Knowledge extraction from archives of natural history collections

Stork, L.

Knowledge Extraction from Archives of Natural History Collections

by Lise Stork

Knowledge Extraction from Archives of Natural History Collections

Proefschrift

ter verkrijging van
de graad van doctor aan de Universiteit Leiden,
op gezag van rector magnificus prof.dr.ir. H. Bijl,
volgens besluit van het college voor promoties
te verdedigen op donderdag 1 juli 2021
klokke 13.45 uur

door

Lise Stork

geboren te Eindhoven, Nederland
in 1990

Promotiecommissie

Promotores:	Prof. Dr. A. Plaat	
	Prof. Dr. Ir. F.J. Verbeek	
Co-promotor:	Dr. K.J. Wolstencroft	
Overige leden:	Prof. Dr. M.M. Terras	University of Edinburgh
	Prof. Dr. T.R. van Andel	
	Prof. Dr. H.C.M. Kleijn	
	Prof. Dr. M.S.K. Lew	
	Dr. V. de Boer	Vrije Universiteit Amsterdam
	Dr. A. Weber	University of Twente

Cover illustrations by G. van Raalten and P. van Oort[1]
Cover design by Dirk Meijer, Marius van Leeuwen and Lise Stork
Layout design by Hao Wang and Lise Stork
Printed by: Gildeprint, Enschede, the Netherlands

This work is supported by the Netherlands Organisation for Scientific Research (NWO) and Brill publishers, grant number 652.001.001 (the Making Sense of Illustrated Handwritten Archives Project). The project was a collaboration between the Leiden Centre of Data Science (LCDS), Naturalis Biodiversity Center (NBC), the university of Leiden (LIACS), Twente (STePS), Groningen (ALICE), and Brill publishers as creative industry partner.

[1]Collection Naturalis Biodiversity Center, MMNAT01_AF_NNM001000192 (front) MMNAT01_AF_NNM001000236 (back). Images free of known restrictions under copyright law (Public Domain Mark 1.0).

Abstract

Natural history collections (NHCs) provide invaluable sources for researchers with different disciplinary backgrounds, aspiring to study the geographical distribution of flora and fauna across the globe as well as other evolutionary processes. They are of paramount importance for mapping out long-term changes: from culture, to ecology, to how natural history is practiced.

This thesis describes computational methods for *knowledge extraction* from archives related to NHCs—here referring to handwritten manuscripts and hand-drawn illustrations. As we are dealing with heterogeneous real-world data, the task becomes exceptionally challenging. Small samples and a long-tailed distribution, sometimes with very fine-grained distinctions between classes, hamper model learning. Prior knowledge is therefore needed to bootstrap the learning process. Moreover, archival content, such as scientific names and their authors, can be difficult to interpret and integrate. Archival content should therefore be formally described for data integration within and across collections. By serving extracted knowledge to the *Semantic Web*, collections are made amenable for research and integration with other biodiversity resources on the Web.

We demonstrate how to leverage domain expert involvement and prior knowledge, such as the natural world's systematic organisation, in the development of state-of-the-art methods from the fields of *computer vision* and the *Semantic Web* for the task of *knowledge extraction* from natural history archival collections.

Keywords— Natural history, Biodiversity, Semantic Web, Knowledge extraction, Prior knowledge, Computer vision

Contents