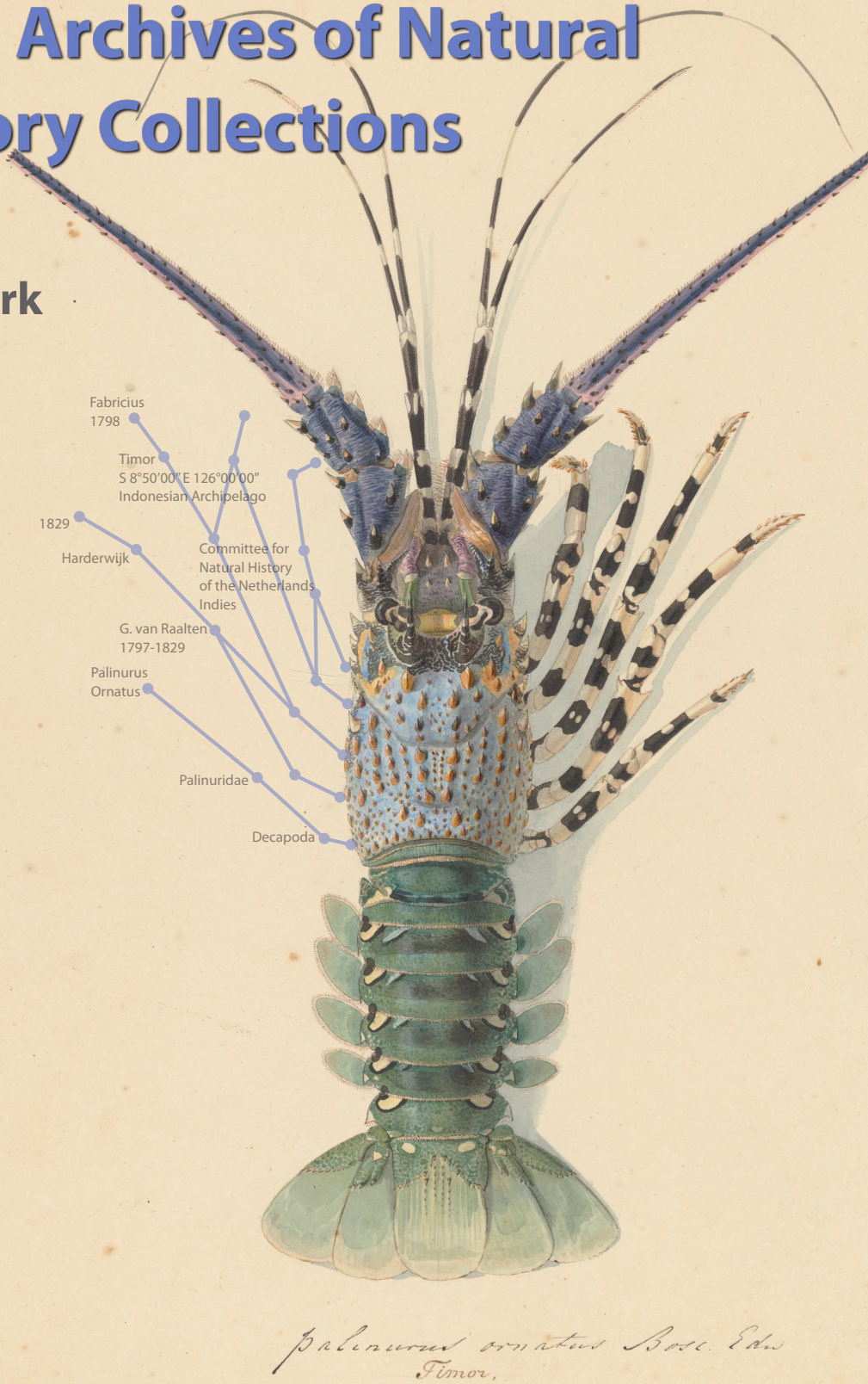
Knowledge extraction from archives of natural history collections

Stork, L.

Knowledge Extraction from Archives of Natural History Collections

Lise Stork



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

CHAPTER 1

Introduction

"Wandering the warren of collection facilities and scientific laboratories that the public rarely sees at the National Museum of Natural History is like peeking into a reconstruction of Noah's Ark. Filling every drawer, cabinet and shelf in sight are millions of taxidermic birds and mammals, preserved worms and fishes, skeletons and fossils, and so much more."

– Maya Wei-Haas, *Smithsonian Magazine*

In the following chapter, we outline the contents of this thesis. We start by describing the main motivation of our work (Section 1.1). We then present the research questions (Section 1.2), and our approach that deals with these questions (Section 1.3). Eventually, we describe the context in which the work was carried out (Section 1.4).

1.1 Access to Archives

1.1.1 Natural History Archival Collections

In 1821, a German naturalist observed a specimen of a bat noting curiously in his field book that "die Zunge ist erstaundend lang, 0,1,3, schmal und pfriemförmig." ("the tongue is surprisingly long, 0,1,3 thin, and awl-shaped.")[1] Based on his observations, he consulted the existing literature, and classified the specimen as the species *Pteropus minimus Geoff*. The naturalist had been part of the Committee for Natural History of the Netherlands Indies ("Natuurkundige Commissie voor Nederlands-Indië") (NC), sent to the Indonesian Archipelago to study natural resources and discover and describe the various species that had their home in the island's rich ecosystem.

Naturalists have been recording field observations—such as the one described above—for several hundred years (1; 2) during expeditions to remote parts of the world, often at

[1] MMNAT01_AF_NNM001001033_004.jpg, a field note taken from the collection of the Committee for Natural History of the Netherlands Indies ("Natuurkundige Commissie voor Nederlands-Indië") (NC), physically kept at the Naturalis Biodiversity Center (NBC).

the risk of their own lives. Field observation records—e.g., field notes and hand-drawn illustrations describing observations of species—contain a wealth of information about scientific practices, important events, and the behaviour, habitat and appearance of organisms and collected specimens. Field observation records are therefore invaluable for increasing the scientific worth of such specimens, which are often accompanied by little descriptive data, see Figure 1.1. They provide detailed accounts of the habitat and behaviour of organisms, colourful histories that specimens alone do not tell. We describe the content of these various resources further in Subsection 2.1.2.

Figure 1.1: A specimen of the *Loligo vulgaris Lamarck, 1798* species from the *Naturalis–Zoology and Geology* catalogues.[1] Images free of known restrictions under copyright law (Public Domain Mark 1.0).

Records from early expeditions are currently held in archives of collection facilities spread out across the globe, such as in natural history museums, botanical gardens and scientific laboratories. One such facility is the impressive 200-feet high tower of the Naturalis Biodiversity Center (NBC) in Leiden, which stores 42 million natural heritage objects, amongst which are specimens—fluid-preserved whole organisms or parts, frozen tissues, pinned insects, pressings, skins, skeletons, nests and other items (3)—as well as archives. You will find a stuffed rhinoceros, drawers with pinned butterflies, a jar with the face of a gorilla, the skeleton of a killer whale,[2] a beautifully detailed illustration of a jellyfish (see back of this thesis), and a handwritten description of the behaviour and dissection of a fruit bat. For colourful accounts of such collections and their history, we refer to the following books: (4) and (5).

An example collection at the NBC is the collection of the Committee for Natural History of the Netherlands Indies ("Natuurkundige Commissie voor Nederlands-Indië") (NC), which consists of roughly 8,000 field notes and 2,000 illustrations, and is related to approximately 10,000 specimens. The specimens and archival materials in this collection provide a unique

[1] https://bioportal.naturalis.nl/multimedia/RMNH.MOL.5009890_0
[2] https://bioportal.naturalis.nl/multimedia/RMNH.MAM.2559_0

view on the distribution of plants and animals in the Indonesian archipelago in the first half of the nineteenth century.

Researchers interested in natural history and biodiversity endeavour to find their way through these vast, labyrinthine collections and archives. In doing so, they are challenged by the complexity of the material. The archival handwritten and illustrated content is especially difficult to process. Even those with expertise in paleography, the study of ancient and historical handwriting, consider natural history archives to be difficult resources. They require researchers to understand and deal with different collection and documentation practices which have evolved over time, and the different European, indigenous and scientific—e.g., Latin and Greek—languages, used to describe the collected flora and fauna.

1.1.2 Collection Digitisation

There are roughly 2.5–3 billion specimens housed in collections worldwide (6; 7; 3), and many of them are accompanied by archival materials and illustrations. Industrial-scale digitisation projects have been set up to produce high-resolution digital renditions of physical collection objects (8; 9; 10; 7), such as the one displayed in Figure 1.1. Over recent years, the World Wide Web has become an important hub for natural history museums,[1] to publish their digitised material (11; 12; 13). Where earlier, artifacts such as specimens, field notes and illustrations, were only accessible to researchers or experts with access to research institute or museum facilities, the Web allows for widespread access to digital high resolution images of collection objects. NHCs have therefore become more widely available, even to the layperson, from any location, at any given time.

Publishing NHCs to the Web introduces new opportunities for data reuse and integration. Museums and other institutions use collection metadata[2] (descriptive data about data) to support accessibility and integration of collection objects. We take the definition of a collection from the Collection Descriptions (CD) interest group (14):

Definition 1.1. *A collection can be seen as any group of items that share some common characteristics such that they are useful to describe as a group.*

Defining these common characteristics computationally—e.g., their implicit semantics—makes collections searchable and allows for the inference of new knowledge. For instance, through indexing of collection metadata in searchable resources, web services, such as search engines, can aid users by focusing search, for instance to collections with a specific

[1] Part of the collection from the NBC can be found online: `https://bioportal.naturalis.nl/`

[2] Metadata are machine understandable information about web resources or other things, `https://www.w3.org/DesignIssues/Metadata`

topic, such as "fishes and birds". Often, archival collection metadata are described with a controlled vocabulary—a fixed set of terms—or ontology, so as to virtually reunite distributed collections (14). Further details are described in Subsection 2.2.2.

For archival resources specifically, we find it useful to distinguish four levels to which metadata can be assigned, see Table 1.1, organised from the coarsest level to the most fine-grained level.

Table 1.1: Levels of archival collection metadata

Metadata levels	Examples	Example metadata
1. collection-level	A collection of objects collected during an expedition	*collection name, associated persons, collection type, temporal coverage, geographical coverage, . . .*
2. item-level	A field book, a diary	*type, title, author, subject, . . .*
3. page-level	A field note, a page from a book	*type (e.g., field note), part of, contents . . .*
4. content-level	A region in a page. A word (e.g., a named entity), a table, a depiction	*type (e.g., a location), language, provenance (e.g., annotator, annotation date) . . .*

In this thesis, the focus lies on metadata coupled to the *content*-level of archives. More specifically, on the transformation from archives—specifically digital renditions of handwritten and illustrated observation records—to machine-readable data.

1.1.3 From Archives to Databases

The growing role of archives in creating a global online knowledge base of biodiversity data, creates new challenges and opportunities for the creation of workflows and best practices for their digitisation. It is often unclear what the term 'digitisation' of archives precisely entails, so we define it here as consisting of four stages: (1) scanning, (2) transcription, (3) representation of knowledge and (4) integration with other historical as well as contemporary digital sources. Below, we first describe general challenges related the process of reading and interpreting archival content. Thereafter, we detail some of the challenges related to stage (2), after which we briefly detail opportunities related to stage (3), using Semantic Web technologies and methods from the field of computer vision, which lead us to our approach and main contributions (Section 1.3).

Challenges Elucidating content of handwritten and illustrated observation records is a complex and intricate task that largely depends on domain expertise. Early records were often written in hard-to-read handwriting as well as in multiple languages (15), and the

evolution of species names, concepts and place names, makes interpretation of the content challenging.

Another challenge has to do with the distribution of collection objects over various institutes and collections. It was, for instance, common for naturalist to trade resources or send them elsewhere for observation or publication (16). Moreover, in the course of the nineteenth and twentieth centuries, natural history museums separated specimens from field notes as well as illustrations. The physical distribution of collection objects across institutes hampers the use of historical observation records for specimen studies. Written references to specimens, literature, illustrations and other field notes are lost and difficult to retrace. Hence, re-establishing links between specimens and archives allows for better integration of their content. These challenges are discussed in more detail in Chapter 2.

To summarise, Table 1.2 shows challenges researchers face when aiming to read and interpret archival content related to NHCs.

Chall.1 Hard-to-read historical handwriting
Chall.2 Evolving scientific paradigms
- (a) The evolution of concepts
- (b) Changing scientific practices
- (c) The evolving (visual) style of a single alphabet

Chall.3 Multilingualism: the use of multiple languages within collections, often even within one page.
Chall.4 Term ambiguity
- (a) Homonymy, polysemy and synonymy
- (b) Abbreviations

Chall.5 Physical distribution of collection objects related to NHCs

Table 1.2: Challenges that come with reading and interpreting the content of archives of NHCs

Without the aid of computational processes for search and integration of data, making sense of such complex and heterogeneous collections becomes an intractable problem.

Manual Transcription Manual full-text transcription is often used to transform handwritten text to digital machine-readable text, e.g., (17; 15; 18), as it produces high-quality data and, through search engines such as Apache Lucene,[1] facilitates computational indexing of terms and full-text search. However, manual full-text transcription is a time-consuming labour-intensive process that heavily depends on domain expertise. Moreover, even though it is not a property of manual transcription, we note that transcriptions

[1] https://lucene.apache.org/

often exist in unstructured or semi-structured text files. Unstructured transcriptions do not stimulate scholarly discussions over challenging or ambiguous content, related to the interpretation-related challenges mentioned above, nor do they facilitate the use of automated methods such as computer vision for Handwritten Text Recognition (HTR), when transcribed words are decoupled from their digital representations (no *ground truth*[1] is created).

Automated transcription Automated methods such as Optical Character Recognition (OCR) offer another solution to the problem, promising to unburden domain experts by taking over part of the transcription process. Even though OCR is seen by many as a solved task, it only allows the processing of homogeneous manuscripts, homogeneous in terms of layout, writing style and lexicon (19; 20; 21; 22; 23). OCR systems rely on the identification of single *characters*, and knowledge about how these are configured to form words and sentences. Therefore, OCR systems are required to know the script of a text, as well as the language it expresses (23). HTR from heterogeneous content—where writings can be multilingual, follow curved lines, are interspersed with depictions and tables, and contain inter-word connections—is still a highly complex task. Compared with manual full-text transcription, HTR systems gain a decrease in transcription time, but sacrifice data quality, as error rates for historical documents with large vocabularies are often high (24; 25). Consequently, Schomaker (2016) (23) argues that the target objective needs to be adjusted: rather than taking a character approach, and positioning HTR as a function for full-text transcription, it should function in word retrieval and search systems.

When taking a word-oriented approach to HTR, out-of-vocabulary (OOV) or zero-shot learning (ZSL) strategies are required to deal with the long-tailed distribution of manuscript content (26; 27). Both strategies aim to recognise content not yet observed by a classifier during training, which is crucial for our purpose. Success of machine learning methods for automated detection and classification (discussed in Subsection 2.2.1) commonly depends on large samples for training, ∼thousands of examples per class. Interesting content in observation records (e.g., nouns such as species names, persons, locations, depictions of rare items or species (28)) lie in the long tail of the distribution, meaning that they have a low occurrence rate compared with common content (e.g., articles, prepositions).

To summarise, Table 1.3 shows challenges researchers face when aiming to use automated methods for the transcription of archival content related to NHCs. Moreover, both aforementioned solutions prove time-consuming and labour-intensive, and commonly

[1]Ground truth in computer vision refers to the mapping between an observation of an object in an image (a whole image or a region of an image) to a discreet categorisation that we use in language to refer to the object.

produce flat, unstructured or syntactically structured data that are difficult to understand, integrate and search computationally. Researchers studying the material are expected to search through indexed keywords or using full-text search, requiring them to have considerable prior knowledge concerning the material. For observation records, search is further complicated by use of multiple languages, and writing in which historical terms, name ambiguity, hypernymy and homonymy are common (29). Such challenges ask for more 'intelligent' web technologies.

Chall.6 Heterogeneous material
- (a) Multiple modalities, often on one page
- (b) Multilingualism (same as **Chall.3**)
- (c) Curved lines (challenging for line segmentation)
- (d) Poor paper quality and bleed-through of ink, varying material
- (e) Intra-class variations (e.g., the evolving (visual) style of a single alphabet, multiple viewpoints)

Chall.7 Labelling or annotation of the content requires domain expertise

Chall.8 Long-tailed distribution of data: many classes having only few instances (and those are often the most interesting classes, such as rare species)

Table 1.3: Challenges that come with the automated extraction of information from archives of NHCs

Knowledge extraction Knowledge extraction is the process of producing knowledge from sources, such as unstructured texts or images, by organising their contents according to some formalised semantic data model. *Semantic annotation*, for instance, is the process of annotating *named entities*[1] in an (often digital) text with semantic concepts such as a class to which the entity belongs—e.g., a person—relevant for text interpretation and document retrieval. Coupling named entities with background knowledge, and linking them through formal, ontological descriptions, provides connectivity throughout the documents (31). Recognising and classifying named entities automatically is a sub-task of *knowledge extraction*, called named entity recognition and classification (NERC). NERC techniques work, similarly to semantic annotation, commonly on digital texts rather than on text images. A first step in the semantic annotation or NERC process is therefore often full-text transcription. Techniques related to knowledge extraction are further described in Subsection 2.2.2.

[1] *Named entity* is a term coined during the Sixth Message Understanding Conference (MUC-6) by R. Grishman & Sundheim in 1996 (30). Named entities are the central units of a text; they form the general semantics of a text. Examples of named entities are person, organization and location names, and numeric expressions including time, date, money and percent expressions.

FAIR data and the Semantic Web Due to technological developments and an increase in the amount of data available on the Web, scholars increasingly rely on computational support to deal with complex research data. To improve the infrastructure that supports the accessibility and reuse of scholarly data, research institutes have set up the Findable, Accessible, Interoperable, and Reusable (FAIR) data principles.[1] Publishing scholarly content to the Web as FAIR data leads to high quality digital publications that facilitate data and knowledge integration and reuse, and thereby cooperation and the discovery of knowledge (32).

FAIR data largely depends on the technical way in which metadata are published and curated, and therefore align for a large part with ideas of the Semantic Web, and the Linked Data principles for publishing interlinked datasets to the Semantic Web:[2]

1. Use Uniform Resource Identifiers (URIs) as names for things (*reusable*).
2. Use HTTP URIs so that names can be looked up (*findable and accessible*).
3. Provide useful information when someone looks up a URI, using standards (*interoperable and reusable*).
4. Include links to other URIs, so that someone can discover more things (interoperable).

The use of globally unique and persistent identifiers—such as URIs[3] or Internationalised Resource Identifiers (IRIs)[4]—uniquely define (meta)data, such as named entities and their metadata, which is crucial when dealing with highly ambiguous names. Using HTTP URIs such as Uniform Resource Locators (URLs) make data findable and accessible, and through the use of accessible and shared formalisations (such as RDF and SPARQL, which we discuss in Subsection 2.2.2) for available metadata, data become understandable by machines, and thereby interoperable. Using the principles of Linked Data to structure the handwritten and illustrated content of archives allows scholarly discussions about their interpretation, as well as comparative research across distributed collections, global studies and studies that span generations.

Computational methods—such as the ones mentioned above—can assist researchers interested in biodiversity and documentation practices in the formation of an encompassing knowledge base with respect to the current and past variability of organisms and their distribution across the globe, and thereby with the formation of a global picture of

[1] https://www.force11.org/fairprinciples
[2] https://www.w3.org/DesignIssues/LinkedData.html
[3] https://www.w3.org/TR/uri-clarification/
[4] IRIs complement URIs and are sequences of characters from the Universal Character Set (Unicode/ISO 10646) https://www.ietf.org/rfc/rfc3987.txt

biodiversity. Such a knowledge base serves as an information source for research in a wide range of other research domains, amongst which are environmental and climate change, public health, wildlife diseases, economics and biosecurity (8; 33; 9; 10). By analysing NHCs, researchers can identify key drivers of changes in biodiversity. Identifying such drivers is crucial given the biodiversity crisis the world is currently facing (34; 7).

1.2 Research Questions

The main objectives of this thesis are: 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**)*. From these objectives, we distilled the most important research questions. Each of the questions relates to our approach (Section 1.3), and to one of the chapters.

Various types of software applications exist that aim to make digital archives computationally searchable: some of these we detailed in Section 1.1. Through the first research question (**Q.1**), we investigate the most common types of software system designs and their trade-offs. From the output of **Q.1**, we propose an approach for knowledge extraction from manuscripts related to NHCs.

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

Through the second research question we investigate what kind of knowledge domain experts aim to extract, in order to facilitate rich queries over the content. We divide this question into two sub-questions. Through the first (**Q.2**a), we investigate what the main semantic concepts mentioned in these complex and heterogeneous archival collections are, in order to organise these according to an ontology (background described in Subsection 2.2.2). Through the second (**Q.2**b), we investigate how we can use the ontology to make the content machine-readable, in order to facilitate rich queries over the content that are in line with 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)

- **(a)** *What are the general semantics of historical species observations and how do they differ from present day observations?*

(b) *How do we extract important content and its semantics (e.g., core elements and their relationships) from the archives so that they become machine-readable, allowing rich queries over their content?*

An example process in biodiversity research that is of paramount importance to the field, is the process of determining the status of scientific names in a contemporary classification system, and their relationship to other names. An example is *taxonomic referencing:* linking a legacy name from a historical field note or other source to a scientific name accepted in current taxonomy (35). Through **Q.3**, we investigate how important links, such as taxonomic referencing, can be created and maintained using the FAIR principles. By doing so, references are made accessible to and reusable by any researcher, allowing references to be subjected to scientific discourse. Examples of other important references are links to literature, depictions, other resources, or named entities on the Web such as people or locations.

