keywords

biodiversity informatics, Chilopoda, Diplopoda, handwritten text recognition

Presenting author

Carlos A. Martínez-Muñoz

Presented at

TDWG 2022

Grant title

Deutsche Forschungsgemeinschaft (DFG) - Project number 326061700, Ministerium für Wissenschaft, Forschung und Kunst (MWK) Baden Würtenberg - Project OCR-BW (2019-2022)

Conflicts of interest

References

- Deutsche Forschungsgemeinschaft (2016) DFG Practical Guidelines on Digitisation. DFG form 12.151 – 12/16. URL: https://www.dfg.de/formulare/12_151/12_151_en.pdf

- Koch M, Kasperek G, Hörnschemeyer T, Mehler A, Weiland C, Hausinger A (2017) Setup of BIOfid, a new Specialised Information Service for Biodiversity Research. Proceedings of TDWG 1 https://doi.org/10.3897/tdwgproceedings.1.19803
- Lücking A, Driller C, Stoeckel M, Abrami G, Pachzelt A, Mehler A (2021) Multiple annotation for biodiversity: developing an annotation framework among biology, linguistics and text technology. Language Resources and Evaluation https://doi.org/10.1007/s10579-021-09553-5
- Martínez-Muñoz C (2019) Proposal of Myriatrix, a virtual research environment for the International Society for Myriapodology. URL: https://www.myriapodology.org/newsletter/newsletter/CIMnewsletter2019.pdf
- Mozzherin D, Myltsev A, Patterson D (2017) “gnparser”: a powerful parser for scientific names based on Parsing Expression Grammar. BMC Bioinformatics 18 https://doi.org/10.1186/s12859-017-1663-3
- Muehlberger G, Seaward L, Terras M, Colutto S, et al. (2019) Transforming scholarship in the archives through handwritten text recognition. Journal of Documentation 75 (5): 954-976. https://doi.org/10.1108/jd-07-2018-0114
- Pachzelt A, Kasperek G, Lücking A, Abrami G, Driller C (2021) Semantic Search in Legacy Biodiversity Literature: Integrating data from different data infrastructures. Biodiversity Information Science and Standards 5 https://doi.org/10.3897/biss.5.74251
- Patterson D, Mozzherin D, Shorthouse DP, Thessen A (2016) Challenges with using names to link digital biodiversity information. Biodiversity Data Journal 4: e8080. https://doi.org/10.3897/BDJ.4.e8080
- Smith V, Rycroft S, Brake I, Scott B, Baker E, Livermore L, Blagoderov V, Roberts D (2011) Scratchpads 2.0: a Virtual Research Environment supporting scholarly collaboration, communication and data publication in biodiversity science. ZooKeys 150: 53-70. https://doi.org/10.3897/zookeys.150.2193