Knowledge extraction from archives of natural history collections

Stork, L.

Summary

Descriptive knowledge about the natural world constitutes an understanding of the various types of entities that inhabit it, how they influence and are influenced by their changing environment, and the processes that bring about their variation. Such knowledge is crucial when it comes to making better informed decisions for policies that impact the world's natural diversity, from organisms to ecosystems. For centuries, naturalists map out expeditions to biodiverse areas to describe, illustrate, and collect various living organisms, in order to acquire knowledge of biodiversity. Resulting collection objects such as specimens, field notes, species illustrations, and other resources now exist in institutes and museums across the globe. Unfortunately, many remain under-explored mostly due to their complex context-dependant nature, implicit knowledge, and physical distribution.

Bringing together the multitude of historical and present-day collections to the Web as one global natural history collection, allows for detailed spatio-temporal analyses into the natural world and changing practices in natural history. Joining and distilling knowledge from large collections of digital and physical natural history objects and storing the result as structured, globally reusable, and accessible knowledge, facilitates cooperation within the biodiversity community and therefore furthers research and the discovery of new knowledge. The Semantic Web provides a framework for storing knowledge in such a way: *giving information on the Web well-defined meaning, better enabling computers and people to work in cooperation.*[1]

In this PhD thesis, we analyse different methods to (i) extract rich knowledge detailed in different resources of archival NHCs and (ii) publish the result on the Semantic Web as machine-readable knowledge for others to take up, reuse, and integrate with their own collection data.

[1] A paraphrased fragment from an article published in May 2001 in the Scientific American, titled "The Semantic Web": "The Semantic Web is not a separate Web but an extension of the current one, in which information is given well-defined meaning, better enabling computers and people to work in cooperation."

SUMMARY

First, we analyse system designs for the extraction of information from species observation records in field notes and scientific illustrations. For the extraction of information and knowledge from field notes, we argue for an approach that favors quality over quantity, rich semantic annotation over full-text transcription. Field notes are challenging to work with due to a variety of factors, such as the evolving visual style of a single alphabet and historical hard-to-read handwriting. Semantic annotation intrinsically motivates domain experts to produce high-quality data, and formalises minimal information required for sufficient digitisation. Following from the approach, we analyse what type of semantic information domain experts use to query collections, and what metadata they would use to integrate their collection data. Subsequently, we analyse how knowledge should be stored in a Findable, Accessible, Interoperable, and Reusable (FAIR) way, to encourage further scientific discourse and discovery. As a result of the process described above, we describe a web application for the semantic annotation of natural history archival collections, the Semantic Field Book Annotator (SFB-Annotator). The semantic annotation application, with its underlying semantic model, is being developed further within the LInking Notes of NAturE (LINNAE)–project.

Additionally, we propose a method for the automation of the semantic annotation process. The manual extraction of knowledge from archives is a time-consuming and labour-intensive process. We show that we can identify and classify scientific names in handwritten field notes, using strong assumptions based on expert's knowledge about the structure and content of observation records.

Finally, we analyse the extraction of knowledge from scientific illustrations. Automated species identification is challenging in general due to the inherently long-tailed nature of data, and the millions of classes in a species taxonomy, making it challenging to create models that can identify common as well as rare species. We propose to tackle the problem with zero-shot learning. Although open issues remain—e.g., distribution shifts between illustration collections, originating from differences in paper types, illustration style and granularity of depicted objects—zero-shot learning facilitates learning from prior information, which we believe to be crucial for automated information extraction from heterogeneous data.