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)

Lastly, extracting information from heterogeneous, historical material is time-consuming and requires domain expertise. Through **Q.4**, we demonstrate methods that exploit prior knowledge for the development of automated methods that can help domain experts with the extraction and organisation of knowledge from archives of NHCs.

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

1.3 Approach and Main Contributions

The main theme of our approach is *prior* or *background knowledge*—e.g., community-developed data standards, domain background knowledge, and auxiliary datasets from the domain—and how it can be leveraged in the development of computational techniques for *knowledge extraction* from archives of NHCs.

By exploring various system design patterns for the transformation from manuscripts to databases, we aim to answer **Q.1**, leading to the first contribution:

C.1 A survey of system designs for turning manuscripts into databases (Chapter 3).

The output of this research has informed our approach for investigating items **Q.2** to **Q.4**, which in turn have lead us to formulate a generic system design for the disclosure of

manuscripts from NHCs. We discuss the approach and its main contributions in the following paragraph.

Semi-Automated Semantic Annotation The observation of a species is based on a specific set of units. Experts in the field have described these with a set of core terms for the purpose of sharing and integration of information about biological diversity (36).[1] Investigating **Q.2** has led us to formulate an approach that leverages these community-defined core terms and their ascribed interpretations for the purpose of knowledge extraction from observation records in manuscripts. Specifically, our approach focuses on the semantic annotation of *salient* named entities in *images* of field notes and illustrations, referring to the principal named entities that are visually used to structure a text such as *scientific names* and their *authors*, or the *location* and *date* of an observation event.

By using formal semantics to describe the named entities in handwritten and drawn observation records, richer querying and reasoning over the content becomes possible. Investigating **Q.3** has led us to formulate an approach to serve the annotated data as Linked Open Data (LOD) to the *Semantic Web*, where it can be interlinked and integrated with other collections using *Semantic Web technologies*. The approach has materialised into our second and third contribution:

C.2 The NHC-ontology[2] (Chapter 4).

C.3 The Semantic Field Book Annotator (SFB-Annotator)[3] (being further developed within the LInking Notes of NAturE (LINNAE) project),[4] (Chapter 4).

Investigating **Q.4** has led us to formulate an approach for NERC from field notes and illustrations detailing species observations. As these carefully employ the systematic organisation of species variations, computational techniques can exploit the systematic organisation of the document content. We use background knowledge on the structure of field notes and nomenclature, to design a computer vision-based deep learning model that can recognise and classify named entities from *text images* of field notes semi-automatically, with input and curation from domain experts. We refer to our approach by the term *salient named entity recognition and classification (SNERC)* to stress the recognition of *principal* entities in *text images*. See Figure 1.2 for an overview of our approach.

C.4 A method for SNERC from text images of field notes. (Chapter 5).

[1] https://www.tdwg.org/standards/dwc/

[2] http://www.makingsense.liacs.nl/rdf/nhc/,https://github.com/lisestork/nhc-ontology/

[3] https://github.com/LINNAE-project/SFB-Annotator,https://www.research-software.nl/software/sfb-annotator

[4] https://github.com/LINNAE-project

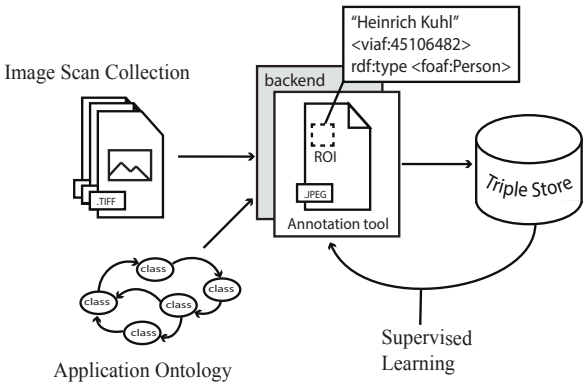


Figure 1.2: The proposed design pattern for the task of SNERC from text images of field notes.

Zero-Shot Classification The classification of scientific illustrations requires specific strategies that rely on prior knowledge, see the automation-related challenges in Subsection 1.1.3. Investigating **Q.4** has led us to formulate an approach for the classification of illustrations for the extreme scenario in which a classifier is asked to classify an image from a class it has not yet seen before during the learning process (zero-shot learning). Such learning relies on the sharing of knowledge between seen and unseen classes, by learning from auxiliary datasets that, jointly, possess knowledge about the union of seen and unseen classes. For our purpose, we use multimodal datasets from the domain (a biological taxonomy, historical texts and photographs) to train a classifier for zero-shot classification of illustrations. The output of the classification process is structured according to a contemporary biological taxonomy. The approach has lead to our fourth and fifth contribution:

C.5 A method for the classification of scientific illustrations from NHCs in a zero-shot classification scenario (Chapter 6).

C.6 The Zoological Illustration and Class Embedding (ZICE) dataset for zero-shot classification of zoological illustrations from NHCs (Chapter 6).

1.4 Project Context

The project *Making Sense of Illustrated Handwritten Archives*[1] ran between 2016 and 2020. The aim of the project was to develop a technologically advanced and user-centered digital environment for the disclosure of handwritten and illustration archives of NHCs, enabling

[1] http://www.makingsenseproject.org

the study of underexplored scientific heritage collections in general. As a use case, the project focused on computationally interpreting the notes and illustrations of the Committee for Natural History of the Netherlands Indies ("Natuurkundige Commissie voor Nederlands-Indië") (NC), which we further discuss in Subsection 2.3.2. The NC collection accumulated between 1820 and 1850 by 17 European naturalists and draftsmen and local helpers, and contains an account of natural historical exploration in the Indonesian Archipelago during that time period. The strength of the project was its interdisciplinary approach. The consortium had domain expertise in cultural history, the history of science (STePS, NBC) taxonomy (NBC), HTR (ALICE), computer vision and knowledge representation (LIACS). Results were consolidated into one digital infrastructure, the environment Natural Committee Online (NCO) (37),[1] co-developed by the publishing house Brill (Leiden).

1.5 Outline of Chapters

This thesis is based on a series of publications. The chapters each present work that has been peer reviewed and published, with the exception of this chapter, and Chapter 2, which serves as background for the thesis. Each chapter is self-contained, but contributes to the overall goal of semi-automated knowledge extraction from archives of NHCs. The thesis is structured as follows:

Chapter 1 has served as an introduction to the thesis, and is loosely based on the position paper of the project *Making Sense of Illustrated Handwritten Archives*:

- Weber, A., Ameryan, M., Wolstencroft, K., **Stork, L.**, Heerlien, M., and Schomaker, L. Towards a digital infrastructure for illustrated handwritten archives. In M. Loannides, editor, Digital Cultural Heritage, volume 10605 of Lecture Notes in Computer Science, pages 155–166. Springer International Publishing, April 2018. https://doi.org/10.1007/978-3-319-75826-8_13

Chapter 2 starts off with a background on natural history research and introduces techniques, datasets and use cases used for analyses in the chapters that follow.

Chapter 3 presents the state of the art relating to digitisation of archival content, and is based on the following conference paper:

- **Stork, L.**, Weber, A., Herik, J. van den, Plaat, A., Verbeek, F., and Wolstencroft, K. From handwritten manuscripts to linked data, In: Méndez E., Crestani F.,

[1] https://labs.brill.com/makingsense/,https://dh.brill.com/nco/

Ribeiro C., David G., Lopes J. (eds) Digital Libraries for Open Knowledge. TPDL 2018. volume 11057 of Lecture Notes in Computer Science, Springer, Cham https://doi.org/10.1007/978-3-030-00066-0_34

Chapter 4 presents our infrastructure and tooling for the transformation of digitised archives to structured knowledge bases. The chapter is based on the following journal paper and conference abstract:

- **Stork, L.**, Weber, A., Gassó Miracle, E., Verbeek, F., Plaat, A., Herik, J. van den, and Wolstencroft, K. Semantic annotation of natural history collections. Web Semantics: Science, Services and Agents on the World Wide Web (2018), https://doi.org/10.1016/j.websem.2018.06.002
- **Stork, L.**, Weber, A., Gassó Miracle, E., and Wolstencroft, K., A workflow for the semantic annotation of field Books and specimen labels, in Biodiversity Information Science and Standards 2: e25839 (2018) https://doi.org/10.3897/biss.2.25839

Chapter 5 presents a method for automating part of the tooling discussed in Chapter 4, and is based on the following conference paper:

- **Stork, L.**, Weber, A., Van den Herik, J., Plaat, A., Verbeek, F., Wolstencroft, K., Automated semantic annotation of species names in handwritten texts, In: Fuhr, N., Azzopardi, L., Stein, B., Hauff, C., Mayr, P. & Hiemstra, D. (eds.) Advances in Information Retrieval: 41st European Conference on Information Retrieval Research (ECIR), 2019. vol. 11437 of Lecture Notes in Computer Science, Springer, Cham. 667-680 14 p. https://doi.org/10.1007/978-3-030-15712-8_43

Chapter 6 discusses the classification of zoological illustrations, and presents an approach to deal with such statistically highly complex data: heterogeneous content, and a long-tailed scenario where many categories have only few or no examples for training. The approach exploits multimodal historical and contemporary data sources from the Web. The chapter is based on the following journal paper:

- **Stork, L.**, Weber, A., van den Herik, J., Plaat, A., Verbeek, F., & Wolstencroft, K. (2021). Large-scale zero-shot learning in the wild: Classifying zoological illustrations. Ecological Informatics, 62, 101222. https://doi.org/10.1016/j.ecoinf.2021.101222

In **Chapter 7**, we conclude our thesis with an overview of findings and their broader implications.

CHAPTER 2

Background

"Whether in literary criticism or scientific investigation, the academic mind is best at taking things apart. The complementary arts of integration are far less well developed. As with any interdisciplinary pursuit, it is the bridging across disparate ways of knowing that is the constant challenge."

– Richard J. Borden, in: *Ecology and Experience*

In this chapter we present the background of the thesis. Owing to its interdisciplinary nature, we cover a number of diverse topics, ranging from *natural history*, which is the application context of this thesis (Section 2.1), to methods from two sub-fields of computer science and artificial intelligence (AI): *symbolic*, and *subsymbolic* artificial intelligence, each with their own academic legacy. We discuss both under the umbrella of the task *knowledge extraction* (Section 2.2). We close the chapter by detailing the use cases and datasets that formed the basis for analyses described in subsequent chapters (Section 2.3).

2.1 Natural History

Biodiversity research aims to understand the whole of life on earth, its evolution and the various factors that generate its diversity. The field is usually subdivided into three levels, in which diversity is measured and researched: species, genetics and ecology. In this thesis, we focus on research into the diversity of *species*. Inherent to species research is the comparison and classification of the various plants and animals that inhabit our world. In order to realise this, naturalists in the field are challenged to develop methods that moderate systematic descriptions. Expeditions to biodiverse areas allow naturalists to record organism observations and identifications, enabling them to extend, improve or challenge existing classifications.

2.1.1 Taxonomy and Nomenclature

In the first part of the 18th century, Carl Linnaeus published his *Systema Naturae*, a system that formed the basis for biological taxonomy and nomenclature. From then on, naturalists and taxonomists started to use *taxonomy* and *binomial nomenclature* for the hierarchical classification and systematic naming of organisms. Therefore, most historical records found today in museums and other institutions (38), as well as contemporary biodiversity datasets, use biological taxonomy and binomial nomenclature to classify and describe their specimens and observation records.

Taxonomy. In the *Systema Naturae*, Linnaeus presented ideas for the *hierarchical classification* of species. By his system of classification, the natural world was organised into three *kingdoms*: the *animal kingdom*, the *plant kingdom*, and the *mineral kingdom*, although his system for the classification of minerals was never widely adopted by the scientific community. Species were grouped based on shared traits into units called *taxa*, which were in turn organised hierarchically into six nomenclatural *ranks* that increasingly share more traits: *kingdom*, *class*, *order*, *genus*, *species* and *variety*, according to a subsumption relationship. For example: a common octopus is an octopod (of the order octopoda), as well as a cephalopod (of the class cephalopoda), as well as an animal (of the kingdom animalia), see Figure 2.1. More recent subdivisions that have been added over the years are *phyla*, *families* and *tribes*, and subranks such as *subspecies*, or *subtribes*.

```
Animalia [kingdom]
└── Mollusca [phylum]
    └── Cephalopoda [order]
        └── Octopoda [class]
            └── Octopodidae [family]
                └── Octopus Cuvier, 1797 [genus]
                    └── Octopus vulgaris Cuvier, 1797 [species]
```

Figure 2.1: Classification of the species *Octopus vulgaris Cuvier, 1797*, a common octopus.[1] Edges represent the "part–of" relationship.

Binomial nomenclature. *Binomial nomenclature* translates into *two-term naming system*, and was introduced to formally name species according to one system. The idea of a two-term naming system was first put forth by Linnaeus in 1753 in his work *Species Plantarum*. A systematic name in binomial nomenclature is called a *binomial name*, also known as a *scientific name*. Octopus vulgaris (Figure 2.1) is an example of a binomial name. The first of the two terms identifies the genus to which the organism belongs, and

[1] https://www.gbif.org/species/2289671

the second is called the *specific epithet*, and points to the specific species within that genus. Commonly, the binomial is followed by the name of the author who published the name, and the date when the name was published in literature. It is also common for a name to have more than one author. Figure 2.2 shows an example of a scientific species name from a field note; it dates back to 1821.

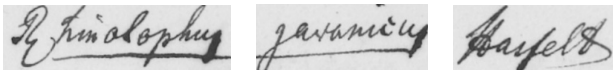


Figure 2.2: A scientific name in binomial nomenclature: (a) *Rhinolophus* (genus) (b) *javanicus* (specific epithet) (c) *Hasselt* (author of the name: *Johan Coenraad van Hasselt*)

Taxonomic Debates and Name Ambiguity. During the development of biodiversity research, methods of biological classification were continuously subject to intense discussion (39). Multiple theories emerged regarding collection practices and classification. In particular in the early nineteenth century and before, naturalists were struggling to find and agree upon one 'true' natural system. NHCs embody this search for a terminological structure which could be used to order, describe and classify nature.

The lack of consensus on biological classifications, as well as the challenges that came with the publication of scientific names—the very act of bringing home the actual observation records as well as tensions that arose through top-down policy-making (16)—resulted in species descriptions that are challenging to analyse within the present scientific paradigm, but also within collections themselves: (i) biological classification systems implied in field books cannot be directly mapped to present taxonomies (ii) taxa have various types of synonyms and homonyms within collections, see Figure 2.3, and Figure 2.4, and (iii) scientific names shift between genera and species (39; 29; 40).

```
Scotophilus kuhlii temminckii (Horsfield, 1824) [current name]
├── Vespertilio temminckii Horsfield, 1824 [synonym]
└── Vespertilio fulvus Kuhl & Van Hasselt [synonym]
```

Figure 2.3: Synonyms of the current taxon `Scotophilus kuhlii temminckii`. Courtesy: E. Gassó Miracle (2016)

Due to this, *taxonomic referencing* (resolving historic scientific names to current scientific names) of historical observation records, as well as establishing links between scientific names in general, are important processes in species research. Thomer et al. (2012) (35) describe taxonomic referencing as the process of linking a legacy name to its valid scientific name. They mention the process is analogous to that of georeferencing for localities.

Similarly to georeferencing, the process helps to integrate data related to the same entities, as well as separate data from unrelated ones.

```
Orestias elegans [hemihomonym]
├── Orestias elegans Garman, 1895 [accepted name]
└── Orestias elegans Ridley, 1887 [accepted name]
```

Figure 2.4: Hemihomonymy, an accepted form of homonymy where two species have the same name, but come from distinct kingdoms: the first referring to an animal (a pupfish), and the second to a plant (an orchid).[1]

2.1.2 Multimodal Field Observations

Early field observations exist in natural history museums as physical **specimens**, accompanied by archival material such as handwritten **field books**, and **illustrations**. Museums keep historical records for comparison with contemporary records. Many collections date back 100 years or more (3). Through more recent next-generation techniques such as photogrammetry, laser scanning, and computed tomography, rich digital representations of specimens as well as manuscripts (e.g., the Dutch Metamorfoze programme[2]) can be created (7). Below we discuss each modality and its characteristics.

Specimens. Specimens that are commonly kept in natural history museums are: fluid-preserved whole organisms and organism parts, frozen tissues, pinned dried insects, pressings and seeds or spores of plants, dried skins, skeletons, nests of birds and eggs of birds and insects (3). Figure 2.5 shows a skeleton of a *Pteropus vampyrus* from the Naturalis Biodiversity Center (NBC).

It is common for such specimens to be accompanied by labels containing metadata regarding the specimen, such as the name of the collector, the scientific name, location and date of collection, although metadata are often limited to a location or naturalist that performed the identification or collected the specimen. In most cases, labels include scientific names, but do not record any scientific context (29), for instance regarding the literature used for classification. As alternative views on taxonomy exist, mentioned earlier in Subsection 2.1.1, one name can point to two very distinct species. Linking the physical specimens to observation records becomes crucial. When a specimen is accompanied by a record of the organism's latent, faded or internal traits and attributes (e.g. behaviour, coat colour, or intestines), identification of the preserved specimen can be revisited and its value for use in long-term scientific studies therefore increases.

[1] https://species.wikimedia.org/wiki/Orestias_elegans
[2] https://www.metamorfoze.nl/kennis-onderzoek/lexicon/preservation-imaging

Figure 2.5: A specimen from the *Naturalis - Zoology and Geology catalogues.*[1]

Field Books. Since the onset of field work in biodiversity expeditions, naturalist have been manually recording species observation data. The containers that preserve these observation records are fittingly named *field books* (41), see Figure 2.6. They provide rich descriptions of species-specific traits such as measurements of specific organs or other body parts, the environmental conditions in which organisms are discovered and information about how organisms were collected, classified and described. Because of this, field books provide rich insight into the daily practices, methods, and results of the research field (33).

The interpretation of historical field records is an intricate and complex task. We demonstrate this complexity with the use of an example. The field note shown in Figure 2.6 describes an occurrence of an organism identified as the *Titthaecheilos javanicus Nobis* (right page, upper left corner).

Nobis is latin for *by us*. The space behind the binomial name is reserved for the author of the species. Therefore, the term *by us* refers to the authors of the field book: according to them, they were the first ones to have identified, described and named the organism. The name *Titthaecheilos javanicus* has, however, never been published in any classification system. Most likely, the name served as a basionym[2] for the published name *Pteropus titthaecheilus Tem.* (upper right corner) believed to have been added to the field note in Leiden, years later, by *Jacob Coenraad Temminck*, a dutch zoologist and museum director. The name can be found in older classification systems as the name *Pteropus*

[1] `https://data.biodiversitydata.nl/naturalis/specimen/RMNH.MAM.33245.a` Images free of known restrictions under copyright law (Public Domain Mark 1.0).

[2] A basionym is a synonym on which a later scientific name is based.

titthaecheilus (Temminck 1825). In turn, that name served as a basionym for the accepted name *Cynopterus titthaecheilus (Temminck, 1825)*.

Moreover, below the scientific name we find another name type: *Buitenzorg*, a place name. Historically, Buitenzorg was the name for the large city of *Bogor*, close to the capital of *Java*, *Jakarta*. The city houses the largest botanical garden in the world, the botanical garden of *Bogor*, which served as the headquarters of the NC. Last, the field note is written in a distinct, historical style, and mixes three languages: the note starts in Dutch, continues in German, and ends in Latin.

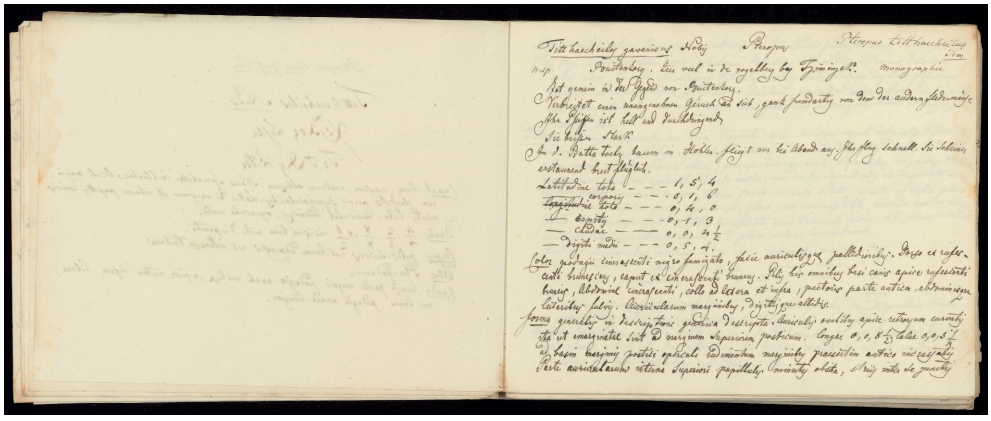


Figure 2.6: A page from the annotated field book describing the species *Titthaecheilos javanicus Nobis*. *Pteropus titthaecheilus Tem* (upper right corner) is believed to have been added later in Leiden by *Jacob Coenraad Temminck*, `http://viaf.org/viaf/69703180`, a dutch zoologist and museum director. The written annotation is thus an additional identification of the observed organism. Collection Naturalis Biodiversity Center, MMNAT01_AF_NNM001001033_013.[1] Image free of known restrictions under copyright law (Public Domain Mark 1.0).

