



Universiteit
Leiden
The Netherlands

Knowledge extraction from archives of natural history collections

Stork, L.

Citation

Stork, L. (2021, July 1). *Knowledge extraction from archives of natural history collections*. Retrieved from <https://hdl.handle.net/1887/3192382>

Version: Publisher's Version

License: [Licence agreement concerning inclusion of doctoral thesis in the Institutional Repository of the University of Leiden](#)

Downloaded from: <https://hdl.handle.net/1887/3192382>

Note: To cite this publication please use the final published version (if applicable).

Cover Page



Universiteit Leiden



The handle <http://hdl.handle.net/1887/3192382> holds various files of this Leiden University dissertation.

Author: Stork, L.

Title: Knowledge extraction from archives of natural history collections

Issue date: 2021-07-01

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

Copyright © 2021 Lise Stork.

Cover illustrations by G. van Raalten and P. van Oort¹

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.

¹Collection Naturalis Biodiversity Center, MMNAT01_AF_NNM001000192 (front) MM-NAT01_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

Abstract	i
1 Introduction	1
1.1 Access to Archives	1
1.1.1 Natural History Archival Collections	1
1.1.2 Collection Digitisation	3
1.1.3 From Archives to Databases	4
1.2 Research Questions	9
1.3 Approach and Main Contributions	10
1.4 Project Context	12
1.5 Outline of Chapters	13
2 Background	15
2.1 Natural History	15
2.1.1 Taxonomy and Nomenclature	16
2.1.2 Multimodal Field Observations	18
2.2 Knowledge Extraction	21
2.2.1 Machine Learning	22
2.2.2 Knowledge Representation and Reasoning	30
2.3 Data Sources and Use-Cases	34
2.3.1 Data Sources	35
2.3.2 Use Cases	35
3 Manuscripts to Databases	37
3.1 Introduction	37
3.2 System Designs	39
3.2.1 Manual Full-Text Transcription.	40
3.2.2 Semi-Automated Transcription	43
3.2.3 Semantic Annotation of Text Images.	45

3.3	More Product, Less Process	46
4	Semantic Annotation	49
4.1	Introduction	49
4.2	Development of a Semantic Model	51
4.2.1	Requirements	51
4.2.2	Semantics for Biodiversity	52
4.2.3	Data Elucidation by Domain Experts	54
4.2.4	The NHC-Ontology	56
4.3	Semantic Annotation	61
4.3.1	System Design	62
4.3.2	The Semantic Field Book Annotator	63
4.3.3	Towards Semi-Automated Annotation	65
4.4	Qualitative Evaluation	66
4.4.1	The Annotation Process	66
4.4.2	The Data	67
4.4.3	Semantic Queries	68
4.5	Conclusions	72
4.6	Ongoing and Future Work	72
5	Automating Semantic Annotation	77
5.1	Introduction	77
5.2	Related Work	78
5.3	Data	79
5.4	Scientific Name Extraction Model	80
5.4.1	Classification of Word Images	81
5.4.2	Semantic Annotation of Word Images	83
5.5	Experiments and Results	84
5.5.1	Experimental Methodology	84
5.5.2	Results and Discussion	85
5.6	Conclusions and Future Work	87
6	Classification of Biological Illustrations	89
6.1	Introduction	89
6.2	Related Work	93
6.3	The Data	95
6.3.1	The ZICE Dataset	95
6.3.2	The Verification-Set	98
6.4	Methodology	98

6.4.1	Zero-Shot Learning Model	98
6.4.2	Image Embeddings	99
6.4.3	Class Embeddings	100
6.4.4	Combining Class Embeddings	101
6.4.5	Hierarchical Prototype Loss	103
6.5	Experimental Setting	103
6.5.1	Dataset Splits	103
6.5.2	Data Augmentation	105
6.5.3	Evaluation Criteria	105
6.6	Experimental Results	106
6.6.1	Supervised Classification and Visualisation	107
6.6.2	Fine-Grained Zero-Shot Learning	108
6.7	Analysis and Discussion	114
6.8	Conclusions	116
7	Conclusions	117
7.1	Research Questions Revisited	118
7.2	Ongoing and Future Developments	122
	Bibliography	125
	Summary	147
	Samenvatting	149
	Curriculum Vitae	153
	List of Publications	155
	Acknowledgements	157