Knowledge extraction from archives of natural history collections

Stork, L.

CHAPTER 7

Conclusions

"Biodiversity starts in the distant past and it points toward the future."

– Frans Lanting

In this thesis, we presented methods for knowledge extraction from archives of NHCs, informed by prior knowledge of the domain. Archives serve as important historical records, and also crucial references for present-day research subjects, such as environmental studies and climate change. The current biodiversity crisis increases the importance of historical studies, as a longer-term view of changes to biodiversity may provide new insights. New approaches to knowledge extraction from archival collections related to NHCs are required to deal with hard-to-read handwriting, heterogeneous content and the change of species names, genera and place names. Links need to be identified between related items within a specific collection, as well as with external historical resources, such as the Biodiversity Heritage Library (BHL), and contemporary resources, such as the Global Biodiversity Data Facility (GBIF), the Catalogue of Life (COL), or iNaturalist, in order to discover new knowledge.

First of all, we provided motivation for a "more product, less process" approach (87), in which we leverage context, in the form of domain expert knowledge and community-developed data standards, for the semantic annotation of digitised manuscripts. We implemented this approach through the development of a semantic model, the NHC-Ontology, and a semantic annotation tool, the SFB-Annotator, which we evaluated on a use case from the domain. Second, we used the output of the semantic annotation process to train a classifier for the detection of scientific names in text images, in which context, in the form of prior knowledge about the syntax and semantics of nomenclature, as well as about field books, drove the learning process. Last, we explored how distributed, multimodal contextual knowledge from data providers within the domain, such as GBIF,

iNaturalist, and BHL, could be used to extract knowledge (hierarchically structured classifications) from biological illustrations.

In the following section (Section 7.1) we will revisit the research questions we introduced in the introductory chapter (Chapter 1), followed by a discussion of the overall approach and its implications, against the backdrop of developments in the fields of Semantic Web and computer vision, and AI in general.

To conclude, we discuss ongoing and future developments (Section 7.2) related to the work in this thesis.

7.1 Research Questions Revisited

The main objectives of this thesis were: to (i) *extract knowledge from archives of NHCs*, given items **Chall.1** to **Chall.8**, to *make them amenable for research*, and (ii) to *publish the digitised archives and the extracted (meta)data online for global access and integration with other collections (related to **Chall.5**)*. In the introduction, we split these objectives into four research questions that guided our work. We will revisit them below.

Q.1 *What are the trade-offs of various system designs for the disclosure of digital archives?* (Chapter 3)

To systematically answer **Q.1**, the first research chapter of this thesis (Chapter 3) discussed three types of systems that transform manuscripts to machine-readable databases. We focussed specifically on the agents that were expected to perform the enrichment (the crowd, experts, or machines), the type of machine-readable data that was being produced (a lexicon, an annotated corpus, or a knowledge graph), and how much of the manuscripts were processed (manual or machine full-text transcription, user-guided labelling of keywords with a focus on searchability, or enrichment targeted to central units such as named entities).

From these discussions, and given **Chall.1** to **Chall.8**, we derived a "more product, less process" approach for knowledge extraction from field books. Instead of full-text or user-guided keyword transcription, we opted for a targeted approach that depends on domain experts for (i) steering the development of a formal application ontology for field observation records, and (ii) using it for the semantic annotation of these observation records.

Omitting full-text transcription means annotating only a small percentage of the hard-to-read multilingual content, and the transcription and annotation process is streamlined: both the verbatim reading of a text as well as the interpretation can be recorded. We do

note that modelling of manuscript content becomes increasingly complex when content is too unstructured to fit an ontology. However, the content of field books and illustrations fit well into an ontology, as these are characterised by their systematic nature. Moreover, we note that semantic annotation is a knowledge-intensive task that depends on an expert community. Nevertheless, we envision that domain experts have higher intrinsic motivation to take on a task that is relatively difficult, and that relates to their field of interest. Additionally, such tasks tap into a feeling of community contribution. Lastly, we note that automating semantic annotation from text images is likely to be a more complex task than from digital texts, as the structural and positional features of digital texts are much more homogeneous than that of text images.