Illustrations. Historically, collectors were accompanied by professional illustrators, who produced detailed drawings of organisms, as shown in Figure 2.7. The habitus illustration—a scientific illustration of a species' physical appearance—was the most important medium to convey a species' characterising traits to other scientists. In illustrations, scientists are capable of delineating and highlighting minuscule details, often more so than photographs. Habitus illustrations were routinely and abundantly created and commonly served as examples for the description of newly discovered species, so-called holotypes. Additionally, they sometimes recorded the habitat or behaviour of an organism.

[1] `https://dh.brill.com/nco/view/nco_NNM001001033_013/makingsense`

In illustrations, the background (natural habitat) is often omitted and species are depicted in the form of collages of multiple (smaller) depictions of their external and internal anatomy (e.g., bones, organs, limbs). These appear in a combination of various views (e.g., frontal, dorsal, lateral). Moreover, illustrations exist as rough pencil sketches and/or detailed colour drawings and commonly contain handwritten captions. Often, they are published in digital archives with limited or no identifications. When illustrations contain captions with handwritten *historical names*, these are mostly unpublished or obsolete within today's taxonomy. The left illustration in Figure 2.7 says *Asterias tesselatus (Lamarck, 1816)*, an unaccepted name, and *Asterias granularis Kuhl*, an unknown name. The current accepted name of the starfish is the *Goniaster tessellatus (Lamarck, 1816)*. The middle photograph has some unreadable text in the upper right corner, and a pencil annotation that most likely reads *Noae Lam.*, appearing to refer to a genus published by Jean-Baptiste de Lamarck.[1] The current accepted name of the species is *Arca noae (Linnaeus, 1758)*.

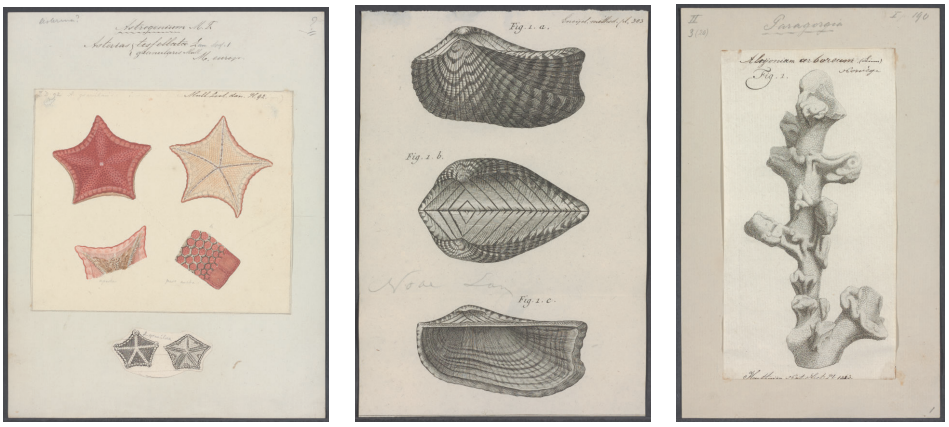


Figure 2.7: Zoological illustrations from Iconographia Zoologica online[2] (best viewed in colour). Images free of known restrictions under copyright law (Public Domain Mark 1.0).

The identification of an organism from a photograph or illustration without reference to a scientific name, is a complex and delicate task, even for domain experts (42).

2.2 Knowledge Extraction

In this section, we introduce the preliminaries used throughout this thesis. Our approach employs techniques from subsymbolic AI (e.g., *computer vision*) well as symbolic AI (e.g., *Semantic Web*) , for the purpose of knowledge extraction.

[1] Jean-Babtiste de Lamarck, a French naturalist. URI: `https://viaf.org/viaf/41849820/`
[2] `https://bijzonderecollecties.uva.nl/gedeelde-content/beeldbanken/iconographia.html`

We take the definition of the term *information extraction* from (43), and use this as a red thread that weaves through the thesis:

Definition 2.1. *"Information extraction is the process of extracting information and turning it into structured data. This may include populating a structured knowledge base with information from an unstructured knowledge source. The information contained in the structured knowledge base can then be used as a resource for other tasks, such as answering natural language queries, or improving on standard search engines with deeper or more implicit forms of knowledge than that expressed in the text".*

Knowledge extraction is a form of information extraction. It uses similar methods, but the main criteria is that results of the extraction process are structured according to formalised semantics such as taxonomies or ontologies (which we will discuss in Subsection 2.2.2).

Examples of information and knowledge extraction tasks are *semantic annotation* and named entity recognition and classification (NERC), both described earlier in Subsection 1.1.3, in which *ontologies*—formal specifications of concepts and their relationships—play a large role in the information extraction process. A similarity between these tasks and our work is that we leverage domain specific ontologies and taxonomies for knowledge extraction. One major distinction between these and our work is that we extract knowledge from *digital images*, whereas commonly, an intermediate step transforms the content of images to digital text, to which then information extraction is applied.

In the following sections, we will detail the most important notations, concepts and techniques used throughout this thesis, which fall under the umbrella of **machine learning**, used to automatically extract patterns (subsymbolic AI, Subsection 2.2.1) and **knowledge representation and reasoning (KRR)**, used to organise the patterns semantically (symbolic AI, Subsection 2.2.2).

2.2.1 Machine Learning

In machine learning, *learning algorithms* learn from data to perform a certain task: e.g., hypothesise to which category y or target value t a sample belongs. The type of learning algorithm or **model** that is used depends on a number of things, such as (i) the **structure** of the **data**, (ii) the **task**, (iii) the kind of **experience** the models are allowed to have during the learning process (simulating certain real-world situations) (44).

Data Structures. Examples of data types often used in machine learning tasks—and that we will use in this thesis—are digital images, sequences such as sentences.

- **Digital images** are digital captions of scenes or pictorial materials. They represent a coherent collection of focus points of light rays coming from an object. A digital image divides the real image into a grid of real numbers, called pixels, which discretise properties of the underlying areas such as *brightness* and *hue*. The process of digitisation of the spatial domain is called *sampling*. Discretising the range in which these real numbers fall is called *quantisation* (45). A gray-scale image is a 2-dimensional (2D) digital image, of which each value represents a pixel that samples the brightness of that pixel. Commonly, the brightness range is encoded in 256 (2^8) levels (values from 0 to 255), corresponding with an 8-bit discretisation. RGB images sample three values per pixel, also called *channels*: the brightness of the red, green and blue values. Similarly to gray-scale images, these three channels are commonly encoded in 256 levels. Both sampling and quantisation depend on the imaging device that is used. The resulting multidimensional array of real numbers can be stored and handled by a digital computer.

  We define a digital image (either gray-scale or RGB), with m rows and n columns, as follows:

  Definition 2.2. *A gray-scale digital image* $\mathbf{X}$ *is a 2D numerical array, or* matrix*, with* $x_{ij} \in \mathbb{R}$ *being the gray value of the pixel in the* i*-th row of and the* j*-th column, see the matrix representation in Equation* (2.1) *below.*

$$\mathbf{X} = \begin{bmatrix} x_{1,1} & x_{1,2} & \dots & x_{1,n} \\ x_{2,1} & x_{2,2} & \dots & x_{2,n} \\ \vdots & \vdots & \vdots & \vdots \\ x_{m,1} & x_{m,2} & \dots & x_{m,n} \end{bmatrix} \tag{2.1}$$

  An RGB image is a multi-channel digital image in which each channel represents a colour layer. This can be considered a 3D numerical array, or tensor*, with* $x_{ijk} \in \mathbb{R}$ *being the value of the pixel in the* i*-th row, the* j*-th column, and* k*-th colour channel. A tensor representation of an RGB image is shown in Equation* (2.2) *below.*

$$\mathrm{X} = \begin{array}{l} \begin{bmatrix} x_{1,1,1} & x_{1,2,1} & \dots & x_{1,n,1} \\ x_{2,1,1} & x_{2,2,1} & \dots & x_{2,n,1} \\ \vdots & \vdots & \vdots & \vdots \\ x_{m,1,1} & x_{m,2,1} & \dots & x_{m,n,1} \end{bmatrix} \\ \begin{bmatrix} x_{1,1,2} & x_{1,2,2} & \dots & x_{1,n,2} \\ x_{2,1,2} & x_{2,2,2} & \dots & x_{2,n,2} \\ \vdots & \vdots & \vdots & \vdots \\ x_{m,1,2} & x_{m,2,2} & \dots & x_{m,n,2} \end{bmatrix} \\ \begin{bmatrix} x_{1,1,3} & x_{1,2,3} & \dots & x_{1,n,3} \\ x_{2,1,3} & x_{2,2,3} & \dots & x_{2,n,3} \\ \vdots & \vdots & \vdots & \vdots \\ x_{m,1,3} & x_{m,2,3} & \dots & x_{m,n,3} \end{bmatrix} \end{array} \tag{2.2}$$

- **Sequences** are digital representations of values that are meaningful in a certain arrangement, such as sentences, digital texts or even sequences of images that represent

words in a sentence.

We define a sequence of values as:

Definition 2.3. *A sequence* ***s*** *is a finite set of values* $(\mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(t)})$ *that are tightly interrelated, where* t *indicates the length of the sequence, and* $\mathbf{x}^{(i)}$ *the* $i-$*th value of the sequence.*

In classical machine learning, "raw" data are first transformed into *features* (numerical representations of raw data) that represent the variations in the data well. This process is called *feature engineering*.

Machine Learning Tasks. Below, we discuss two machine learning tasks used in this thesis, but many more exist (e.g., regression or clustering).

- **Classification:** In a classification task, a machine is asked to infer to which of k categories some input belongs. The learning algorithm is asked to produce a function (*a model*) $f : \mathbb{R}^n \rightarrow \{1, \ldots, k\}$ (44). An example of this is image classification, in which an image $\mathbf{X}$ is mapped to the category $y \in \{1, \ldots, k\}$ to which the machine thinks the image belongs, each number representing a class. For example, an image $\mathbf{X}$ of a *bear* gets mapped to the label $y^{(i)}$ that encodes the *bear* class, or an image of the word *Pteropus* gets mapped to the label $y^{(i)}$, representing the class *Pteropus*. A machine learning model trained for classification generally produces a decision boundary, see Figure 2.8, that separates data from distinct classes (in Figure 2.8, the decision boundary separates instances from the red class from that of the black class). The model classifies a new data point as belonging to a certain class by calculating on which side of the boundary it lies.
- **Classification with structured output:** In *classification with structured output*, a machine is asked to produce, given some input, several values that are all *tightly interrelated*: a sequence **s**. Examples are (i) image captioning, where a machine receives an image as an input and outputs a sentence that describes the image, or (ii) NERC, where a machine receives a sentence, and returns the same sentence with its named entities annotated with terms from a structured knowledge base.

Learning from Experience. Below, we discuss two types of learning strategies that vary in the amount of experience they are allowed to have during the learning process.

- **Supervised learning:** Most types of learning algorithms get to "see" all training examples $x \in \mathbb{R}^n$ and their labels $y \in Y^{tr}$ (classification) or targets $t \in \mathbb{R}^n$ (regression), and are therefore supervised in the sense that they are instructed as to what the output

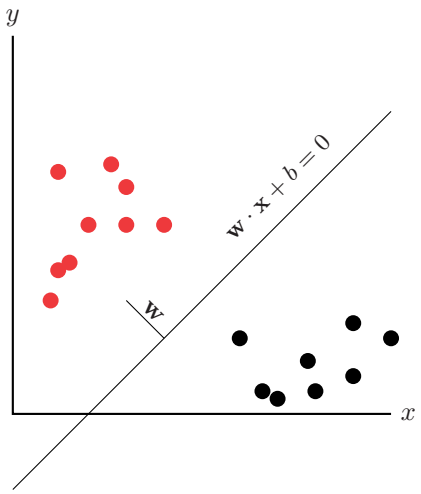


Figure 2.8: Example decision boundary of a binary classifier.

for a certain input should be. From the training data, the machine learning algorithm is asked to learn the mapping between input and output, and use it for extrapolation to new data from a test-set.

- **Zero-shot learning (ZSL):** ZSL is an extension of supervised learning in which the test-set represents a distinct set of classes $y \in Y^{ts}$, for which $Y^{tr} \cap Y^{ts} = \emptyset$. The goal of ZSL is that a classifier learns representations for data from a set of *seen* classes Y^{tr} (seen by the algorithm) with medium to large samples, which are then transferred to classify classes from another set of *unseen* classes Y^{ts}, for which no or small samples are available for training.

  An appealing characteristic of ZSL techniques is that it is possible to exploit data from auxiliary data sources to share representations between classes, and hereby push the boundaries of automated recognition for a specific problem. As with regular supervised learning, it can be difficult in some cases to control which features are shared.

Other popular types of learning are unsupervised learning and semi-supervised learning, but these are out of the scope of this thesis.

Deep Learning Models. Deep learning is a subfield of machine learning that brings forth a specific type of machine learning models, called *deep (artificial) neural networks (DNNs)*. DNNs are able to learn representations of data from data (46), replacing part of the feature engineering pipeline. These learned representations then allow computers to perform a machine learning task. Artificial neural networks (ANNs) are models that are

inspired by the human brain, as they are trained to strengthen and weaken connections between input variables, much like the brain's networks of neurons. DNNs are types of ANNs that are trained to learn *deep*, *hierarchical* representations of data—with multiple levels of abstraction (46). For the purpose of classification, they define a mapping between an input array $\mathbf{x} = (x_1, \ldots, x_n)$ and an output y, $y = f(\mathbf{x}; \boldsymbol{\theta})$, and learn the value of parameters $\boldsymbol{\theta}$ that defines the best function approximation.

Below we discuss types of DNNs that are used in this thesis.

- **Multi-layer perceptrons (MLPs):** An MLP is an example of the simplest type of DNN, see Figure 2.9. It is a type of feed-forward neural network, meaning that the multiplications flow 'forward' in one direction through the network. Although the network represented here has one level of abstraction (one *hidden layer*), DNNs usually have many. By adding multiple hidden layers, we increase the network's *depth*. In Figure 2.9, each connection represents a multiplication with a weight. In this figure, there are two weight matrices $\mathbf{W}^{(i)}$, the superscript denoting the i-th weight matrix, one between the input and the hidden layer, and one between the hidden layer and the output. We call these layers *fully connected*, as each node in one layer is connected with every other node in the next layer, i.e., has its own weight.

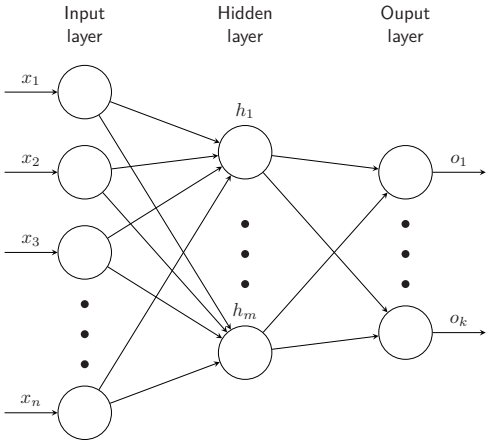


Figure 2.9: An example of an MLP with one layer of abstraction. The network has an n-dimensional input, m hidden nodes and k outputs (classes).

See Equations (2.3) to (2.5) for the network's mapping between $\mathbf{x}$ and y. Equation (2.3) and (2.4) are called node *activations*. They show an additional parameter $b_j^{(i)}$, which stands for *bias* and serves to shift the activation function and thereby the classification boundary by adding a constant.

$$h_i = \sum_j w_{ij}^{(1)} x_j + b_j^{(1)} \tag{2.3}$$

$$o_i = \sum_j w_{ij}^{(2)} x_j + b_j^{(2)} \tag{2.4}$$

$$y = argmax(\mathbf{o}) \tag{2.5}$$

While an MLP is a linear model, most ANNs employ at least one layer in which a non-linearity function g, called *activation function*, is applied to the activation of each of the neurons in a hidden or output layer, as in Equation (2.6). Common activation functions are the *ReLU* or the *softmax* functions.

$$h_i = g\left(\sum_j w_{ij}^{(1)} x_j + b_i\right) \tag{2.6}$$

The activation of the last layer (often a softmax function applied to the activation of the last layer) produces a distribution over output classes, also called *confidence values*, which correspond to the distance of an instance to the decision boundary, and thereby how confident the classifier is about that class being the one represented in the data. This intuitively makes sense, as when our instance lies very close to the decision boundary (or boundaries for multi-class classification), it is more likely to actually belong to the other class than when it's further away.

- **Convolutional neural networkss (CNNs):** Whereas the MLP is used for processing 1D arrays, CNNs are used for processing grid data. They are often used in *computer vision* for classification of images. Characteristically, they use a mathematical operation called a *convolution*. In image classification, a convolution is a function that extracts features from regions of pixels in an image (47). The kernel $\mathbf{K}$, a small matrix of integers, acts as a sliding window that is moved over the digital image $\mathbf{X}$ and produces a weighted average with the underlying pixels, see Equation (2.7) (44).

$$S(i,j) = (\mathbf{X} * \mathbf{K})(i,j) = \sum_m \sum_n \mathbf{X}(i-m, j-n)\mathbf{K}(m,n). \tag{2.7}$$

Hence, instead of acting on the full input, kernels act on subregions of an image. Digital images exist of large 2D or 3D arrays, so employing kernels makes networks easier to train: parameters are shared over multiple regions of the input, so a much smaller number of parameters need to be optimised then when the layer would be fully connected.

Convolving regions in an image with kernel $\mathbf{K}$ results in an output matrix called a *feature map*, which gives an indication of whether or not the features in the kernel are present in certain regions of the image: similar regions produce similar output values. CNNs are types of DNNs as they stack layers of convolutional operations to extract image features on various levels of granularity, from fine-grained features such as *corners* and *edges* to coarser, class-specific features such as *eyes*, *feathers*, a *beak*—even though these coarser features are never that clear-cut. Similarly to an MLP, classification happens in the last fully connected layer, see Figure 2.9. The array of 2D feature maps is re-arranged to a 1D array and act as input to a fully connected layer. Often, engineers employ some fully connected layers before the final classification layer.

- **Recurrent neural networks (RNNs):** RNNs (48) are other types of DNNs for processing sequences of values $\mathbf{t} = (\mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(t)})$. Like CNNs, they share weights and perform a single computation multiple times over a different part of the input, here called $\mathbf{x}^{(t)}$ with time step t. Different from CNNs, they have a *recurrence* property, which means they use output at time t to serve as auxiliary input to a hidden layer at the next time step $t+1$. Effectively, the recurrence property serves as a *memory* that uses past computations $\mathbf{h}^{t-1}$ to influence present computations $\mathbf{h}^t$. Their basic operations are detailed below in Equation (2.8) and (2.9):

$$\mathbf{o}^t = f(\mathbf{h}^t; \boldsymbol{\theta}) \tag{2.8}$$

$$\mathbf{h}^t = g(\mathbf{h}^{t-1}, \mathbf{x}^t; \boldsymbol{\theta}) \tag{2.9}$$

where $\mathbf{o}^t$ is the output at time t, $\mathbf{h}^t$ the state of the hidden layer at time t, and $\mathbf{x}^t$ the input array at time t. f and g serve as activation functions.

Long short-term memory networks (LSTMs) are types of RNNs that overcome some of the issues that occur with regular RNNs. They have a *bilateral* variety, the bilateral long short-term memory network (BLSTM), that can additionally use *future* computations $\mathbf{h}^{t+1}$ to influence present computations $\mathbf{h}^t$.

- **Prototypical neural networks:** Prototypical networks are networks developed for low-shot learning strategies such as few-shot learning (FSL) or ZSL (49). They compute M-dimensional class representations $\mathbf{c}_k \in \mathbb{R}^M$ called *class prototypes*. In contrast to the other DNNs that we discussed, classification does not happen based on a distribution (softmax activation) over the last fully connected layer. Instead, the last fully connected layer maps instances to a metric space, and a distribution over distances

from an instance to class prototypes is produced. Example distance functions are *euclidean distance*, see Equation (2.10), and *cosine similarity*, see Equation (2.11).

$$d(\boldsymbol{p}, \boldsymbol{q}) = \sqrt{\sum_{i=1}^{n} (q_i - p_i)^2} \tag{2.10}$$

$$\cos(\boldsymbol{p}, \boldsymbol{q}) = \frac{\sum_{i=1}^{n} p_i q_i}{\sqrt{\sum_{i=1}^{n} (p_i)^2} \sqrt{\sum_{i=1}^{n} (q_i)^2}} \tag{2.11}$$

To calculate prototypes for FSL, *support points* (example datapoints) are mapped to a metric space, and per-class averages of the resulting embedded support points are calculated, see Equation 2.12. In Equation 2.12, $\mathcal{S}_k$ refers to the set of support points for class k, and $\mathbf{c}_k$ refers to its calculated prototype. We further refer to this metric space by the term *prototype space*.

$$\mathbf{c}_k = \frac{1}{|\mathcal{S}_k|} \sum_{(\mathbf{x}_i, y_i) \in \mathcal{S}_k} f_\phi(\mathbf{x}_i) \tag{2.12}$$

For ZSL, Snell et al. (49) mention that rather than embedding support points in prototype space, prototypes can be constructed by embedding auxiliary information, such as class embeddings in the form of attribute annotations (50; 51), in prototype space. Attribute embeddings encode whether a certain attribute—from a set of predefined attributes—is present for a specific class. Attribute embeddings can be either binary or continuous, e.g., *{wing: 0.1, red: 0.4, tail: 0.7}*.

