Registering biodiversity-related vocabulary as Wikidata lexemes and link their senses to Wikidata items (Topic 5)

Aim/problem/goal

A lexeme is the basic lexical unit of a language consisting of one or more words. Wikidata collects lexemes as structured data in any language. They allow for precise definitions and could potentially be used to extract meaning from texts during text mining. However, to populate Wikidata lexemes workflows are needed from text to Wikidata. This topic aims to create a workflow from TreatmentBank into Wikidata lexemes where they can then be enriched by the community.

Method

Two taxonomic treatments were selected as test input, one modern English taxonomic publication (Wongkamhaeng et al. 2020) and the other a 18th century German text on plant development (Goethe 1790). TextImager (Hemati et al. 2016) and TextAnnotator (Abrami et al. 2020) were used for Natural Language Processing (NLP) to extract biodiversity-related words and phrases, these were then uploaded as lexemes to Wikidata, there they were curated by adding information about their forms, senses and usages. We also visualised the lexemes in Ordia to assist with quality control, prioritisation and data exploration (Nielsen 2019).

Results

During the hackathon about thirty lexemes were created and annotated. An example is the word glabrous (https://www.wikidata.org/wiki/Lexeme:L593539) and the phrase deciduous forest (https://www.wikidata.org/wiki/Lexeme:L594039). These have then been annotated manually, for example by linking the lexemes ‘deciduous’ and ‘forest’ as the parts of the ‘deciduous forest’.

Conclusion

Using existing natural language processing pipelines for part-of-speech tagging and semantic annotation of a given knowledge domain seems to be a viable approach to enable the automatic recognition and extraction of biodiversity-relevant terms and to convert them into Wikidata lexemes. If this is to be scaled up, further clarification is needed on which point in the workflow the community should be involved. Should users themselves be able to upload and process texts in the NLP pipeline? Should the service, including data output, be selectable for existing corpora (e.g. Biodiversity Heritage Library, Pensoft)? Which output format should be offered to keep the data transfer to Wikidata as simple as possible? Who offers and maintains the NLP service?

References

- Abrami, Giuseppe, Manuel Stoeckel, and Alexander Mehler. 2020. ‘TextAnnotator: A UIMA Based Tool for the Simultaneous and Collaborative Annotation of Texts’. In Proceedings of the 12th Language Resources and Evaluation Conference, 891–900. Marseille, France: European Language Resources Association. https://aclanthology.org/2020.lrec-1.112.
- Goethe, W. 1790. ‘Der Versuch die Metamorphose der Pflanzen zu erklären.’ Ettinger, Gotha. https://www.projekt-gutenberg.org/goethe/metamorp/metamorp.html.
- Hemati, Wahed, Tolga Uslu, and Alexander Mehler. 2016. ‘TextImager: A Distributed UIMA-Based System for NLP’. In Proceedings of COLING 2016, the 26th International

Conference on Computational Linguistics: System Demonstrations, 59–63. Osaka, Japan: The COLING 2016 Organizing Committee. https://aclanthology.org/C16-2013.

- Nielsen, Finn Årup. 2019. 'Ordia: A Web Application for Wikidata Lexemes'. In The Semantic Web: ESWC 2019 Satellite Events, edited by Pascal Hitzler, Sabrina Kirrane, Olaf Hartig, Victor de Boer, Maria-Esther Vidal, Maria Maleshkova, Stefan Schlobach, et al., 11762:141–46. Lecture Notes in Computer Science. Cham: Springer International Publishing. https://doi.org/10.1007/978-3-030-32327-1_28.
- Wongkamhaeng, Koraon, Pongrat Dumrongrojwattana, Myung-Hwa Shin, and Chaichat Boonyanusith. 2020. 'Grandidierella Gilesi Chilton, 1921 (Amphipoda, Aoridae), First Encounter of Non-Indigenous Amphipod in the Lam Ta Khong River, Nakhon Ratchasima Province, North-Eastern Thailand'. Biodiversity Data Journal 8 (March): e46452. https://doi.org/10.3897/BDJ.8.e46452.