The research questions that followed, were targeted to the kind of knowledge that needed to be extracted, how formal ontologies could be employed to do so, and whether resulting knowledge graphs could be used to answer domain expert's research questions:

Q.2 *What types of research questions do domain experts formulate regarding archives of NHCs, and how can we make the content of these archives machine-readable to facilitate such queries?* (Chapter 4)

- **Q.2a** *What are the general semantics of historical species observations and how do they differ from present day observations?*
- **Q.2b** *How do we extract important content and its semantics (e.g., core elements and their relationships) from the archives so that it becomes machine-readable and facilitates rich queries?*

First, qualitative interviews and a test annotation procedure were set up to answer research question **Q.2**a. Experts were asked to note down research questions and concepts that were related to the content of field books and illustrations, and subsequently to annotate the digital manuscript pages with these (or new) self-defined semantic concepts.

To answer **Q.2**b, technologies from the field of knowledge representation and reasoning (KRR) were used for the transformation of manuscripts to machine-readable knowledge in the form of knowledge graphs. The concepts defined by domain experts were used for the development of an ontology that represents the content of historical species observations. Through the development of a semantic annotation tool based on the application ontology, domain experts can elucidate the important named entities and their relations, and make them available through a queriable triple store. Qualitative evaluations demonstrated that the tool is usable by domain experts, both for the process of creating structured annotations, as well as answering common research questions. We do note that a larger

"crowd" is required to evaluate the tool and model quantitatively, for instance by measuring inter-annotator agreement (IAA).

Importantly, annotations are produced and published in a FAIR way that stimulates reuse of data and repetition of scholarly experiments. This relates to our third research question:

Q.3 *How can we accommodate a transparent and FAIR approach to enriching the archival content of NHCs, facilitating and encouraging scientific discourse over the content?* (Chapter 4)

Requirements (**R.3** and **R.4**) were set up for publishing the content of manuscripts from NHCs to the Semantic Web as FAIR data. Classes and relations from well-established domain ontologies and vocabularies were selected to represent expert user-defined concepts, in line with the FAIR data principles and the vision of the Semantic Web (which encourages knowledge sharing and reuse). We argued that provenance of annotation is often overlooked, albeit being a very important step in the life of any digital object or statement, as it contributes to meaning, value and reproducibility of experiments. To track the provenance of semantic annotations, we used the Web Annotation Vocabulary[??] and accompanying data model.[1] By tracing and publishing the provenance of annotations on the Semantic Web as LOD, important links, such as those from a *taxonomic referencing* process (the annotation of a legacy name with a reference to an accepted name in a present-day biological taxonomy) become accessible by any researcher, and can be fruitfully discussed. We should stress that an infrastructure for publishing and discussion of such statements in a FAIR way is not yet available in the SFB-Annotator, but this will be taken up in future developments.

Lastly, extracting information from heterogeneous, historical material is time-consuming and requires domain expertise. Through **Q.4**, we investigated how we could exploit context-driven automated methods to help domain experts with the extraction of knowledge from field books and illustrations.

Q.4 *How can we use automated methods for knowledge extraction from archives of NHCs?* (Chapter 5 and 6)

First, we developed a deep-learned model for the recognition and classification of scientific names in field books. The model was based on structural (visual) and positional features (salient named entity recognition and classification (SNERC), a term we use to define a type of NERC in which entities that are visually *salient* in text *images* are recognised

[1] https://www.w3.org/TR/annotation-model/

and classified). Our methods show applicability even though the dataset contained four authors with different handwriting styles and different processes of recording their species observations. We do realise that our experiments were based on limited data, as the semantic annotation tool was not yet available for use by a small crowd of experts, which limited the number of available domain experts that could be deployed for annotation. Moreover, the experiments serve as a proof of concept: only a small percentage of the classes were used for automated semantic annotation, and named entities were annotated semantically, so far without transcription.