Training and Evaluation. After model selection, the model's parameters $\boldsymbol{\theta}$ are learned through iterative minimisation of the *training error*. One iteration commonly consists of minimally three basic steps.

1. The model is applied to a batch of training data (data from the *training-set* $\mathcal{T}^{tr}$).
2. A *loss function* is applied to the output of the model, which calculates the training error—a function over the difference between the output y and the desired output $\hat{y}$.
3. The training error is propagated backwards through the model using the *backpropagation algorithm*, and parameters $\boldsymbol{\theta}$ are adjusted via an optimisation algorithm (such as *gradient descent*).

How well the model performs on real-world data should be evaluated on a dataset that is separate from the training-set, the *test-set*, $\mathcal{T}^{ts}$. It often happens that the representations learned by the model too closely fit the variation in the training-set, and will therefore generalise poorly to new data. This phenomenon is called *overfitting*, and can for instance happen when samples in the training-set are too small to obtain a good representation, or when there are too many parameters in the model, causing it to learn too much of the variance in the data. A metric that is most used for classification is the *average accuracy*, see Equation (2.13) (in percentages).

$$\textit{Accuracy} = \frac{n \text{ correct predictions}}{n \text{ total predictions}} * 100 \tag{2.13}$$

The average accuracy metric is not always the best choice, as it does not correctly portray the predictive power of a model, especially when data are imbalanced (52). Let us first consider a binary classification problem, and a dataset with a uniform distribution over its classes. We produce a naive model, let us call it model g, that for every input predicts the class c^m, the class in which the majority of the data resides, i.e., the *majority guess*: $g : \mathbb{R}^d \rightarrow c^m$. The estimated average accuracy on the test-set would already be as high as 50%. Imagine now a classifier h, that is trained on a 5-class classification problem with a similar uniform distribution over its classes. If such a classifier h produces an average accuracy of 50%, it will have learned a much better data representation than our naive binary classifier, even though the average accuracy produced would be equal.

The accuracy metric is especially vulnerable to bias that skewed data introduces. If we would apply our naive classifier g to a dataset where 90% of the data are of class 0 and the rest of class 1, we would obtain an average accuracy of 90%.

Further details of specific models, learning strategies, evaluation metrics and other, can be found in the respective chapters.

2.2.2 Knowledge Representation and Reasoning

The field of knowledge representation and reasoning (KRR) is quite extensive, so we limit ourselves to techniques for structuring data and data about data (metadata), with a focus on the representation of data in the form of **knowledge graphs**. Through the use of **schemas** and **ontologies**, which impose constraints or assign attributes to data, new knowledge can be inferred (*reasoning*). We furthermore discuss the principles of Linked Data (LD), which allow knowledge graphs served on the Web to link together, forming a Web of semantic data, called the **Semantic Web**.

Structured data, in contrast to *unstructured data*, are data that are structured according to some data model, and can therefore be interpreted by machines. Unstructured data, such as free text, can be made machine understandable by adding structure to capture the implicit semantics. Below we define what it means to make implicit semantics of data accessible to machines (53), turning data into machine-understandable *knowledge*.

We distinguish three levels for structuring data, that vary based on their capability to express implicit semantics (54):

- **Controlled vocabularies:** Controlled vocabularies include shared terminologies and nomenclatures and define terms to formally describe concepts within a domain.

  In the biodiversity domain, for instance, community collaboration is used to create shared knowledge representations, in order to make effective use of existing data (36). For broad-scale analyses, biodiversity information must be readily available in digital form, published as FAIR data. Below, in listing 2.1, we show a piece of Extensible Markup Language (XML), taken from the Simple Darwin Core documentation[1] that structures biodiversity data according to the Darwin Core (DwC)[2] standard, a glossary of terms (properties) for the description of biodiversity records. Such a basic form of structuring data allows intelligent machine search over data, for example the aggregation of scientific names across distributed collections using the term `dwc:scientificName`.

```
<?xml version="1.0" encoding="UTF-8"?>
<SimpleDarwinRecordSet
    xmlns="http://rs.tdwg.org/dwc/xsd/simpledarwincore/"
    xmlns:dc="http://purl.org/dc/terms/"
    xmlns:dwc="http://rs.tdwg.org/dwc/terms/"
    xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
    <SimpleDarwinRecord>
        <dwc:basisOfRecord>Taxon</dwc:basisOfRecord>
        <dwc:scientificName>Centropyge flavicauda Fraser-Brunner 1933
        </dwc:scientificName>
        ...
        <dwc:kingdom>Animalia</dwc:kingdom>
        <dwc:phylum>Chordata</dwc:phylum>
        <dwc:class>Osteichthyes</dwc:class>
        <dwc:order>Perciformes</dwc:order>
        <dwc:family>Pomacanthidae</dwc:family>
        <dwc:genus>Centropyge</dwc:genus>
        <dwc:specificEpithet>flavicauda</dwc:specificEpithet>
        <dwc:taxonRank>species</dwc:taxonRank>
        ...
    </SimpleDarwinRecord>
</SimpleDarwinRecordSet>
```

Listing 2.1: A piece of simple darwin core

[1] `https://dwc.tdwg.org/simple/`
[2] `https://dwc.tdwg.org/`

Other than an unique definition of terms within a domain, the XML document does not capture a lot of meaning that can be interpreted by machines. A machine does not know how the term Centropyge relates to the term flauvicauda, or would not know the difference between the specific epithet flauvicauda, part of the genus Centropyge, and the epithet flauvicauda belonging to a different genus (should such a scientific name exist).

- **Taxonomies** extend controlled vocabularies with "is–a" (subsumption) relationships between terms and thus add hierarchy to 'flat' controlled vocabularies.

  A biological taxonomy, used to structure the scientific name in Figure 2.10, shows an example of how pieces of data can be related through the subsumption relationship, where a class lower in the hierarchy is connected to the one above it using the *is–a* relationship.

```
Animalia
└─ Chordata
   └─ Osteichthyes
      └─ Perciformes
         └─ Pomacanthidae
            └─ Centropyge
               └─ Centropyge flavicauda (Fraser-Brunner 1933)
```

Figure 2.10: A hierarchy of the species Centropyge flavicauda (Fraser-Brunner 1933), where the edges (from top to bottom) refer to the "is–a" subsumption relationship

- **Schemas and Ontologies** further extend taxonomies by distinguishing (hierarchically organised) types and properties (relationships), and allow the modelling of constraints, axioms and rules.

  Ontology in philosophy is the study of existence. More specifically, the study concerns itself with questions that relate to what types of entities exist, and how they relate to one another. In computer science, an *ontology* refers to a data structure that can be processed by machines (11; 55):

  Definition 2.4. *An ontology is a formal, explicit specification of a shared conceptualisation (56).*

  - formal: *an ontology has well-defined syntax and semantics,*
  - explicit: *an ontology can be represented and processed algorithmically*
  - shared: *an ontology is agreed upon in a community and facilitates communication between its member agents, and*
  - conceptualisation: *an ontology presents a model of the real world*

Similarly to philosophy, an ontology in computer science consists of a formally defined set of terms, and relationships (properties) that define how the terms are related (53). We will denote these consistently in the same script throughout this thesis: e.g., `class` for classes, *property* for properties, `instance` for instances of classes, and "*literal*" for literals (a value of some datatype).

Knowledge graphs use ontologies and database schemas to structure data according to a directed graph data model. We use the following definition:

Definition 2.5. *A knowledge graph is a graph of data intended to accumulate and convey knowledge of the real world, whose nodes represent entities of interest and whose edges represent relations between these entities (57).*

The Semantic Web. The *Semantic Web* is a network of shared, structured data and metadata. It is based on directed labelled graphs as data models for resources and their relationships, and IRIs such as HTTP URLs[1] for the description of these resources and relationships. The Resource Description Framework (RDF) is the World Wide Web Consortium (W3C)[2] recommended data format for graph data on the Semantic Web. The RDF data model uses triples for the description of resources, in the form:

$$\langle subject, predicate, object \rangle \tag{2.14}$$

An example being: $\langle$`:img.jpg`,*:creator*,"*Heinrich Kuhl*"$\rangle$. Each such triple forms a single arc in a directed labelled graph.

RDF Schema[3] and the Web Ontology Language (OWL)[4] are formalisms that provide vocabularies for structuring knowledge for various levels of expressiveness. Properties of terms can be used for *reasoning* over data, i.e., inferring new facts from a set of asserted axioms.

A simple example term that can be used to infer new facts from assertions is the term `owl:TransitiveProperty`, and it is defined as follows:

Definition 2.6. *A transitive relation $\xrightarrow{x}$ is a relation specifying that if A $\xrightarrow{x}$ B and B $\xrightarrow{x}$ C, then A $\xrightarrow{x}$ C.*

[1] Web-resolvable URIs

[2] The World Wide Web Consortium (W3C) is an international community for the development of standards on the Web. `https://www.w3.org/Consortium/`

[3] `https://www.w3.org/TR/rdf-schema/`

[4] `https://www.w3.org/OWL/`

Instances of the owl class `owl:TransitiveProperty` inherit this transitive property. *rdfs:subclassOf*, for example, is an instance of `owl:TransitiveProperty`. If we define each term from Figure 2.10 as a class that is connected to the class above it with the property *rdfs:subclassOf*, machines can infer from this statement that the class `Centropyge` is also a subclass of every class above it, such as the class `Chordata`.

The SPARQL Protocol and RDF Query Language (SPARQL)[1] is one of the query languages with which RDF graphs can be queried, making use of their graphical structure.

Through shared formalisms, distributed directed labeled graphs on the Web are linked together. The collection of directed labeled graphs on the Web are referred to as Linked data. Tim Berners-Lee, one of the inventors of the Web, suggests a 5-star scheme with which to deploy Linked Data, in which each step assumes employment of the previous step(s):[2]

★☆☆☆☆ *publish data on the Web in any format (e.g., PDF, JPEG) accompanied by an explicit Open License (expression of rights).*

★★☆☆☆ *publish structured data on the Web in a machine-readable format (e.g., XML).*

★★★☆☆ *publish structured data on the Web in a documented, non-proprietary data format (e.g., CSV, KML).*

★★★★☆ *publish structured data on the Web as RDF (eg Turtle, RDFa, JSON-LD, SPARQL)*

★★★★★ *In your RDF, have the identifiers be links (HTTP URLs) to useful data sources.*

These also relate to the Linked Data principles, mentioned earlier in Subsection 1.1.3. Serving data to the Web of Linked Data has many benefits: (i) the adoption of HTTP URIs for the representation of entities (such as named entities) ensures more accurate content descriptions and thereby allows for the computational disambiguation of terms, (ii) through the use of shared conceptualisations, data become interoperable, and (iii) through use of Semantic Web services, data can be federated and integrated between distributed collections (11).

2.3 Data Sources and Use-Cases

In this section, we discuss contemporary online biodiversity data sources relevant for this thesis (Subsection 2.3.1), and detail two NHC use-cases (Subsection 2.3.2) used for

[1] `https://www.w3.org/TR/sparql11-query/`

[2] `https://dvcs.w3.org/hg/gld/raw-file/default/glossary/index.html#x5-star-linked-open-data`

analyses in further chapters.

2.3.1 Data Sources

- **The Global Biodiversity Data Facility (GBIF)** is an international network and data infrastructure for biodiversity data on the Web (58). Through the use of data standards, the organisation has pulled together hundreds of millions of species occurrence records. These include record types from multiple sources, such as, museum specimens, scientific expedition data and photos taken by amateur naturalists. GBIF provides an API, allowing data to be downloaded as *Darwin Core archives*,[1] a Biodiversity Information Standards (TDWG) standard. Additionally, they provide other resources such as the GBIF backbone taxonomy (59), a single taxonomy that organises all names included in GBIF according to one taxonomic system, which integrates information from external resources such as the Encyclopedia of Life (EOL),[2] the International Union for Conservation of Nature and Natural Resources (IUCN)[3] and GenBank.[4] It includes knowledge such as whether a name is accepted, what synonyms it has, what its higher classifications are.
- **The Biodiversity Heritage Library (BHL)**[5] is an online library that provides open access to biodiversity heritage literature from all over the world (13). The library has employed OCR and automatically recognises scientific names in order to improve access to the printed literature through full-text search.
- **iNaturalist** (60), is a citizen science project,[6] that allows amateur naturalists to upload photographs of organism encounters in the wild together with an identification and geo-location. For their mobile application, they employ image recognition to help naturalists with the identification of observed organisms.

2.3.2 Use Cases

- **Committee for Natural History of the Netherlands Indies ("Natuurkundige Commissie voor Nederlands-Indië") (NC):** The NC was founded by King William I of the United Kingdom of the Netherlands. Consisting of a group of naturalist, draftsmen and preparators[7] from the Netherlands as well as German-speaking countries

[1] http://rs.tdwg.org/dwc/
[2] https://eol.org/docs/what-is-eol
[3] https://www.iucn.nl/over-iucn-nl
[4] https://www.ncbi.nlm.nih.gov/genbank/
[5] https://www.biodiversitylibrary.org/
[6] https://www.inaturalist.org/
[7] In the field of natural history preparators are those responsible for preparing plants and dead animals so that they could be used for research. Those preparing only animals are also known as taxidermists

and France (16), the committee was sent to the Indonesian archipelago. Their primary task was the collection of information on natural resources in the Dutch Indies. In addition, they were deployed to observe and describe the local flora and fauna. As a result, many specimens, biological illustrations and observation descriptions were brought back to the Netherlands for closer investigation, with the aim to publish results on the natural diversity of the Dutch Indies.

Currently, the physical collection is stored at the Naturalis Biodiversity Center in Leiden. In 2008 the archival part of the collection was digitised (scanned through the Metamorfoze programme[1]), leading to a digitised collection of roughly 8,000 field book pages, and 2,000 illustrations.

- **The Iconografia Zoologica collection (IZ):** The Iconographia Zoologica[2] (short: IZ) is a 19th century collection of biological illustrations from the Artis Library of the University of Amsterdam. The collection was formed by three collectors: the well-known collector and naturalist Th. G. van Lidth de Jeude, the zoologist R.T. Maitland and the curator of the shell collection at the Amsterdam Zoo, Abraham Oltman, together with the Amsterdam society *Natura Artis Magistra*. In the 21st century, the collection was digitised and labelled with either complete binomial species names (genus and specific epithet) or corresponding genera. The full online collection contains over 26,500 sketches and drawings.

[1] https://www.metamorfoze.nl/kennis-onderzoek/lexicon/preservation-imaging
[2] https://bijzonderecollecties.uva.nl/gedeelde-content/beeldbanken/iconographia.html

CHAPTER 3

Manuscripts to Databases

"You who read me—are you certain you understand my language?"

– Jorge Luis Borges, *The Library of Babel*

Searching through historical manuscript collections can seem like an insurmountable task. Misreading one word can change the entire reading of a text, and even a correct reading of a historical text might not give any direct clues as to its meaning, as historical content needs to be understood in the spirit of its own time. Tying images of handwritten text to their symbolic representation (such as digital text), allows for computational exploration of the content and facilitates their correct interpretation.

In this chapter, we aim to answer research question **Q.1** (*What are the trade-offs of various system designs for the disclosure of digital archives?*).

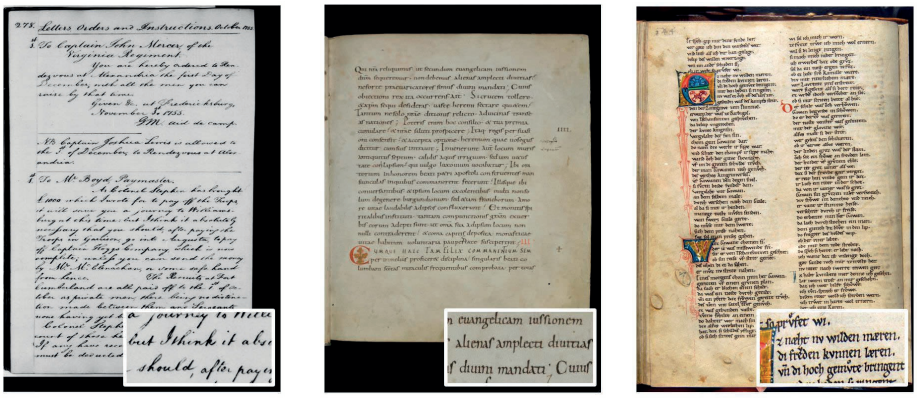
3.1 Introduction

Galleries, Libraries, Archives and Museums (GLAMs) often provide web-accessible, digitised images of historical manuscripts from various domains, e.g., medieval manuscripts,[1] natural history field books,[2] works on philosophy and jurisprudence,[3] ancient religious manuscripts,[4] notarial acts,[5] or biodiversity literature.[6]

In order to computationally access the content of text images, they can be transcribed and/or annotated by the public at large through crowdsourcing (61; 17), or by human domain experts through nichesourcing (62; 63). By utilising human-generated transcriptions,

[1] https://dlmm.library.jhu.edu/en/digital-library-of-medieval-manuscripts/
[2] https://siarchives.si.edu/about/field-book-project
[3] https://blogs.ucl.ac.uk/transcribe-bentham/
[4] https://www.deadseascrolls.org.il/about-the-project/the-digital-library
[5] https://alleamsterdamseakten.nl/
[6] https://www.biodiversitylibrary.org/

automated techniques such as HTR (23; 64) and keyword spotting (65) can further take up transcription. Figure 3.1 shows example historical manuscripts from three different datasets available from the comprehensive IAM-HistDB[1] research database:



(a) (b) (c)

Figure 3.1: Sample pages from the: (a) George Washington, (b) Saint Gall, and (c) Parzival datasets, taken from the IAM-HistDB research database.

Computational systems that produce machine-readable content from historical manuscripts, such as the ones in Figure 3.1 commonly contain three components that each digest the output from the previous component (see also Figure 3.2) (66):

Comp.1 Pre-processing of the heterogeneous content through *document image analysis* (DIA): e.g., segmentation of the heterogeneous content into page elements such as paragraphs, lines and word zones.

Comp.2 Manual or automated transcription of the segmented lines or word zones.

Comp.3 Some form of information extraction or retrieval techniques. The former often by means of natural language processing (NLP) techniques over transcribed texts.

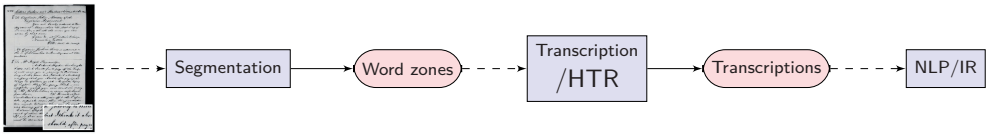


Figure 3.2: Three typical steps in historical document processing. Blue square boxes indicate processes while red rounded boxes indicate output of these processes.

[1] https://diuf.unifr.ch/main/hisdoc/iam-histdb

In this chapter we discuss various systems used in the literature for the enrichment of historical manuscripts. We divide the systems into three groups based on a set of properties that we define (Section 3.2). Based on a final discussion, we propose an approach for knowledge extraction from digital images of field books and scientific illustrations (Section 3.3).

3.2 System Designs

We analyse systems for the enrichment of manuscripts in a slightly less conventional way, for the purpose of optimising and streamlining knowledge extraction. In the literature, systems are often discussed based on types of algorithms used for **Comp.1** (related to binarisation, segmentation, text-line normalisation (66)) and for **Comp.2**, techniques for HTR and OCR and their performance on standard benchmarks (such as the ones from Figure 3.1). **Comp.3** is often looked at separately, after realisation of **Comp.1** and **Comp.2**. We focus on component **Comp.2** and **Comp.3** in conjunction, and look at three properties in specific: **agents** that aid in the transcription and annotation process, the **proportion** of the content that is transcribed, and **richness** of content descriptions:

- **Agents:** The agents that are involved in the process of digitisation of the text: (1) the public at large, (2) the expert community, (3) a machine.
- **Proportion:** The proportion of text that is transcribed, whether it is attempting full verbatim transcription or retrieval of keywords, in which each step includes the previous step(s): (1) named entities, (2) keywords (3) full text.
- **Richness:** The level of richness with which the content is described, in which each step assumes employment of the previous step(s): (1) verbatim, (2) locally defined semantic tags, (3) terms from controlled vocabulary or schemas, (4) IRIs, (5) terms from an ontology.

We define a set of terms in the context of manuscript enrichment, as the terminology may vary between studies:

- **Transcription:** The digital representation of a written text. *Transcribing* in this context is the act of transforming the verbatim handwritten text in a digital image of a manuscript to digital text.
- **Label:** The representation of a region of interest (ROI) in a digital image as digital text. *Labelling* in this context is the act of "attaching" a digital label to a ROI using some computational system. The ROI together with its representation as digital text can be used as training data for machine learning.

- **Annotation:** digital or written notes or comments added to an image or digital text; they point to a specific ROI (for images) or range (for digital text), and add comments or metadata such as a free text description or a semantic type (*semantic annotation*).
- **Keyword:** A word that is key in describing the content of a document, such as a word that would be used to search a set of documents using a search engine.
- **Named entity**: "Information units like names, including person, organization and location names, and numeric expressions including time, date, money and percent expressions (30)."

