2018 IEEE 14th International Conference on e-Science

Linking Natural History Collections

Lise Stork*
Leiden Institute of Advanced Computer Science
Leiden University
Leiden, the Netherlands
l.stork@liacs.leidenuniv.nl

Andreas Weber*
Department of Science, Technology, and Policy Studies
University of Twente
Enschede, the Netherlands
a.weber@utwente.nl

Eulàlia Gassó Miracle
Naturalis Biodiversity Center
Leiden, the Netherlands
eulalia.gassomiracle@naturalis.nl

Katherine Wolstencroft
Leiden Institute of Advanced Computer Science
Leiden University
Leiden, the Netherlands
k.j.wolstencroft@liacs.leidenuniv.nl

I. Introduction

Over the last two decades, natural history museums, archives, and libraries all over the world have spent major efforts to digitise natural historical collections [1]. Usually, such collections consist of handwritten field notes, drawings, published descriptions, and actual specimens. Therefore, disclosing and linking the content of such digitized collections is challenging. Next to difficulties in deciphering historical handwriting in different languages, researchers struggle with the evolution of concepts (fig. 1). In particular, the change of species names, genera and place names makes it difficult to identify links between related items within a specific collection, and also with external historical resources, such as the Biodiversity Heritage Library (BHL), and contemporary resources, such as, the Global Biodiversity Information Facility (GBIF). Up to now, manual linking methods, in combination with entity recognition, have been applied to such content, *after* full text transcription [2] [3] [4]. Although such a procedure often leads to high-quality data, it is also a labour-intensive, time-consuming and therefore a costly way of opening up natural history collections [5]. In this paper, we describe tooling and infrastructure that enables direct and collaborative semantic annotation of field book content, without the requirement for full transcription. These tools enable a more streamlined approach to the creation of rich, integrated archives that can be interlinked with other cultural history resources in the field. In our use case we are annotating and enriching data from expeditions undertaken by the Committee for Natural History (1820-1850) [6]. The collection contains approximately 10,000 specimens and 8000 handwritten field book pages. Due its vast size and heterogeneity, a full text transcription is no viable option. Despite digitization, the collection has thus remained inaccessible to scholars and the general public.

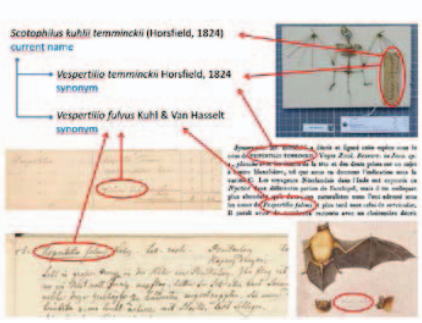

Fig. 1. Evolution of a taxonomic name of a bat.

II. The Semantic Fieldbook Annotator

At the core of our paper is the Semantic Fieldbook Annotator (=SFB-A) which enables researchers, collection holders and possibly also citizen scientist to interact with digitized natural history collections [6]. The SFB-A allows users to draw bounding boxes, or Regions of Interest (ROIs), over the image scans to which semantic annotations, e.g. semantic classes of words, can be attached. The SFB-A forms part of a broader project which examines the options and limitations of using semantic technologies in the domain of digitised biodiversity heritage [7] [8]. The SFB-A allows for direct annotation of named entities in digitized images of natural historical and similar collections.

Instead of transcribing all text, we use the SFB-A to annotate salient named entities in field notes of which the semantics are defined in a formal ontology. Examples are taxonomical names or geographical locations. By doing so we aim to save time and cost, and also preserve a direct link to the original document image. The SFB-A is pre-populated with information on naturalists who participated in the expedition and places visited during their work. The SFB-A interface guides users in their choices of semantic classes and instances, using autocomplete for terms in the underlying ontology. The ontology, which has been fully evaluated elsewhere, describes the concepts expected in any field book record and the relationships between them [6] [7]. It is an application ontology that integrates several existing domain ontologies, for example, the Darwin Core Semantic Web and the Uberon anatomy ontology. The SFB-A also allows researchers to examine the original text and annotations to determine if they agree with the interpretations. Potentially, this opens up natural history collections for meaningful collaborations among experts and citizens worldwide [9]. It also enables scholarly discourse over specific parts of the source material. The resulting semantic annotations can be stored and served as linked data, in order to interlink content with other resources. Taken together, this paper shows that direct semantic annotation is able to produce an integrated resource that can be related to present-day biodiversity results, from resources such as GBIF, and other historical collections, such as those from the BHL. In addition, when annotated named entities are interlinked by a semantic model of the domain, this method can disclose an integrated, searchable and FAIR (Findable, Accessible, Interoperable and Reusable) dataset. In our presentation we will demonstrate the working of the SFB-A and discuss the challenges we encountered.

* Lise Stork and Andreas Weber share the first-authorship of this abstract. Funding for this research is provided by NWO and Brill publishers.

978-1-5386-9156-4/18/$31.00 ©2018 IEEE
DOI 10.1109/eScience.2018.00113

REFERENCES

[1] M. Heerlien, J. van Leusen, S. Schnörr, S. De Jong-Kole, N. Raes, and K. van Hulsen, "The natural history production line: an industrial approach to the digitization of scientific collections," *J. Comput. Cult. Herit,* vol. 8, pp. 3-11, 2015.

[2] M. van Erp, "Acessing natural history: Discoveries in data cleaning, structuring, and retrieval," PhD thesis Tilburg University, 2010.

[3] H. Pethers, and H. Huertas, "The Dollmann collection: a case study of linking library and historical specimen collections at the Natural History Museum, London," *Linnean*, vol. 31, pp. 18-22, 2015.

[4] J. Wettlaufer, Ch. Johnson, M. Scholz, M. Fichtner, and S.G. Thoteompudi, „Semantic Blumenbach: Exploration of text-object relationships with semantic web technology in the history of science," *Digital Scholarship in the Humanities*, vol. 30, suppl. 1, pp. 187-198, 2015.

[5] T. Causer, K. Grint, A.-M. Sichani, and M. Terras, "'Making such a bargain': Transcribe Bentham and the quality of cost-effectiveness of crowdsourced transcription," *Digital Scholarship in the Humanities* (2018), pp. 1-21.

[6] https://github.com/lisestork/SFB-Annotator

[7] L. Stork, A. Weber, E. Gassó Miracle, F. Verbeek, A. Plaat, J. van den Herik, and K. Wolstencroft, "Semantic annotation of natural history collections," *J. of Web Semantics*, accepted for publication, in press.

[8] A. Weber, M. Ameryan, K. Wolstencroft, L. Stork, M. Heerlien, and L. Schomaker, „Towards a Digital Infrastructure for Illustrated Handwritten Archives," in *Digital Heritage*, M. Ioannides (ed.), Lecture Notes in Computer Science (LNCS), vol. 10605, Cham: Springer Nature, 2018, 155-166.

[9] J.A. Drew, C.S. Moreau, and M.K.J. Stiassny, "Digitization of museum collections holds the potential to enhance researcher diversity," *Nature Ecology & Evolution* vol. 1, 1788-1790, 2017.