Second, we explored methods for the classification of biological illustrations. Historical names that accompany historical biological illustrations are often unpublished or obsolete within biological taxonomies that exist today. To aid the domain experts in the identification of their biological illustrations as taxa from an established taxonomy (such as the GBIF taxonomy backbone), we explored ZSL methods based on multimodal background knowledge from multiple data providers within the domain, namely GBIF, iNaturalist and BHL. Although results demonstrated the complexity of the task, we believe that automated methods that map biological illustrations to scientific names within a contemporary taxonomy can act as decision support for the identification of rich historical illustrations.

To conclude, we argue that the results discussed in our experimental chapters are encouraging. Methods driven by prior knowledge can build on the legacy of expert domain knowledge, such as domain ontologies or models trained for ZSL, which are better suited to deal with ambiguous content and limited data, and indicate potential for use of such models in an expert support system for semantic annotation of field books and illustrations. At the same time, the results stress the difficulty of our task, and specifically show a necessity for research into methods that are able to learn from small samples and heterogeneous content, especially for a field in which semantic modelling or generation of training data heavily depend on domain expert's involvement.

Archives of NHCs are crucial sources for research in a wide range of other subjects such as environmental and climate change. The technologies proposed in this thesis aim at building a technological infrastructure that will allow users to semi-automatically extract knowledge from historical manuscript collections, and to present the extracted knowledge in a FAIR way to researchers and the public at large. Using Semantic Web technologies for the transformation of manuscripts to knowledge graphs allows users to construct rich semantic queries or aggregate informative content across archival collections. Automated methods such as HTR, NERC and ZSL can users to semi-automatically extract and organise the content. It thus opens up new opportunities for scientific research, heritage institutions and

publishers, while reducing the need for costly human intervention. Moreover, reconciling historical and contemporary biodiversity data opens op possibilities for mapping out long-term changes in biodiversity.

7.2 Ongoing and Future Developments

Currently, we are working on the implementation of an online version of the SFB-Annotator, as more extensively discussed in Section 4.6. When published online, a small user-base of experts can be deployed for annotation, which will, in turn, extend the annotation knowledge graph. With access to a larger annotation knowledge graph, learning algorithms can be deployed to infer new knowledge. We envision using learning over graphs to predict links between multimodal resources (details discussed in Subsection 2.1.2) (*entity linking*), or for instance between named entities that refer to the same entity (*named entity disambiguation*).

Furthermore, we aim to further our SNERC implementation to include the transcription of named entities, using techniques from HTR (preferably with ZSL for the recognition of unseen *out-of-vocabulary* words) (165; 166).

Moreover, we aim to publish valuable statements about the content of field books and illustrations—e.g., resolved ambiguous taxonomic names or locations—online as FAIR data, thereby stimulating scholarly discussions over the content, and envision publishing such statements as micro-contributions on the *NanoBench*[1] for nano-publications.

The methodologies presented in this thesis have implemented what we call a "serving hatch" approach to the combination of techniques from subsymbolic and symbolic AI. What we mean by the analogy is that techniques from both fields are deployed to fruitfully pass information back and forth. In our case, an application ontology informs a classifier to look for instances of certain classes, and how these should be related, and the classifier learns from experience where these are. The output of the classifier therefore allows for some form of interpretation and reasoning. We argue that this is a first step in the creation of an infrastructure that facilitates *hybrid* AI—in which techniques from both families work together through combined inference and reasoning. In future work, we would like to research hybrid techniques for knowledge extraction from archives of NHCs. Such techniques could improve and accelerate learning from small samples and heterogeneous data through the exploitation of the strengths of both fields. For instance, we envision reasoning-based handwriting recognition and semantic annotation, in which inference is

[1] `https://github.com/peta-pico/nanobench`

performed through a dialogue between both bottom-up induced (learned), and top-down deduced (reasoned) facts.