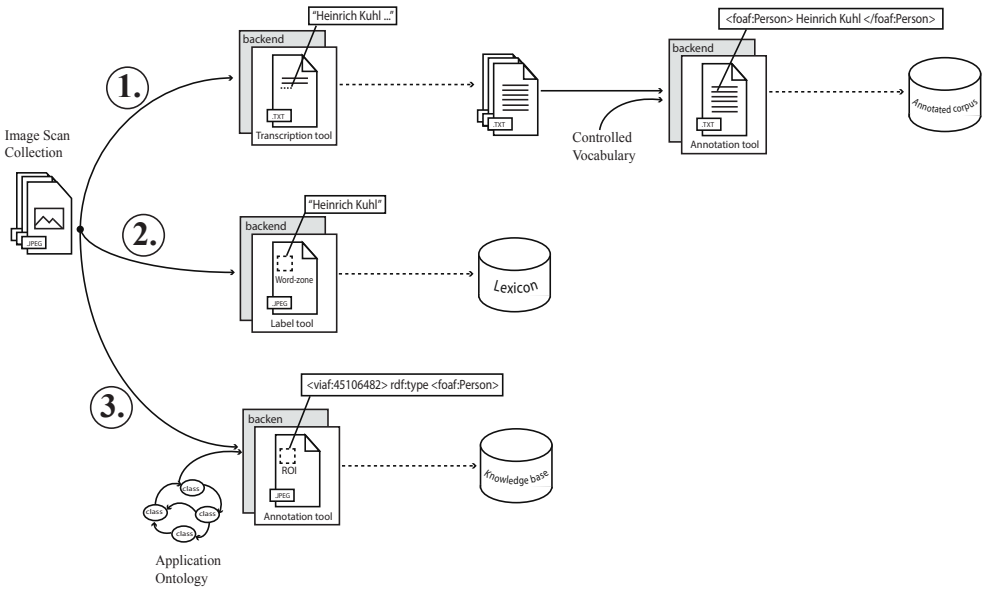


Figure 3.3: Manuscript enrichment design patterns.

We discuss our predefined properties for three types of systems: (1) manual full-text transcription, (2) semi-automated transcription, and (3) semantic annotation of text images, which we graphically represent in Figure 3.3. In total, we discuss a selection of 10 systems that in our opinion represents the breadth of the literature well. In the coming subsections we describe each system type, for which we discuss example frameworks and projects.

3.2.1 Manual Full-Text Transcription.

GLAMs around the world are beginning to notice the potential of crowdsourcing for full-text transcription (see Figure 3.3, system type (1)). In crowd- or nichesourcing, scholars,

experts or the public at large, take on the task of digitisation of the verbatim content of manuscripts (17; 61; 62). Many examples of initiatives now exist that manually transcribe manuscripts in full. We discuss three examples below, specifically including two that digitise handwritten field books:

- The *Field Book Project*[1] (15) is a project set up by the Smithsonian Institution Archives in collaboration with the National Museum of Natural History. The project uses a crowd of what they call "volunpeers"[2] to harvest full-text transcriptions from field books (67), through their transcription center.[3] Controlled vocabularies such as the Natural Collections Description (NCD) are used to describe metadata on levels above content-level (see Figure 1.1). Their approach, described in (68), mentions the use of geo-tagging for future work, to disambiguate localities.

- The *Transcribe Bentham* initiative[4] has digitised and, through crowdsourcing, successfully transcribed 24,833 (update: 27th of November 2020) manuscript pages from jurist Jeremy Bentham (1748-1832), stored in the University College London digital archive, through a customised version of the MediaWiki[5] transcription interface (17; 61).[6] Manuscripts are transcribed, and transcriptions are marked-up with Text Encoding Initiative (TEI)[7]-compliant XML. They indicate a survey pointed out most volunteers took an interest in the history and life of *Bentham*, and that reasons which kept volunteers from transcribing were difficulties deciphering the hand of Bentham. Within another project, *tranScriptorium*[8] (22), the transcriptions are used to further transcribe the manuscripts using HTR techniques.

- The project *From Documents to Datasets* (35) provides a design for the conversion from digitised handwritten field books to datasets, see Figure 4.2, structured according to terms from the DwC standard. They propose first to fully transcribe the texts together with experts, then upload those texts together with the image scans to a MediaWiki[9] server. Via templates, the *taxa*, *locations* and *dates*, are annotated by researchers through a crowdsourcing initiative. Annotators can resolve verbatim names to current ones (taxonomic referencing) during the semantic annotation process. The

[1] https://siarchives.si.edu/about/field-book-project

[2] A combination of the word *volunteer* and *peer*. The term is coined by Meghan Ferriter of the Smithsonian Transcription Center, and is used to refer to a skilled volunteer working at a professional level. https://siarchives.si.edu/blog/growing-community-volunpeers-communication-discovery

[3] https://transcription.si.edu/

[4] https://blogs.ucl.ac.uk/transcribe-bentham/

[5] https://www.mediawiki.org/wiki/MediaWiki

[6] https://blogs.ucl.ac.uk/transcribe-bentham/

[7] TEI is a standard for the representation of texts in digital form, in order to represent structure and content of the text, such as page layout and physical properties https://tei-c.org/

[8] http://transcriptorium.eu/

[9] https://wikisource.org/

annotations are then extracted and converted manually to DwC terms, in order to publish them to the GBIF [1] data server (69).

Agents. Full-text transcription offers a good solution for GLAMs aiming to digitise their manuscript collections, but we note that manuscripts with heterogeneous hard-to-read historical handwriting and content can be too challenging to transcribe by the public at large (**Chall.6**, in line with **Chall.7**). Multiple crowdsourcing techniques exist that secure data quality,[2] but motivation can drop when tasks are too challenging. Although transcription projects often mention they leverage the crowd, most valuable effort appears to come from the community (domain enthusiasts, volunpeers, domain experts, citizen scientists, amateur experts). Transcription and annotation of heterogeneous, multilingual, hard-to-read manuscripts is a knowledge-intensive task, and (amateur) experts have more domain knowledge to perform the tasks, and are intrinsically motivated to produce high-quality data (62). In this sense we note that the term crowdsourcing is an ambiguous one, as there is a significant distinction between the public at large, and the smaller community crowd. We therefore prefer to use the term *nichesourcing* (coined by de Boer et al. (62)) to refer to the act of leveraging a smaller "crowd" of domain (amateur) experts for such knowledge-intensive tasks.

Proportion. As the term *full*-text transcription suggests, the aim of most crowd- or nichesourcing efforts through transcription tools aim at transcribing a text in full. One thing to note is that full-text transcription is time-consuming, and success depends on many factors, such as the complexity of the material and the involvement (motivation) of the community crowd. Full-text transcription can mitigates semantic enrichment, since the manipulation of digital text is computationally more straightforward than the manipulation of text images. However, much of the digitised textual content serves human comprehension, the "glue" that connects the truly interesting pieces of information, and is often not used as search terms.

Richness. Although some systems discussed above employ some form of semantic enrichment (richness level 2), most transcription systems in the literature, however, produce unstructured or semi-structured—usually based on syntax rather than semantics—XML files. This is useful for further searching and processing (e.g. using text mining techniques), but does not enable content to be semantically queried, or integrated with other collections.

[1] http://www.gbif.org/

[2] http://manuscripttranscription.blogspot.com/2012/03/quality-control-for-crowdsourced.html

3.2.2 Semi-Automated Transcription

Transcription can be partly taken on by HTR techniques (see Figure 3.3, system type (2)). Human experts take on the task of labelling segmented lines or word zones (ROIs containing written words), which are in turn used to automatically increase searchability of other parts of the text. An increase in human-generated transcriptions invokes an increase in the ability of HTR and word spotting techniques to accurately transcribe words in other parts of the texts. Common techniques include supervised deep learning methods such as BLSTMs for classification of characters, full words or sentences, or clustering techniques such as keyword spotting, where "clouds" of visually similar word zones are labelled by experts, rather than single word zones. In our discussion we omit systems that employ OCR, as the content of historical manuscripts is too heterogeneous (see **Chall.6**) for OCR to produce any usable results.

- The *HisDoc* project[1] is an example of a HTR system: experts transcribe individual text lines, and these are used as input to a supervised learning system that aims to learn models for single characters (64). As their system performs HTR at word level, a lexicon (a set of valid words) is required for automated transcription. As an alternative, they experiment with lexicon-free word spotting techniques (65). In the literature, keyword spotting is referred to as a *recognition-free* approach (70): word images are matched to visually similar images, often through a form of clustering of word images in a feature space (71). In order to deal with name variants and misspellings, they define word confusion candidates as synonyms (72).

- A 17th-century botanical manuscript "*Historia de las plantas*" has been digitised (73), using the the Computer Assisted Transcription of Text Images (CATTI) framework (74; 75). The framework performs layout analysis and allows users to transcribe the extracted line segments. The framework also offers HTR technology as an "assistant" that helps users transcribe the text. The HTR technology is based on Hidden Markov Models (HMMs) that operate on single characters, and language models that use as input N-grams. Toselli et al. (73) indicate that the CATTI system primarily aims at producing high-quality professional manuscripts, but indicate that potentially, the crowd could be leveraged, as was done in the *Transcribe Bentham* project.

- *Transkribus* is a platform developed for the enrichment and searching of historical documents (76). A user can transcribe sentences which are then used for training using HTR (21). Similarly to the HisDoc project, Transcribus uses keyword spotting techniques that allow users to search the texts. The project implements a form of

[1] https://diuf.unifr.ch/main/hisdoc/

semantic enrichment: users can use locally defined, user-created semantic tags to label transcriptions or segments.

- The *MONK* system is a search engine for processing multilingual, multi-script historical text, developed by Schomaker (23). It implements HTR as a function for word retrieval. The goal of MONK is therefore not necessarily full-text transcription, but rather to create a searchable index (77). The system has already processed many documents, amongst which the *Dead Sea Scrolls*;[1] Hebrew manuscripts encountered in the Qumran Caves near the Dead Sea.

Agents. Machines, through HTR techniques, can take part in a transcription effort, but have trouble transcribing content that is too heterogeneous (see **Chall.6**), as good results rely on many human-labelled examples. Character-based methods rely on language models and are therefore dependant on a statistical language model or lexicon, whereas an object recognition approach that looks at whole words (such as the one taken by MONK, or word spotting techniques) has to deal with **Chall.8**, as interesting words lie in the long tail of the word distribution. Historical handwriting recognition is far from solved (23), and especially for heterogeneous content, often produces poor results that are difficult to interpret.

Proportion. It appears that, for many HTR systems and their users, the eventual goal is full-text transcription of complete manuscript collections. Other systems aim at creating a searchable index, which does not necessarily require all content to be transcribed. Ultimately, the process is never linear for HTR systems: more transcriptions lead to an increasing number of accurately recognised words. A partly transcribed collection can also be published online as a "living" document of which the proportion of machine-readable content continues to grow.

Richness. The main goal of HTR systems is verbatim transcription (richness level 1), although some allow for semantic enrichment, often no further than richness level 2. It is worthwhile to note that automated tasks such as NERC that further enrich the verbatim content to capture any implicit semantics commonly rely on NLP, a technique that relies on the context of words rather than words in isolation, and therefore depends on the transcription of that context. Although full-text transcription is not required to make a text searchable (not many scholars would be interested to find all instances of the word "*the*" in a collection), we do argue that *undirected* (as in: unguided by formalisms) word-zone

[1] https://www.deadseascrolls.org.il/about-the-project/the-digital-library

labelling or keyword spotting limits or hampers automated extraction of any semantics before manuscripts are fully transcribed.

3.2.3 Semantic Annotation of Text Images.

GLAMs make increasing use of Semantic Web technologies to enrich and publish their collection items (78; 79; 80). Several systems on the web aim for semantic annotation of textual resources (31; 81), but digitised manuscripts are not often enriched in the same way. There are, however, a couple of example systems that directly annotate text images with semantic concepts. Similarly to word-zone labelling, scholars, experts or the public at large can be employed to semantically annotate online documents (see Figure 3.3, system type (4)).

- *Accurator*[1] is an example of a web application that uses an expert crowd to annotate digital images, in specific digitised items from cultural heritage collections, such as paintings. Web users can help museums describe their collection items by providing expert knowledge. They are prompted to annotate digital renditions of items from cultural heritage collections with terms from controlled vocabularies, carefully selected for the target domain of the collection. For each collection, experts were even involved in the process of determining a goal for proper enrichment, in order to improve access to the collection in question. Annotations are stored in RDF format and linked to the digital images using the Web Annotation Vocabulary[2] (82).

- Ebert et al. (2010) (83) perform ontology-based information extraction (OBIE) from handwritten documents. They are one of the first ones to introduce the topic to the field of HTR. Interestingly, their system employs a dialogue between a component that deals with HTR and a OBIE component. Their system is based on digital ink as input (using the MyScript[3] system for HTR) and the scope of their experiments is homogeneous handwriting (they experiment with modern English handwritten texts) rather than the heterogeneous material from historical manuscript collections, which additionally needs to deal with historical multilingual text (**Chall.6**).

- Adak at al. (2016) (84) perform named entity recognition (NER) on unstructured handwritten text images, without employing any character or word recogniser. After word segmentation, they extract engineered structural and positional features from word zones, which are used in a BLSTM for NER. Classification of the named entities is out of the scope of their paper. The methodology presented in the paper does not

[1] http://www.accurator.nl/
[2] https://www.w3.org/TR/annotation-vocab/
[3] https://developer.myscript.com/docs/concepts/introduction/

increase searchability of the text, but can be combined with a controlled vocabulary for NERC to automatically enrich the handwritten content semantically. We therefore included it in this section. The article presents a nice overview of how relevant page elements such as named entities can be identified in text images with hard-to-read historical texts.

Agents Semantic annotation of texts is a more knowledge-intensive task than mere verbatim transcription of a text, as a level of interpretation is required. Therefore, human (amateur) experts are required to take part in the annotation process. Additionally, quite some time is spent selecting or re-engineering vocabularies or ontologies to fit the target domain. However, an application ontology formalises the minimal information required for annotation, thereby driving the enrichment process. Moreover, machines can take part in the semantic annotation process, as is shown by Adak et al. (84).

Proportion The systems mentioned above operate on text (or multimodal) images, and focus on the annotation of information units, such as named entities, rather than just any word or full text. Prior to the annotation effort, the expert community decides on interesting concepts and their meanings, and use these to semantically enrich ROIs through a nichesourcing initiative, which users eventually use to navigate and understand the resulting knowledge base, and join distributed collections.

Richness At a minimum, semantic annotation systems annotate texts or text images with semantic concepts, for instance through a combination of supervised HTR and NERC from features of the handwritten text (85; 84; 86) (richness level 2). Examples exist that even use terms from controlled vocabularies or schemas (richness level 3), or that use HTTP URIs for better content descriptions (richness level 4) (82; 31; 81).

3.3 More Product, Less Process

Coming back to **Q.1** (*What are the trade-offs of various system designs for the disclosure of digital archives?*), we note that the enrichment of manuscripts is often a highly time-consuming process that depends on community engagement. This is no different for field book manuscripts, which are exceptionally challenging to make sense of, given **Chall.1** to **Chall.5**.

At the same time, if we look back at **Chall.6** to **Chall.8**, we note that it seems unavoidable that humans play a large part in the enrichment process, although machines can be employed

to speed up this process, given that their results are presented in a transparent, human-understandable way. Systems with high recall but low precision[1] increase retrievability of words, but results can clutter the enrichment process when not presented well. Moreover, unless character-based out-of-lexicon methods are employed, words that occur more often are the first to be recognised accurately, while they are more likely to be less relevant. A third thing to note is that enrichment efforts often result in unstructured or syntactically structured digital text, that require a crucial enrichment step in order to be understood and reused by scholars and the general public.

We have observed in Subsection 2.1.2 and systems discussed in the previous section, that the content in manuscripts from NHCs is organised around a systematic regularity that is intrinsic to the field of biodiversity, in which researchers attempt to systematise the natural world. This systematic organisation is not commonly encountered in other manuscripts. At the same time, community standards are set up to formalise these systematics.[2] In terms of efficiency; should "volunpeers" not maximise their impact by focussing not only on transcription, but also on systematics, using standard formalisms from the domain?

Greene et al. (87) already noted in 2005 in their article *More Product, Less Process* that there exists a huge backlog of unprocessed archival material (for the most part the authors refer to cataloguing of archives on a collection- and item-level for minimal collection access, but we argue that the same concerns apply to enrichment of and access to archival content). They mention that processing of archival material should: *"describe materials sufficient to promote use."* To strengthen their argument, they quote an article already published three decades ago on the same topic:

> We rarely ask the question: when is *this* collection processed? Instead, we process all collections to an ideal standard level. The second problem is that by processing all collections to the ideal standard level, we cannot keep up with the collections we have on hand or with the new collections coming in. The result tends to be a small number of beautifully processed collections available for use and an extensive backlog of collections that are closed while they wait to be processed (88).

This idea is in line with the idea of Minimum Information about a Digital Specimen (MIDS)[3] from the Collection Descriptions (CD) interest group, on the formalisation of sufficient digitisation:

> A harmonizing framework captured as a TDWG standard can help clarify levels (depth) of digitization and the minimum information captured and published at each level.This would help to ensure that enough data are captured, curated and published against specific requirements so they are useful for the widest range of

[1] recall refers to the percentage of *all* words that is correctly retrieved, while precision refers to the percentage of words that is correctly retrieved from all *retrieved* words.

[2] `https://www.tdwg.org/`

[3] `https://www.tdwg.org/community/cd/mids/`

> possible purposes; as well as making it easier to consistently measure the extent of digitization achieved over time and to set priorities for remaining work (89).

We extend these ideas to the digitisation of manuscript *content*. We claim that at a minimum, information extraction from manuscripts should promote document understanding, rather than full-text transcription of each manuscript to an ideal level.

We therefore opt for a targeted approach, in which the expert community decides the semantic concepts relevant for document understanding and search, maps these to existing ontologies and IRIs, and uses these to guide the annotation effort by semantically annotating and transcribing the relevant word zones in text images through a nichesourcing initiative. Texts are made searchable, pointing users to interesting bits of the text documents, while ground truth is generated for semi-automated semantic annotation (similar to NERC) as well as verbatim transcription. In an end-to-end approach, a named entity recogniser can then benefit from output of the handwriting recogniser, and vice versa.

Although some extra work is required to semantically annotate texts with Linked Data (LD), omitting full-text transcription means having to annotate only a small percentage of the content; e.g., focussing on the transcription and semantic annotation of those *named entities* that allow users to construct rich semantic queries or aggregate informative content across archival collections.

Pre-populating knowledge bases with background knowledge, such as collection-specific locations from the Geonames database or collection-specific persons from the Virtual International Authority File (VIAF) authority IRIs, helps annotators to use the correct named entities for annotation. Using LD for annotation helps remove ambiguity as IRIs contain rich descriptions. The name "*Heinrich Kuhl*", for instance, is ambiguous. If we instead use the IRI `https://viaf.org/viaf/45106482/`, we agree on the reference of the verbatim name to the person "*Heinrich Kuhl*" (1797-1821), a German zoologist.

Lastly we argue that annotation *provenance* is a dimension that is often overlooked, but should be seen as a critical step in the elucidation process. With data provenance we refer to data concerning the lineage of data: why, when, and how they were produced or changed, and measures of their quality (90; 91; 92). Storing provenance of annotations contributes to publishing annotation knowledge graphs in a FAIR way, allowing scholarly discussions over the content and reproducibility of hypotheses and results.

CHAPTER 4

Semantic Annotation

> *"There is no true interpretation of anything; interpretation is a vehicle in the service of human comprehension. The value of interpretation is in enabling others to fruitfully think about an idea."*
>
> – Andreas Buja, *as quoted in: The Elements of Statistical Learning*

Semantics concerns itself with meaning, or reference. David Lewis, a famous American philosopher of the twentieth century, wrote on the topic of semantics the following:

> I distinguish two topics: first, the description of possible languages or grammars as abstract semantic systems whereby symbols are associated with aspects of the world; and, second, the description of the psychological and sociological facts whereby a particular one of these abstract semantic systems is the one used by a person or population (93).

In Section 2.1, we have discussed the languages and grammars used in historical and contemporary species research, in the light of challenges **Chall.2** and **Chall.4**. In this chapter, guided by domain experts, we extract references hidden in historical field books (*implicit semantics*). We discuss how we can use machines to make these implicit semantics accessible to researchers, allowing for scholarly discussions over the content, through a process called *semantic annotation*.

Specifically, this chapter aims to answer two research questions **Q.2** (*What types of research questions do domain experts formulate regarding the archival content of NHCs, and how can we make the content machine-readable to facilitate such queries?*) and **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?*).

4.1 Introduction

We have established in earlier chapters that interpretation of field observation records is challenging, even for domain experts (see challenges **Chall.1** to **Chall.5**). Ideas should

therefore be developed for the use of computational processes to disclose collection content and semantics in a transparent way. Doing so ensures that interpretations of field book content not only exist in inaccessible ledgers or text files of individual researchers, but also somewhere accessible and understandable by the public at large, biodiversity researchers as well as those studying natural and cultural history.

Through the emergence of digitisation projects (8; 15), new possibilities arise to disclose hand-written manuscript collections with digital tools. Some initiatives, such as the *Field Book Project* (discussed in Chapter 3), use manual full-text transcription to make collections available to the general public. In this chapter we propose to disclose archives, in the domain of natural history, through *semantic annotation* of the content. Many definitions exist but we take it to be the process of producing structured annotations from the named entities in texts. These named entities form the general semantics of these texts. Coupling them with background knowledge, and linking them through formal descriptions, provides connectivity throughout the documents (31).

Work has already been done linking collections on a *collection*- and *item*-level using controlled vocabularies (see Table 1.1), the principles of Linked Data, and/or ontologies, not only regarding biodiversity collections (13; 68), but cultural heritage (CH) collections in general (94; 95; 96; 79; 97; 98; 99). This is also the case for collections of manuscripts, but fewer examples exist that semantically link the multimodal field observations on a *content*-level. Such an approach would facilitate content aggregation as well as the use of structured queries and reasoning over the content, and, through the use of IRIs, disambiguation of named entities, which is crucial in the field of biodiversity. Therefore, this chapter makes the following contributions to the field:

1. We provide a semantic model, an application ontology written in OWL,[1] to structure drawing captions and historical occurrence records in field books. Relevant concepts were defined by domain experts, and modelled by integrating ontologies developed for the biodiversity domain, a geographical database, and for annotation provenance.

2. We present a semantic annotation tool, the SFB-Annotator, which uses the application ontology, and enables domain experts to produce structured annotations from digitised natural history archival collections using the ontology. In addition, the tool documents the provenance of annotations.

3. We provide the results of a qualitative evaluation of the proposed model and annotation process. The annotations will subsequently inform the development of an

[1] `https://www.w3.org/OWL/`

adaptive learning approach leading to semi-automated annotation, which we discuss in Chapter 5.

We show the applicability of the ontology and annotation system on a selection of field notes from the digitised NC collection (mentioned in Subsection 2.3.2), which contains approximately 8,000 field note scans.

This chapter is structured as follows: in Section 4.2 we discuss the model development method and process, Section 4.3 describes the semantic annotation approach using the model, and in Section 4.4 we evaluate the approach qualitatively and discusses annotation data acquired from semantically annotating a collection of field book pages from the use NC use case. Lastly we discuss results, describe limitations and outline future work in Section 4.5.

4.2 Development of a Semantic Model

The development process for the semantic model followed the ontology development process described by Fernández et al (100). The emphasis in the development process of our model was on the re-use and re-engineering of existing semantic models. We thus followed the ontology development process as outlined in scenario 4 of the NeOn methodology for ontology engineering (101). Furthermore, we support a user-centered design, where the focus is on the needs of the end user, similar to a method for database design described by Gray (102), where questions of domain experts become requirements for the design and evaluation of the system.

4.2.1 Requirements

The requirements for the semantic model describe user requirements for elucidating content from text images, and requirements for adhering to the principles of sharing data in the Semantic Web.

Elucidating Content

R.1 The model should formalise the general semantics of species observations described in field books and illustrations.

- **(a)** The model should include the named entities that domain experts use when constructing queries in order to answer their research questions.
- **(b)** The model should reveal relations between the named entities and their characteristics, for instance, hierarchical or transitive relations, so that these can

be exploited in rich content queries. The model should thus be written in an ontology language such as the recommended W3C[1] standard language, OWL.

R.2 The model should be able to deal with variants of terms and their context. Examples are historical terms, synonyms and homonyms, scientific names and their vernacular names, and abbreviations.

(a) Standardised terms for resources, such as IRIs, should be used to represent named entities so that name variants can be linked and dissimilar entities with a similar name can be disambiguated.

(b) The context of name variants should be made explicit so that name variants are understandable in their context, for domain experts as well as automated reasoners.

Serving Structured Annotations to the Semantic Web

R.3 The model should re-use existing ontologies and vocabularies to facilitate data aggregation on the web.

R.4 The model should store annotation provenance to enable the sources of annotations to be traced and to facilitate scientific discourse over the content.

(a) The annotations should track metadata regarding the annotation process; annotator, date/time, and interpretation.

(b) The annotations should store metadata regarding their span in text images: multiple pages, single pages or fragments from pages, to keep track of the provenance of annotations in relation to the collection. Linking image fragments to their annotations and annotation metadata can be used in further research for salient named entity recognition and classification (SNERC), and facilitates repetition of experiments by other researchers.

4.2.2 Semantics for Biodiversity

Below we discuss available state-of-the-art standards and ontologies regarding semantics for biodiversity.

[1]The World Wide Web Consortium (W3C) is an international community for the development of standards on the Web. https://www.w3.org/Consortium/

The Darwin Core. The biodiversity data standard that is most commonly used to model species occurrences is the DwC standard (36). It has been developed through community consensus and thus describes which concepts in observation records are most important to the community. The DwC describes these key concepts with standardised terms. Its main classes are: `dwc:Organism`, `dwc:Taxon`, `dwc:Identification`, `dwc:Occurrence` and `dwc:Event`. The standard therefore satisfies **R.1**, and thus proves to be a suitable baseline for our model.

For our purpose, the DwC alone does not suffice. Firstly, the DwC does not satisfy **R.1b**. Although the terms from the DwC were converted to be used with RDF (103), the standard does not allow all properties to be used within its `dwciri:` namespace, adopted to refer to IRIs (103). This means that not all relations can be used to point to IRIs, hindering the linking of entities from handwritten observation records during an annotation effort. The current standard lacks properties to interconnect its main classes and does not exceed the semantics of RDF Schema. This means it does not include types of properties and property axioms that we require, such as equivalence and transitivity.

Moreover, the DwC does not model taxonomies explicitly, so reasoning algorithms cannot benefit from their inherently hierarchical nature. It models classification systems by connecting a taxon identifier to a literal through a rank property, e.g.,: `nc:taxon1` *dwc:order* "*Chiroptera*". Finally, the DwC's use of literals for named entities does not fulfill our requirements. As literals are multi-interpretable, they do not serve as unique identifiers within RDF. In the field of biological taxonomy, and especially historical taxonomy, where multiple interpretations of species and naming conventions exist, being able to disambiguate between terms with the same name is crucial (29). In these respects, the DwC does non satisfy **R.2a** and **R.2b**.

The Darwin Core Semantic Web. The Darwin Core Semantic Web (DSW)[1] ontology extends the DwC by providing properties to link the main classes of the DwC (104). It hereby addresses the limitations of the DwC regarding **R.1b**. The DSW also introduces a new class, the `dsw:Token` class, to link the graphical model to evidence in the form of a `dwc:Specimen`, `dwc:HumanObservation` or other class on which the identification of an organism during an occurrence event is based. However, the DSW ontology does not allow biological taxonomies to be graphically modelled, a requirement that is included in **R.1b**. Finally, to the extent of our knowledge, the applicability of the DSW ontology has not yet been demonstrated on large datasets.

[1] `https://github.com/darwin-sw/dsw`

TaxMeOn. The TaxMeOn[1] Meta-Ontology of Biological Names is an ontology that models biological taxonomies (105). The ontology uses IRIs for taxa and introduces hierarchy by connecting the taxa to each other using the transitive $\overrightarrow{isPartOfHigherTaxon}$ property. This property is made transitive so that logically inferred, the scientific name is not only a part of its own higher taxon, but all higher taxa. This way of modelling classification systems is suitable for our purpose: taxa can be linked during the annotation process, recreating the historical taxonomy and allowing subsequent querying of the archive for all species from a certain class or order. Moreover, the instances are modelled as IRI, avoiding name ambiguity. Its conceptualisation, however, is subtly different than the DSW ontology: TaxMeOn models taxa as instances of a rank class such as genus whereas the DSW ontology only models taxa as instances of the class `dwc:Taxon`.

In summary, present-day biodiversity records can be described using terms from the DwC and the DSW, but some alterations need to be considered for the description of NHCs. Domain experts' interests were explored to complement the existing vocabularies to satisfy (**R.1a**) and to address **R.1b**, the DSW ontology was re-structured so that the biological taxonomies could be modelled based on the structure of the TaxMeOn ontology. Furthermore, the terms in the field books were linked to standardised terms from other datasets. This accommodates the linking of different spellings and abbreviations (**R.2a**), the inclusion of context metadata (**R.2b**) and enables data aggregation on the web (**R.3**). Finally, the storage of provenance metadata of annotations (**R.4**) was addressed. The modelling process is explained in the coming subsections.

4.2.3 Data Elucidation by Domain Experts

To inform the design process, the interests of domain experts were assessed via qualitative interviews and a test annotation procedure, addressing **R.1a**.

Seven domain experts participated in the interviews that were set up to acquire knowledge about interesting concepts in field books; two cultural historians, two information specialists handling collection queries from within the Naturalis Biodiversity Center (NBC) and three biologists interested in taxonomy and the history of biodiversity. A subset of 59 pages from our use case was selected for inspection. These pages contained all species descriptions within the collection belonging to the order *Chiroptera*, an order of mammals that consists of the bats. The subset consisted of 40 pages of observation descriptions and 19 drawings.

[1] `http://schema.onki.fi/taxmeon/`

First, participants were asked to describe their research interests and denote research questions they would like to address with access to a natural history archive. Examples included "*Are the species named directly in the field or do they receive a number or a temporary name?*" and "*Did specific naturalists have a specialisation, such as the description of plants?*". Subsequently, they were asked to note down conceptual elements they would expect to find in historical observation records that would help them answer their research questions. Being primed to think in concepts, they were asked to use these concepts to annotate the field book pages and depictions with a digital tool, to allow the addition of new concepts to the semantic model should these be discovered during the annotation process.

Table 4.1: Conceptual elements domain experts expected to find in observation records, organised by topic. Similar concepts were merged, e.g., *Linnean Name* and *Species Name*. The number *c* indicates how often the concept was used for annotation of the field note subset, accumulated for all participants, and the number **n-7** indicates that n of the 7 participants used the concept for annotation.

Topic	**Annotated Concepts**	*c*, (**n-7**)
Classification	Linnean name: *30*, (**7-7**) Literature used: *2*, (**2-7**) New namings: *3*, (**2-7**) Additional class.: *6*, (**4-7**)	Vernacular name: *2*, (**2-7**) Synonyms: *6*, (**4-7**)
Species	Rarity: *5*, (**2-7**) Range: *5*, (**2-7**)	Use by locals: *0*
Expedition	Person: *23*, (**7-7**) ◆ Collector: *2*, (**1-7**) ◆ Author: *6*, (**2-7**) ◆ Companion: *0* ◆ Local person: *0* ◆ Illustrator: *5*, (**3-7**) Observation place: *22*, (**7-7**)	Role of indigenous population in knowledge retrieval: *0* Collection practices: *2*, (**2-7**) Drawing property: *5*, (**3-7**) Language peculiarity: *0* Observation date: *10*, (**7-7**) Publication: *0*
Organism	Link to specimen: *1*, (**1-7**) Drawing *17*, (**7-7**) ◆ parts *7*, (**2-7**) ◆ views *4*, (**3-7**) Preservation *0* Measurement: *5*, (**5-7**) Quality: *14*, (**7-7**) ◆ Colour: *2*, (**2-7**) ◆ Behaviour: *8*, (**2-7**) ◆ Morphology: *5*, (**5-7**)	Link to Drawing: *2*, (**1-7**) Condition: *0* ◆ Living: *0* ◆ Dead: *0* Anatomy: *40*, (**7-7**) Gender: *1*, (**1-7**) Count: *1*, (**1-7**) ◆ Specimen *0* ◆ Anatomy term: *1*, (**1-7**)

Table 4.1 lists the concepts that were identified by the domain experts, followed by a number *c* indicating how often the concept was used for annotation of the subset, accumulated for all participants, and a number **n-7** indicating how many of the 7 participants used

the concept for annotation. If a more specific subclass was used for annotation, it was included in the count for both the general class as well as the more specific class. They can be broadly divided into concepts relating to species classifications, their abundance and use, expedition details and characteristics of the observed organism.

Within our experiment, cultural historians appeared most interested in expedition practices, more than in the specimens or species described. During the annotation process, they were searching for clues in the text as to why certain languages were used interchangeably, in what ways knowledge was recorded, which indigenous people were helping to find new species, what methods naturalists used to find and gather the specimens or what adjectives were used to describe the behaviour or appearance of organisms. The biologists appeared to be more interested in classification systems, naming conventions, species characteristics and literature used for classification. The output from the interviews and annotation procedure was used to aid the design process of the semantic model. The questions from domain experts were used to test the output of the annotated field book in Subsection 4.2.4.

The most important named entities from table 4.1 which were extensively annotated by the experts in the field books, but which are not included in the DSW ontology, are dates, additional classifications (synonyms and later classifications), additional occurrences (species range and rarity), and structured organism descriptions (anatomical parts, qualities and measurements). We thus adopt these in the final model.

4.2.4 The NHC-Ontology

In this section we explain further design choices for the natural history collection (NHC)-Ontology (NHC-Ontology[1]) and describe the adoption and application of the classes and properties. The ontology extends the DSW ontology with two classes and seven properties in order to address the remaining limitations mentioned in Subsection 4.2.2. Figure 4.1 provides a graphical overview of the model. Two classes and all new properties are added within our own namespace, indicated by the dashed lines and the nhc: namespace.

Classifications and Taxonomies. The class `nhc:TaxonRank` connects to the DSW model. All taxa are modelled as instances of the class `dwc:Taxon` and all taxon ranks as instances of the class `nhc:TaxonRank`. We adopt a derivative of the DwC property *dwc:taxonRank*, see figure 4.1. As the DwC standard does not have an analogous property in the `dwciri:` namespace, we adopt it in our namespace. To represent hierarchy in the classification system we created the transitive property *nhc:belongsToTaxon* to link a

[1] `http://www.makingsense.liacs.nl/rdf/nhc/`,`https://github.com/lisestork/nhc-ontology/`

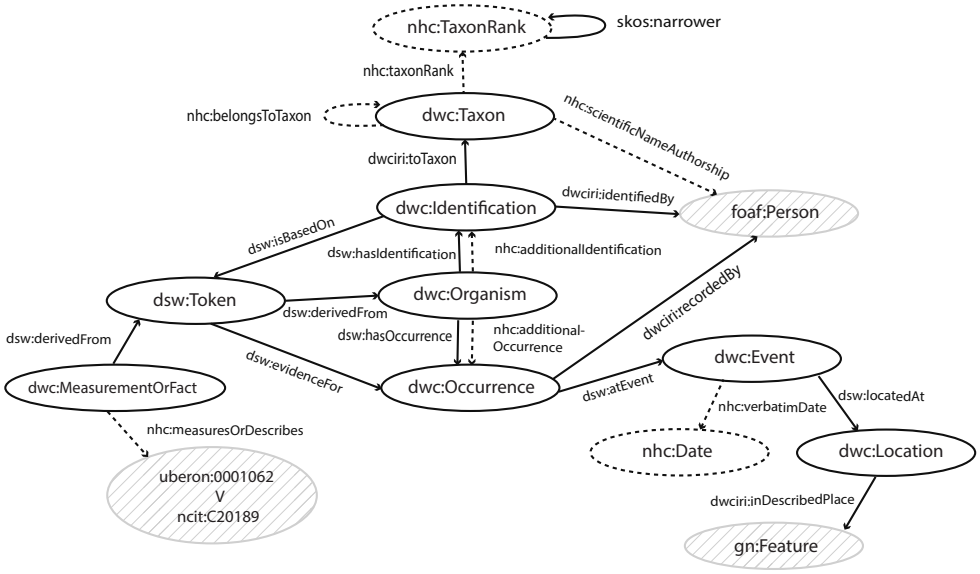


Figure 4.1: The NHC-Ontology, an extension of the DSW ontology for annotating NHCs. Gray striped classes indicate classes from external ontologies, whereas classes and properties with a dotted line pattern indicate additions to the DSW ontology.

taxon to a taxon higher in rank. Because of this transitive property we can, for example, query a collection for all families belonging to a specific order, e.g., "*Show me all families that belong to the order Chiroptera*".

In the semantic model, we model a scientific name (discussed in Subsection 2.1.1) as a single unit representing a species.[1] The author of the scientific name is linked separately, as domain experts indicated they have special interest in retrieving authors and their scientific names. For instance, all taxonomic names from a specific author to obtain knowledge concerning which species they named and to establish personal naming conventions. To link the publisher to the scientific name, we use the DwC term *scientificNameAuthorship* which we also adopt in our namespace as it does not yet have an equivalent in the `dwciri:` namespace.

When writing up observation records in field books, authors sometimes use the term "*Nobis*", Latin for "*by us*", or any other place holder for the name of the scientific publisher, as discussed in Subsection 2.1.2. "*Nobis*" in this case refers to a scientific author name, namely the writers of the field book. Annotating the region with the class `foaf:Person`,

[1] Exceptions where a genus is modelled individually are field book pages that describe characteristics of a specific genus without mentioning a species.

and linking it to the taxon with the property *nhc:scientificNameAuthorship* is useful, as placeholders can be matched with the names of the authors of the field book, allowing the taxonomic names to be resolved.

Evidence for Identification. In the DSW ontology, the class `dwc:Token` is used to link an identification to the resource on which the identification was based. This class can be replaced with the more specific `dwc:PreservedSpecimen` or `dwc:HumanObservation` class. The human observation represents a single observation record from a field book or a drawing. Therefore, we let an instance of the `dwc:HumanObservation` class point to multiple field book pages describing one record. This way, users can retrieve observation records, drawings and specimen relating to their research interests, e.g., "*show me all observations recorded on Java*".

As domain experts were interested in the measurements used for classification of an organism, as is visible in Table 4.1, we adopt the `dwc:MeasurementOrFact` class in the ontology, a class taken from the DwC standard. The `dwc:MeasurementOrFact` class is connected to the `dwc:Token` class with the *dsw:derivedFrom* property or its inverse *dsw:hasDerivative* to indicate that it is derived from, or a part of, the observation record, see Figure 4.1. As the *dsw:derivedFrom* property is transitive, the measurement is also derived from the specific organism, beneficial for querying and reasoning. We use the `dwc:MeasurementOrFact` class to annotate measurement tables or paragraphs with organism fact descriptions that cover full paragraphs. We adopt the property *nhc:measuresOrDescribes* in our model to link an instance of the class `dwc:MeasurementOrFact` to a term relating to an anatomical entity (`UBERON:0001062`), such as "*liver*", or a property or attribute (`ncit:C20189`) of the organism, such as a "*colour*", which are measured or described in the table or paragraph. To omit annotation of a full paragraph, we can annotate only the entity that is being described. This way, we can use the entity to point users to a table or free text description of an organism's characteristic. One cultural historian was, for instance, interested in the adjectives used when describing the colour and morphology of anatomical entities. Pages describing a specific anatomical entity could be retrieved in single query e.g. "*Show me all observation records from person X that measure a liver*".

Verbatim Date. A further addition is the class `nhc:Date` . This class is used to annotate verbatim dates: An instance of the class, e.g., `nc:date1` is given a label such as "*10 Apr. 1821*" or "*Sept*". It is connected to the `dwc:Event` class using the *dwc:verbatimEventDate* to indicate this. The verbatim date will be converted to a standard format and linked to the `dwc:Event` class using the *dwc:year*, *dwc:month* and *dwc:day* properties. This way, dates can be used for querying using filters. Dates are an important part of species

descriptions and are easily annotated as they are formally formatted and have a prominent position on the page.

Written Annotations. Field books often contain manual annotations or revisions written above or adjacent to the original text. Types of annotations that occur a lot in our use case relate to the classification of an observed organism or an additional observation. A naturalist, for instance, classified an observed organism as a different taxon at a later date, based on further research of the described traits and anatomical parts or based on other literature. Whether this represents a shift in naming conventions, a new interpretation of the metadata or merely additional information or synonymy is unclear. Additionally, naturalists made side notes of observations of the same species by different naturalists at different locations, such as "*In Batavia according to Diard*".

In our qualitative analysis, biologists indicated that they were interested in exploring these annotations. They indicated that it was relevant for them to be able to discern which text was written at the time of the original observation, belonging to the original record, and which was added later. To emphasise these structures we added two properties; the *nhc:additionalIdentification* and the *nhc:additionalOccurrence* property. These are both added as sub-properties of the property *nhc:additional* such that all additional annotations can be accentuated or queried using this property.

Linking to External Ontologies and Datasets. The ontology connects to classes from other ontologies and thesauri (indicated by a striped fill in Figure 4.1) such as Uberon[1] for anatomical entities (106) and the NCI Thesaurus[2] for species attributes (107), both used for the identification of a taxon, the GeoNames Database[3] for geographical locations (108) and VIAF[4] for referring to persons (109) as instances of the class `foaf:Person` from the Friend Of A Friend (FOAF) language,[5] a vocabulary of properties and classes that makes use of the RDF technology. Linking to these vocabularies gives us three benefits. (1) the entities can be resolved, (2) queries can utilise the structures of these ontologies for querying and reasoning purposes, (3) the ontologies provide extra metadata. Instances from the GeoNames Database, for instance, are mapped to different historical name variants, abbreviations and modern names. As an example, the entity `http://sws.geonames.org/1648473` is linked to the modern name "*Bogor*" and simultaneously to the historical name "*Buitenzorg*", a term used in the field books.

[1] `http://purl.obolibrary.org/obo/`
[2] `https://ncit.nci.nih.gov`
[3] `http://sws.geonames.org/`
[4] `http://viaf.org/viaf/`
[5] `http://www.foaf-project.org/`

They distinguish a *gn:alternateName* with a language tag such as <gn:alternateName xml:lang="id">Kota Bogor</gn:alternateName> from a *gn:name*, revealing indigenous namings. Further, the property *gn:shortName* is used for abbreviations and *gn:officialName* for official names.

We choose not to link to IRIs from biological taxa in external datasets, as the same scientific name can sometimes refer to different organisms (discussed in Subsection 2.1.1). Disambiguation of species names requires metadata such as place of observation, date and biologist who performed the classification. We propose to create unique identifiers for each taxon within the namespace of the collection. After a careful analysis of the annotation data after the annotation process, these taxa can be resolved and linked to each other and taxa from external datasets. This preserves the verbatim content of the field books and allows scholars to link to distinct taxonomic datasets and species after the process of taxonomic referencing, should this be required to represent different theories.

Documenting Provenance of Annotations. Provenance is crucial in the disclosure of archival collections. The provenance of data extracted from collections contributes to their interpretation and value, and allows researchers to repeat experiments. To link semantic annotations to their provenance, the Web Annotation Vocabulary[1] was used. Reasons for adoption of the model are the use of the principles of Linked Data, its ability to address segments or fragments of media sources, and the fact that it is a W3C recommendation. Using the provenance data model, we can link instances of classes from the ontology depicted in Figure 4.1 to the image scans. Listing 4.1 shows an example annotation.

```
@prefix ex: <http://example.org/terms/> .
@prefix oa: <http://www.w3.org/ns/oa#> .
@prefix dcterms: <http://purl.org/dc/terms/> .

<http://example.org/anno54> a oa:Annotation ;
    oa:hasBody <https://viaf.org/viaf/45106482/>;
    oa:hasTarget ex:image1.jpg#xywh=x,y,h,w ;
    dcterms:created "2020-10-13T13:00:00Z" ;
    dcterms:createdBy <https://orcid.org/0000-0002-2146-4803> ;
    oa:motivatedBy oa:linking .
```

Listing 4.1: An example annotation

The resulting application ontology, a combination of the NHC-Ontology and the Web Annotation Vocabulary, provides a framework for annotating important named entities in the data. It is made accessible to users through a semantic annotation tool, the

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

SFB-Annotator, that enables the semantic annotation of digitised images of hand-written text and illustrations. The tool is discussed in the next section.

4.3 Semantic Annotation

In recent years, projects that create platforms for the storage, transcription and annotation of digitised historical documents on the web have begun to emerge. The *Field Book Project* (15), discussed in Subsection 3.2, was formed in 2010 as a joint initiative between the Smithsonian National Museum of Natural History (NMNH) and the Smithsonian Institution Archives (SIA). The project was set up to bring together field books from multiple NHCs and make them available for the general public.

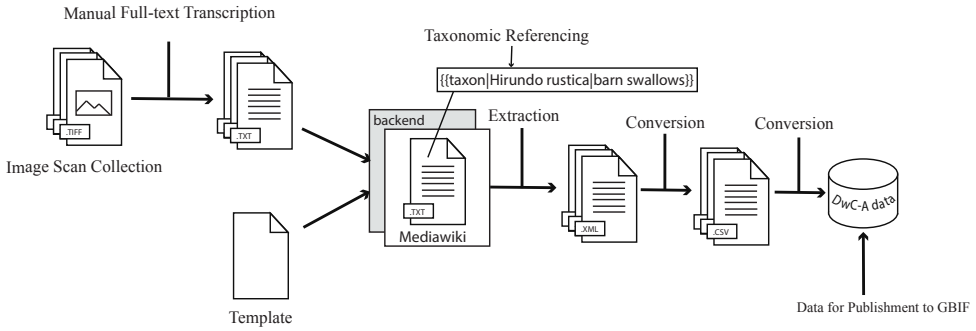


Figure 4.2: *From Documents to Datasets* (35) system design

The Field Book Project makes use of the NCD[1] standard for storing metadata on a *collection*-level. Further, the project uses the Metadata Object Description Schema (MODS)[2] to create item-level metadata (68). The BHL[3] describe their data using XML and MODS or Dublin Core (DC).[4] None of the above mentioned projects, however, aims to annotate the *content* from items within NHCs. Responding to this need, the project *From Documents to Datasets* (also discussed in Subsection 3.2) (35) provides a design for the conversion from digitised handwritten field books to a semi-structured annotated corpus, see Figure 4.2, using terms from the DwC standard. They propose first to fully transcribe the texts together with experts, then upload those texts together with the image scans to a MediaWiki[5] server. Via templates, the *taxa*, *locations* and *dates*, are annotated

[1] http://rs.tdwg.org/ontology/voc/
[2] http://www.loc.gov/standards/mods/
[3] http://www.biodiversitylibrary.org/
[4] http://dublincore.org/
[5] https://wikisource.org/

by researchers through a crowdsourcing initiative. Annotators can resolve verbatim names to current ones (taxonomic referencing) during the semantic annotation process. The annotations are then extracted and converted manually to DwC terms, in order to publish them in the GBIF [1] data server (69). This project provides an excellent methodology to structure named entities from field books. We thus build upon this methodology and extend it to fit our needs.

4.3.1 System Design

Similar to the projects mentioned at the beginning of Section 4.3, we use the NCD standard and the DC to enrich NHCs on a collection and item level. On a content level, our approach differs from the approach in Figure 4.2. In a similar fashion, semantics are added to the named entities. However, we use IRIs to describe the named entities, we link the IRIs together where possible to form a connected graph, and add hierarchical descriptions of classes and properties. The data become readable and interpretable by machines and can be interlinked and aggregated with other biodiversity data on the web, such as GBIF (see Subsection 2.3.1). To link the named entities together we use the NHC-Ontology, described in Subsection 4.2.4, which also enables rich querying and reasoning. Our system design is shown in Figure 4.3.

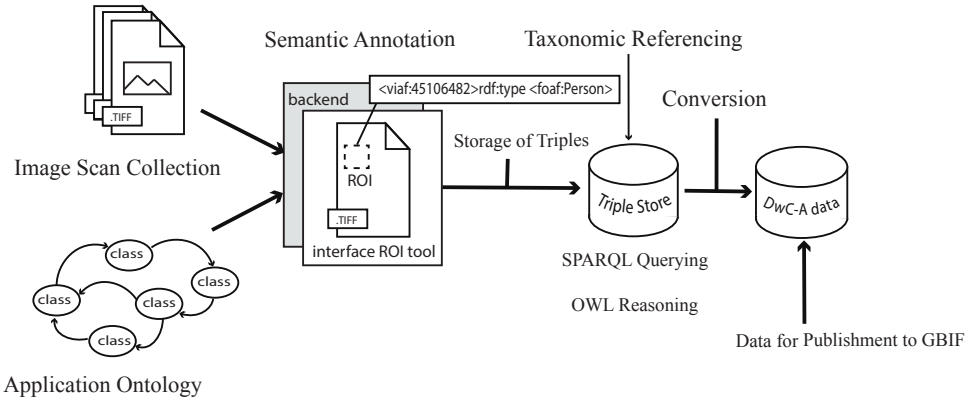


Figure 4.3: The proposed system for semantically annotating manuscripts from NHCs.

In contrast to design pattern (3) (see Section 3.2 and Figure 3.3), our approach omits the step of full-text transcription, and allows users to directly annotate text images (pattern (4)). To the best of our knowledge, no other system exists that uses an ontology to

[1] http://www.gbif.org/

annotate named entities in digital images of manuscript pages. We argue that annotation of the most important entities from the field books already allows biodiversity researchers to create models and search the texts, simultaneously minimising annotation efforts.

Furthermore, we suggest that the process of *taxonomic referencing* of species and genera should occur *after* all named entities from a field book or collection are annotated and linked. As mentioned earlier, fully linked field books allow for a thorough comparison between different taxonomies and naming conventions. After a careful analysis, these taxa can be resolved and linked to other taxa, but we argue that this should be decoupled from the first stage of the annotation process. Moreover, we argue that, especially with historical biodiversity data, multiple interpretations of the data should be able to exist in parallel. We therefore choose to annotate classification hierarchies in the collection verbatim, to facilitate multiple researchers adding their own layers of interpretations.

Additionally, researchers can attach free-text metadata to classes from the application ontology, using the properties from the DwC standard such as *dwc:habitat* or *dwc:samplingProtocol* which can be attached to the `dwc:Event` instance, *dwc:organismRemarks* to an instance of the class `dwc:Organism` or *dwc:identificationReferences* to add literature referenced in the manuscripts to the `dwc:Identification` class.

4.3.2 The Semantic Field Book Annotator

The Semantic Field Book Annotator (SFB-Annotator) is a web application, developed for domain experts, to harvest structured annotations from field books using the NHC-Ontology and proposed design.

Users can draw bounding boxes over ROIs in image scans, as shown in Figure 4.3 and 4.4, to which annotations can be attached. The ROI tool makes use of the *Annotorious* annotation Application Programming Interface (API)[1] to select a ROI and create an annotation object, see Figure 4.4. The annotation object is connected with its provenance and metadata: a target—a page or a ROI—and a body which links the ROI to either a transcription or an IRI. The geometry of the ROI is connected to the annotation object using *oa:hasSelector* and `oa:FragmentSelector`, see also Figure 4.5. In order to make the manuscript images zoomable, Annotorious is used together with the OpenSeaDragon API.[2]

[1] `https://annotorious.github.io/`
[2] `https://openseadragon.github.io/`

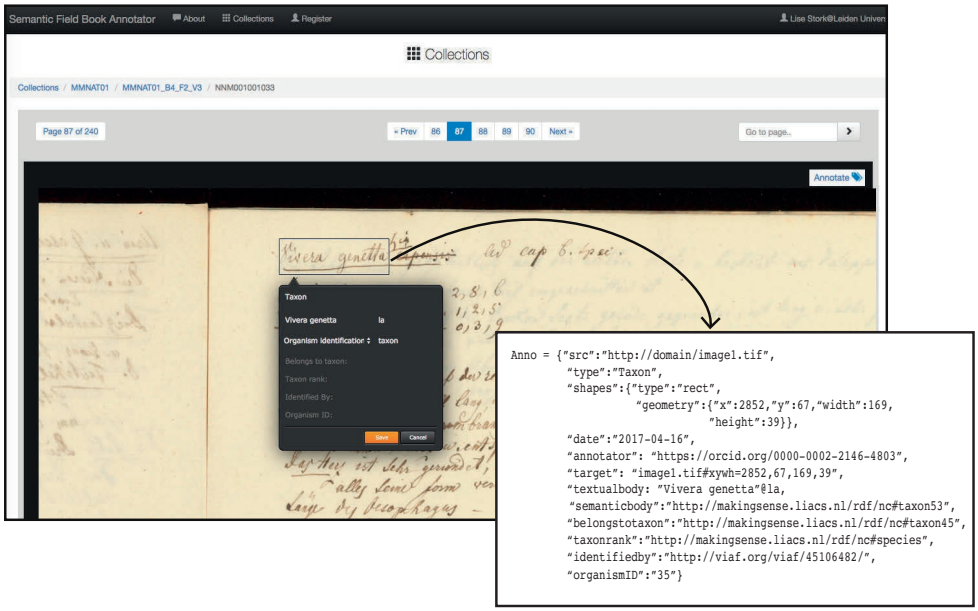


Figure 4.4: The annotation process using the *Semantic Field Book Annotator*

For storage, we use a servlet that pushes the annotation to an annotation server. In the servlet, annotation objects written in JavaScript Object Notation (JSON) are converted to RDF triples using the RDF4J API, an open source Java framework for processing RDF data. For storage of annotations we use the Virtuoso quad store as it is a well evaluated store for data-intensive server applications (110). Moreover, it can be accessed via the RDF4J API.

In the annotation process, a distinction is made between explicit and implicit classes. Explicit classes, in comparison to implicit classes, refer to the group of named entities that are easily observed in the field books, and therefore can be pulled out of the text more easily by annotators, and finally by automated processes. We refer to these with the term *salient* named entities. These are: the *taxonomic name, location, date, scientific publisher, writer, anatomical entities, properties* and *tables*. The implied classes serve to connect the explicit classes. However, they can also be used to link to class-specific meta-data encountered in the field books. The Darwin Core (DwC)'s *dwc:organismRemarks* can, for instance, be used to store free text descriptions from the field book about the organism under observation, as is also mentioned at the end of Subsection 4.3.1.

During the annotation process, a user first links a ROI to a class c from the set of *explicit* classes $C^e = \{c_1, c_2, ..., c_n\}$ of the application ontology. In figure 4.4 this is

the `ncit:C20189` or *property or attribute* class. The user then specifies a predicate p from the set of predicates $P = \{p_1, p_2, ..., p_n\}$, although this is only required in the case where multiple predicates are possible such as with the class `foaf:Person`. We however argue that it makes the annotation process more transparent and thus less error-prone. The predicates are displayed in a readable way, e.g., *Measures or describes:* `property or attribute`, such as visible in Figure 4.4, or for instance *Additional occurrence recorded at:* `location`. When a class and predicate are specified, optional metadata fields appear such as: `uberon:` IRI, in case of an anatomical entity.

During annotation, a single occurrence is given a unique code through the property *dwc:occurrenceID*. To create connections between all entities in one record that belong to a single occurrence, every time an instance is annotated, the entire base model, excluding the measurements, is instantiated, as visible in Figure 4.1. Unique identifiers for instances are created based on the unique occurrenceID, such as `nc:identification+occurrenceID`, such that new information will be added to the same organism occurrence graph. Even if entities are missing, IRIs exist but remain without a label until they are annotated by the user. More information about the SFB-Annotator and the annotation procedure can be found online.[1]

4.3.3 Towards Semi-Automated Annotation

As a first step towards semi-automated annotation, we pre-populated the knowledge base (a triple store) with domain knowledge concerning the collection, such as locations and names of researchers that participated in the expeditions. This contextual knowledge can aid annotators with the annotation process using autocomplete to retrieve candidate instances, such as `http://viaf.org/viaf/69703180/`, the VIAF record for Coenraad Jacob Temminck. The user can choose to annotate the verbatim text with a IRI from a set of candidate IRIs that exist in the triple store. If no instance yet exists or if it is an implicit instance such as one from the organism class, a (globally) unique IRI is created.

In Chapter 5, we further research methods for semi-automated annotation, using salient named entity recognition and classification (SNERC) for automated identification and classification of explicit salient named entities in digital field note images. The identification of these entities and their classifications can guide the retrieval of candidate instances for semantic autocomplete.

[1] `https://github.com/LINNAE-project/SFB-Annotator`

4.4 Qualitative Evaluation

In concordance with a domain expert from the field of natural history, one of the field books from the NC collection, named *'Manuscripten van de leden der Natuurkundige commissie: Mammalien, van Kuhl'*, was semantically annotated using the SFB-Annotator. This book contains observation records of species from three different orders: the order *Chiropterae*, or bats, the order *Quadrumana*, Latin for *the four-handed ones*, referring to the apes, and lastly the order *Falculatae*, a historical order referring to a collection of mammals such as the shrew, the badger and the bear. The coming sections will qualitatively evaluate the annotation process (Subsection 4.4.1) the resulting data (Subsection 4.4.2), and possibilities for querying using the concepts and questions composed by the domain experts, mentioned in Subsection 4.2.3.

4.4.1 The Annotation Process

Annotating a page from the field book using the *Semantic Field Book Annotator* ranged between approximately 1 to 10 minutes, depending upon the amount of named entities on the page and the difficulty of interpreting a named entity. Taxonomic names such as the one in Figure 2.6, (*Titthaecheilos javanicus*) can be difficult to read. When the order of pages is shuffled, the correct interpretation of links between entities is further hampered. Other names, however, are easier to read and connect to related named entities. As the layout of the document hints to the location of the named entities, the annotation process quickly becomes easier. Taxonomic names, scientific publishers of names, and locations are likely to appear on the top of a page.

As the time spent annotating a named entity largely depends upon its readability and interpretability, we argue that the biggest difference between our approach and the one in Figure 4.2 is the omission of one processing step. Where other approaches first transcribe the entire text and then look for named entities to be semantically enriched, we omit the first step and directly search for named entities to be enriched. Consequently, we argue that this results in faster processing of field books to graphs in a knowledge base. We do realise that linking to other entities might be a process that can prove more challenging than merely annotating the class of an entity.

4.4.2 The Data

From the annotated field book, 98 single pages[1] were semantically annotated and their annotations validated by a natural history expert. Table 4.2 shows the number of named entities that were extracted from the field book pages, the size of the triple store and the *per page*, *per class* and notable *per predicate* statistics.

In the case that a named entity is absent in a linked observation record, for instance if an annotator omitted the annotation of a named entity, querying the data is not hampered and can even, together with graphic visualisations of the data, help control data quality. When a named entity is not annotated, for instance the location of the organism observation, the IRI lacks a label, a link to an annotation object and thereby a span in the image (a ROI), as mentioned at the end of Subsection 4.3.2. Observation records of which the location is absent or not yet annotated can be found by querying the knowledge base for locations without a label or annotation.

Table 4.2: Annotation specifications

Total Annotations

Pages	**Size MB**	**Observ. Records**	**NEs**	**Triples**	**NEs per page**	
					μ	σ
98	1.5	34	371	9921	5	2.8

Annotations per class

Class	***n***	**Class**	***n***
`dwc:Taxon`	52	`nhc:Date`	6
`foaf:Person`	47	`uberon:0001062`	160
`dcterms:Location`	15	`ncit:C20189`	28
`dwc:MeasurementorFact`	13	*Total*	371

Predicate specifics

`Class`	*Predicate*	***n***
`foaf:Person`	*nhc:scientificNameAuthorship*	41
	dwciri:recordedBy	35
	dwciri:identifiedBy	39
`dwc:Organism`	*nhc:additionalOccurrence*	3
	nhc:additionalIdentification	15

[1] During the digitisation process, the field notes were scanned two pages at a time. One page here refers to one *physical* page containing text, rather than one digital image.

4.4.3 Semantic Queries

In this section we evaluate, using the annotated data, which questions are common in terms of search requirements, determine if and how the questions can be answered using SPARQL and the NHC-Ontology, and demonstrate the gain in comparison to full-text search.

Domain Expert's Queries. The evaluation in Subsection 4.2.3 resulted in a list containing 53 research questions.[1] 18 questions were from biologists, 28 from cultural historians and 7 from information specialists.

To estimate the nature of common research questions, the questions were grouped together on the basis of types of named entities. Most common questions were: a question combining a type of resource and a person name, e.g., "*Show me all field notes from person X*", and a question combining the person class and a taxon name, e.g., "*Did specific naturalists have a specialisation such as plants or animals?*". The entities used in the queries were all covered by the model, except for some more specific person classes such as a local helpers or illustrators.

From the 53 questions, 7 did not relate to the content of the field books and were therefore excluded from the question set. They could potentially be addressed with other parts of the archive. For instance, "*How was a day organised*" relates to the field observation practices, something that is more likely to be found in the diaries within the archive. Another example is "*Are there letters from person X to person Y in the collection?*". Such a question could be answered by querying the collection for both person X and Y, making use of their IRI to overcome name ambiguity. Both diaries and letters are however beyond the scope of this paper.

Four of the questions related specifically to specimens and their preservation. Although we did not annotate specimens, the semantic model does allow these type of queries. The label of a physical specimen or its digital image can also be used for semantic annotation, as mentioned in Subsection 4.2.4. The class `dwc:PreservedSpecimen` is then used instead of `dwc:HumanObservation`.

For clarification a distinction is made between six types of queries, see Table 4.3. The table includes a count of how often each type of question occurred in the question set. "Which" and "Where" questions were often seen as entity retrieval tasks, except in the case of "which page" or 'where in the archive', and open questions were seen as document retrieval tasks. Closed questions that can be answered with a "yes" or "no" were also seen

[1] `https://github.com/lisestork/NHC-Ontology/blob/master/Questions_orderedByEquality.xlsx`

as document retrieval tasks, as these are usually questions that require further inspection of a document. For both query variants, queries were evaluated with regards to relevance of the search results and if extra effort is required by the user after retrieval.

Table 4.3: Types of expert queries

Query type	Count
T1: "All *documents* containing keyword *k*."	1
T2: "All *documents* matching structure *s*."	18
T3: "All *documents* matching structure *s* and keyword *k*."	7
T4: "All *entities* containing keyword *k*."	0
T5: "All *entities* matching structure *s*"	7
T6: "All *entities* matching structure *s* and keyword *k*	13

Structured vs. Full-Text Queries Where structured query-languages such as SPARQL are better at querying the *structure* of the data, full-text queries are better at querying the *content* (111). Here, we demonstrate that in the case of field books, structured or hybrid queries (112) using the NHC-Ontology are able to provide more relevant query results than full-text queries.

It is notable from table 4.3 that few questions involved simple keyword searches. The only question that can be answered directly using a keyword is: "*Show me all resources (lists, drawings and observations concerning a specific species* ***k****.*" *k* being the keyword, as no limit is imposed on the type of resource that should be retrieved. For 5 of the questions of type T3, full-text search can also provide an answer, although not directly. Examples are the following questions: "*What did person* ***k*** *find?*" or "*Which drawings were made by person* ***k***". However, *all* resources that in any way relate to person **k** would be retrieved, thus retrieving irrelevant documents alongside relevant ones.

Most common queries are structured queries retrieving specific documents (T2) such as "*Show me all drawings with a head of a fish*" and hybrid queries retrieving named entities (T6) such as "*Which anatomical entities were used for the classification of the family Pteropodidae*". When transformed to hybrid queries, 25 out of 46 queries will provide a direct answer to the original question. For the remaining 21 of 46 queries, document pages are presented to the user that will likely contain an answer to their question, an example being: "*How were habitats described in the collection between dd-mm-yyyy and dd-mm-yyyy?*". The semantic query can point a user to the pages that adhere to these date restrictions, but the user will have to inspect them to answer his or her question.

Listing 4.2 to 4.5 below presents 4 of the 46 questions in SPARQL form, two for cultural history two for biology research. Listings 4.2 and 4.3 are example SPARQL queries for

cultural history research, and provide an indirect answer to the questions mentioned in the listing captions:

```
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>
PREFIX dwciri: <http://rs.tdwg.org/dwc/iri/>
PREFIX dsw: <http://purl.org/dsw/>
PREFIX viaf: <http://viaf.org/viaf/>
PREFIX oa: <http://www.w3.org/ns/oa#>
SELECT ?label ?page WHERE {
      ?identification dwciri:toTaxon ?taxon .
      ?taxon rdfs:label ?label .
      ?organism dsw:hasIdentification ?identification .
      ?occurrence dwciri:recordedBy viaf:45106482 .
      ?occurrence dsw:hasEvidence ?observationRecord .
      ?anno oa:hasBody ?observationRecord .
      ?anno oa:hasTarget ?page }
```

Listing 4.2: How were species collected by Heinrich Kuhl, viaf:45106482?

```
PREFIX nhc: <http://makingsense.liacs.nl/rdf/nhc/>
PREFIX dwc: <http://rs.tdwg.org/dwc/terms/>
PREFIX dsw: <http://purl.org/dsw/>
PREFIX oa: <http://www.w3.org/ns/oa#>
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>
SELECT ?page ?label WHERE {
      ?event dwc:year ?year
      FILTER ( ?year >= 1820 ) .
      FILTER ( ?year <= 1821 ) .
      ?event nhc:verbatimEventDate ?date .
      ?date rdfs:label ?label .
      ?event dsw:eventOf ?occurrence .
      ?occurrence dsw:hasEvidence ?observationRecord .
      ?anno oa:hasBody ?observationRecord .
      ?anno oa:hasTarget ?page }
```

Listing 4.3: How were habitats described in the collection between *1820* and *1821*?

Listings 4.4 and 4.5 below are examples of queries for biology research, and provide a direct answer to the questions mentioned in the captions. More example queries can be found online.[1]

```
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>
PREFIX nhc: <http://makingsense.liacs.nl/rdf/nhc/>
PREFIX nc: <http://makingsense.liacs.nl/rdf/nc#>
PREFIX dwc: <http://rs.tdwg.org/dwc/>
PREFIX dwciri: <http://rs.tdwg.org/dwc/iri/>
PREFIX dsw: <http://purl.org/dsw/>
PREFIX viaf: <http://viaf.org/viaf/>
PREFIX oa: <http://www.w3.org/ns/oa#>
PREFIX gn: <http://www.geonames.org/ontology#>
SELECT DISTINCT ?label WHERE {
      ?taxon rdfs:label ?label .
      ?taxon nhc:taxonRank nc:species .
      ?taxon nhc:belongsToTaxon  ?order .
      ?order rdfs:label ?Chiropterae .
      FILTER regex(?Chiropterae, "Chiropterae") .
```

[1] https://github.com/lisestork/NHC-Ontology

```
    ?identification dwciri:toTaxon ?taxon .
    ?organism dsw:hasIdentification ?identification .
    ?occurrence dsw:occurrenceOf ?organism .
    ?occurrence dwciri:recordedBy viaf:45106482 .
    ?occurrence dsw:atEvent ?event .
    ?event dsw:locatedAt ?location .
    ?location dwciri:inDescribedPlace ?place .
    ?place gn:parentFeature ?parent .
    ?parent gn:alternateName ?name
    FILTER regex(str(?name), "Java", "i") }
```

Listing 4.4: Which chiroptera species were collected by Heinrich Kuhl, viaf:45106482, on Java?

```
PREFIX dwciri: <http://rs.tdwg.org/dwc/iri/>
PREFIX dsw: <http://purl.org/dsw/>
PREFIX uberon: <http://purl.obolibrary.org/obo/>
PREFIX ncit: <http://identifiers.org/ncit/>
PREFIX nhc: <http://makingsense.liacs.nl/rdf/nhc/>
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
SELECT DISTINCT ?label2 ?uberon
WHERE { ?identification dwciri:toTaxon ?taxon .
        ?taxon rdfs:label ?label
        FILTER regex(?label, "Pteropus")
        ?identification dsw:isBasedOn ?token .
        ?token dsw:hasDerivative ?measurement .
        ?measurement nhc:measuresOrDescribes ?anatomy .
        ?anatomy rdfs:label ?label2 .
        ?anatomy rdf:type ?uberon .
        ?uberon rdfs:subClassOf uberon:UBERON_0001062 }
```

Listing 4.5: Which anatomical entities were used for the classification of the genus Pteropus?

We finally argue that, as Virtuoso is equipped with full-text indices that can be queried via SPARQL (110), queries can be formulated both as full-text, semantic or hybrid queries. However, as most queries make use of the structure of the data *in combination* with keywords, making use of semantic queries is beneficial for the retrieval process.

We note that the average user should not be required to write complex SPARQL queries. To take on this problem, methods have been developed that bridge the gap between the Semantic Web and the domain expert users (113; 114; 115).

For further observation, the ontology can be found online together with the domain experts' questions, the questions transformed to queries and a visualisation of one fully linked observation record.[1] The semantic annotations can be accessed through a SPARQL endpoint[2] which can be queried using a SPARQL query editor.[3] The code for the SFB-Annotator and annotation guidelines can also be found online,[4] and will be updated once newer versions are available.

[1] https://github.com/lisestork/NHC-Ontology
[2] http://makingsense.liacs.nl/rdf4j-server/repositories/NC
[3] An example query editor is the Yasgui editor: http://yasgui.org/, accessed: 30-03-2018
[4] https://github.com/LINNAE-project/SFB-Annotator

4.5 Conclusions

In this chapter, we presented a semantic model and tool for the semantic annotation of field books. Through the semantic annotation of one field book, we evaluated the model and demonstrated the annotation approach. This approach will eventually lead to a structured dataset constructed from the NC collection, available through a SPARQL endpoint. It is an example of how the content of historical collections in general could be disclosed using semantic annotation.

The qualitative evaluations demonstrated that the application ontology adheres to our requirements and is usable by domain experts both for the process of creating structured annotations as well as answering common research questions. Answers to structured queries will either point users to specific pages, to enable closer inspection of the original text, or provide them with lists or graphical output. However, as the model we propose is centered around the observation and collection of organisms from field books, it currently serves the requirements of the biologists and taxonomists better than the cultural historians. We anticipate that extensions to the model will be required when annotating other artifacts in the collection. Letters and diaries from the collection, for example, describe the economy, villages, cultures and inhabitants of colonial Indonesia, and accompanying drawings depict environmental conditions. A base model for these resources would provide a useful addition to the semantic model we propose.

4.6 Ongoing and Future Work

Recently, the SFB-Annotator has become part of a project called the LInking Notes of NAturE (LINNAE).[1] Within this project, we worked together with a research software engineer from the eScience center[2] to bring the SFB-Annotator online for use in the biodiversity domain (116).[3] Amongst others, developments include the refinement of the data model (exemplified with an example annotation in Figure 4.5), packing of the application in a Docker container[4] to ease installation, and the migration of the tool's infrastructure to the International Image Interoperability Framework (IIIF),[5] which is becoming a standard for viewing and annotating cultural heritage manuscripts online, see Figure 4.6.

[1] https://github.com/LINNAE-project
[2] https://www.esciencecenter.nl/
[3] https://research-software.nl/software/sfb-annotator
[4] https://www.docker.com/
[5] https://iiif.io/

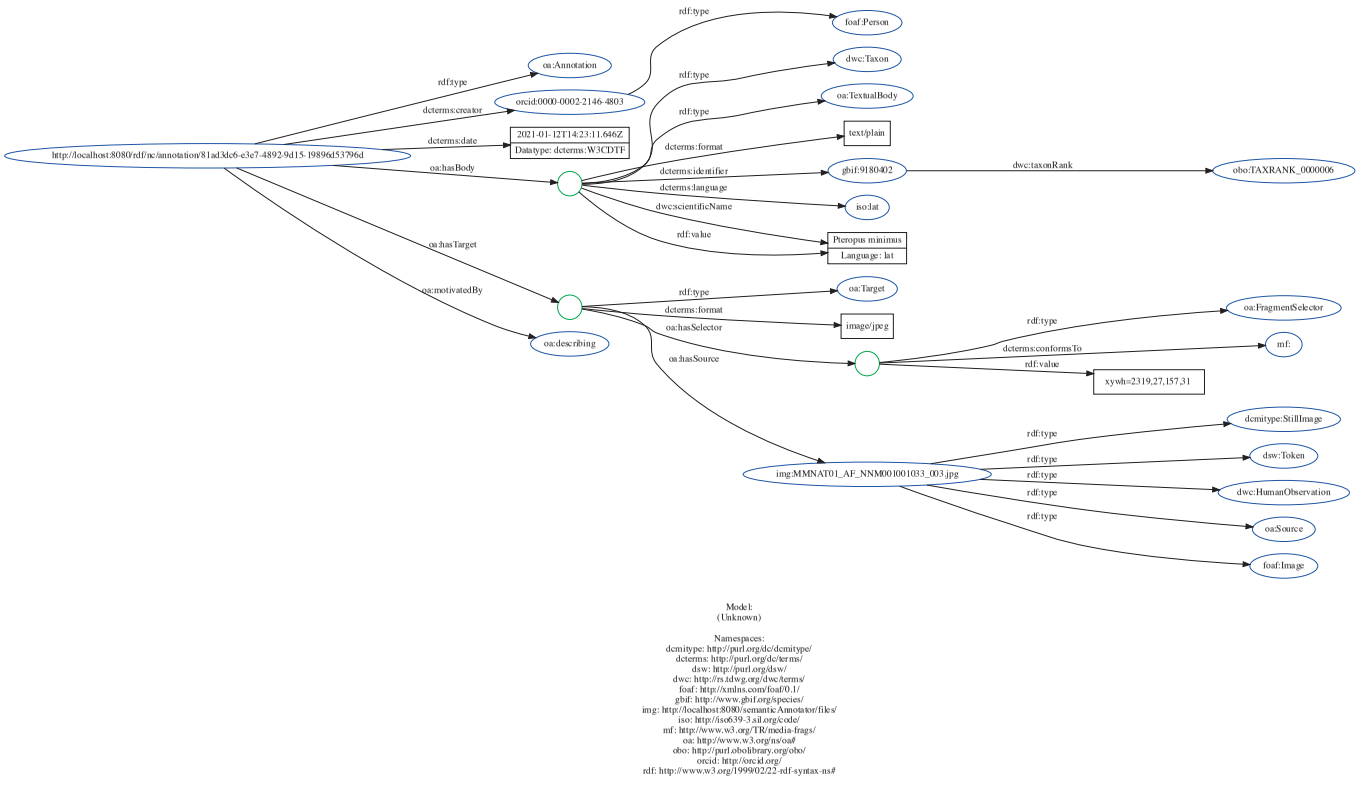


Figure 4.5: Example semantic annotation of an annotation, using the refined data model. Courtesy: A. Kuzniar (2021). Other example annotations can be found here: `https://github.com/LINNAE-project/SFB-Annotator/tree/master/doc/models`

Through the Cantaloupe image server, images and their metadata are retrieved, converted to JPG and sent to the IIIF viewer. RDF annotations can be retrieved through the IIIF manifest server and appended to the `manifest.json`, a template to present images in the viewer, although this is still ongoing work. As an image viewer, we depend on the Mirador IIIF viewer,[1] which includes OpenSeaDragon for zoomable images and uses the Web Annotation Data Model[2] for annotations. To query the final knowledge graph, we employ the GRLC tool (117), which translates SPARQL queries to Linked Data Web APIs.[3] This work is supported by the Netherlands eScience Center (Grant Number: 27019P01).

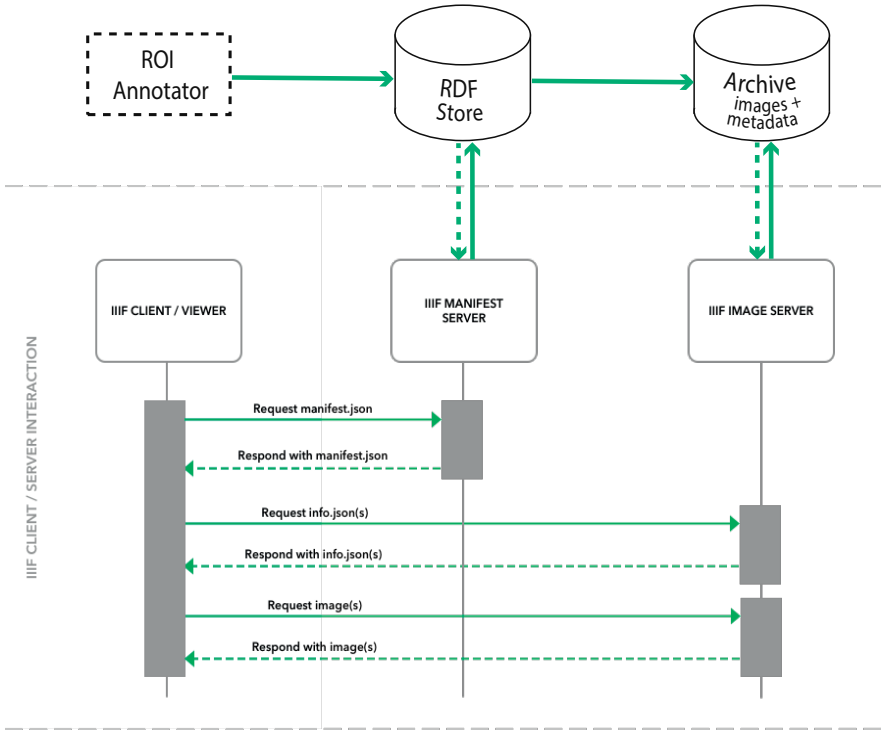


Figure 4.6: Proposed architecture of the SFB-Annotator.[4] Courtesy: A. Kuzniar (2020)

In our next steps, the usability of the SFB-Annotator will be further improved; we will continue to evaluate the model with a small expert crowd to assess if the annotation task is well defined and to retrieve more accurate annotation time estimates.

The annotations that were harvested during the first evaluation of the SFB-Annotator (see

[1] `https://projectmirador.org/`
[2] `https://www.w3.org/TR/annotation-model/`
[3] `https://github.com/CLARIAH/grlc`
[4] Figure is derived from `https://iiif.github.io/training/intro-to-iiif/SOFTWARE.html`

Subsection (4.4.2) will serve as a dataset for automating part of the annotation process. With fully transcribed texts, NLP can be used for the purpose of semi-automated semantic annotation. As we use text images instead of digital texts, we require alternative, computer vision methods for NERC, which rely on structural and positional features of words for annotation (84; 118; 119). We present first experiments of this process in the following chapter, Chapter 5.

CHAPTER 5

Automating Semantic Annotation

"Perhaps the deepest accomplishment of cognitive development is the construction of larger-scale systems of knowledge: [...] Building these systems takes years, much longer than learning a single new word or concept, but on this scale too the final product of learning far outstrips the data observed."

– Joshua B. Tenebaum, Charles Kemp, Thomas L. Griffiths, Noah D. Goodman, in: *How to Grow a Mind: Statistics, Structure, and Abstraction*

Biological nomenclature and systematics (discussed in Chapter 2), forms the basis of worldwide scientific discourse about the biodiversity of our planet. Employing such prior knowledge about biological structures in machine learning models, enables the process of learning to retrieve these structures accurately from only small data samples that encode them. At the same time, historical field observations, even more than contemporary ones, contain fierce discussions about systematics and nomenclature. Biological taxonomies, once extracted from archives, can be used to search historical records. Systems can exploit extracted taxonomies through query expansion techniques, or allow users to semantically query, or browse through, archival collections.

In this chapter, we aim to answer research question **Q.4**: *How can we use automated methods for knowledge extraction from archives of NHCs?* by aiming to automate part of the pipeline for knowledge extraction from *field books*.

5.1 Introduction

Automatic knowledge extraction from field book manuscripts and illustrations is challenging as content is too heterogeneous to process using common HTR techniques (**Chall.6**). HTR is one of the more challenging tasks within the field of Document Image Analysis and Recognition (DIAR), mainly due to the huge variety in writing styles and languages, paper degradation, overlapping words and historical handwriting. Creating labelled examples

for HTR requires domain expertise (**Chall.7**), and interesting words lie in the long tail of the distribution of words (**Chall.8**). Examples of interesting content that lies in this long tail, are scientific names. In Chapter 4, we saw that domain experts were interested in retrieving such names, see Table 4.1.

Here, we use computer vision and Semantic Web technologies to (i) identify the elements of scientific species names in handwritten document images, and (ii) link and structure the elements, using an ontology for species observations. We use the MONK handwriting recognition system (23) to segment the document images into single word images. Our main contribution is the automatic identification and semantic annotation of word zones in manuscripts that contain species names, and the goal is to integrate such a system with a system for HTR, together tackling the task of named entity transcription and salient named entity recognition and classification (SNERC).

We build on work described in the previous chapter (Chapter 4), where an ontology and software for semantic annotation of species observation records was constructed and tested with domain experts. Here, we advance these methods by automating the process of semantic annotation. We present a a novel approach to identify *scientific names* in historical *handwritten* document images. Rather than first transcribing the text and performing NERC afterwards on the digital text, we exploit characteristics of the document images for identifying the domain specific salient named entities, using terms from the NHC-Ontology[1] to classify and organise them. We argue that the ability to quickly index handwritten document images based on scientific names, ranks and authors, helps users to navigate through large collections of documents in online libraries, such as the Biodiversity Heritage Library (BHL).[2] It opens up possibilities for faceted search, semantic querying and semantic recommendations. Additionally, maintaining a link to the word image and location in the full document image is important to generate ground truth for repetition of image processing experiments as well as to allow researchers to view the original document and therefore the extracted text in context.

5.2 Related Work

Organisations and researchers that dedicate themselves to the preservation of natural history collections, such as *IdigBio*[3] or the BHL (13), continuously develop new methods to digitise specimen collections in a cost-effective and sustainable way, in order to facilitate ongoing species research.

[1] http://www.makingsense.liacs.nl/rdf/nhc/,https://github.com/lisestork/nhc-ontology/
[2] https://www.biodiversitylibrary.org/
[3] https://www.idigbio.org/

The automatic extraction of scientific names from text is essential for the management of archival resources. Therefore, there are several examples of methods for extracting and disambiguating species names from printed texts, but extracting the same information from handwritten texts is much more of a challenge. TaxonGrab (120), for example, automatically extracts species names from printed biological texts. The BHL, which aggregates scans of biodiversity publications and field notes, indexes scientific names extracted from the publications—printed text—in their collection, to improve accessibility for taxonomists. They match the text, extracted via OCR, with the Taxonomic Name Server (TNS) to identify likely scientific names (13).

Similarly to the BHL, other researchers and institutes are exploiting the power of automatic text processing for the digitisation of natural history collections. Software has been developed to parse OCR output of printed text to formalised DwC entries for archival and retrieval purposes (121). Drinkwater et al. (20) investigate the aid of OCR in the digitisation of herbarium specimen labels, and found a significant increase in time effectiveness using OCR output to (i) sort specimens prior to database submission, and (ii) to add transform labels to minimal database records. Drinkwater et al. explicitly note that OCR is currently only possible for typed and printed labels and not for handwritten text.

As HTR is one of the more challenging tasks within the field of DIAR, mainly due to the huge variety in writing styles and languages, paper degradation, overlapping words and historical handwriting (**Chall.6**). The recognition of named entities can help document understanding and searchability of the text, and can potentially aid HTR (86). Formerly, NERC was a task solely used on digital text, but it has recently also been applied directly to handwritten text (85; 84; 86). Especially when few instances of words exist and a collection consists of many different hands and connected words, making it difficult to create character-based representations, the identification of key words can help make the text searchable, and potentially aid HTR. Moreover, in many cases, full-text transcriptions of entire pages of field books are not required in order to make them digitally accessible.

5.3 Data

Transcribed field books exist online, but (to the best of our knowledge) no segmented and annotated images of handwritten species observations are available online. For this purpose, word images from 240 field notes from a natural history collection have been segmented and semantically annotated. The process of annotation has been carried out in the context of this work. However, the process of segmenting digital images into word

zones has been carried out by the MONK system for the project *Making Sense of Illustrated Handwritten Archives*[1] (19), and this is reflected in Figure 5.1.

From a field book on mammals, we selected field notes from four different writers, to account for different handwriting styles and structures, ensuring a representative dataset to demonstrate how the automated methods perform on heterogeneous, real-world data. The segmented word images were obtained from a nichesourcing effort, with the help of a handwriting recognition system MONK and a group of domain expert labellers. The word images were subsequently manually annotated using four classes, as shown in Table 5.1. Two of four classes are taxonomic entities. The third class refers to the publisher of the taxonomic name, and lastly we have the class *Other*, which includes all words that do not belong to any of the previously mentioned classes.

Table 5.1: Dataset class count

class	Genus	Species	Author	Other	Total
y	0	1	2	3	
n	177	167	144	17309	17797

The final counts of examples per class are shown in Table 5.1. The process of labelling and annotating words is time-consuming and, in our case, requires expert knowledge. Therefore, limited training data is available. As machine learning methods generally require a very large number of labelled samples, methods have to be adjusted to the dataset size to acquire a predictive model that generalises well. These adjustments are described in Section 5.4 and 5.5. This is also one of the challenges of projects working with real-world data where obtaining labelled data is expensive or simply not feasible. Models that use prior knowledge are better able to generalise from noisy data and small samples. The dataset used in this work can be found online.[2]

5.4 Scientific Name Extraction Model

Below we describe our contribution. The full pipeline is shown in Figure 5.1, the blue rectangle indicating the scope of this work.

We used the MONK handwriting recognition system (23; 26), developed by Schomaker, for word segmentation (122; 123; 124). First, the system segments handwritten document images into lines and second, relative to those lines, into word zones that potentially hold words. The system allows the labelling of word images and transcription of sentences by

[1] http://www.makingsenseproject.org
[2] 10.5281/zenodo.2545573

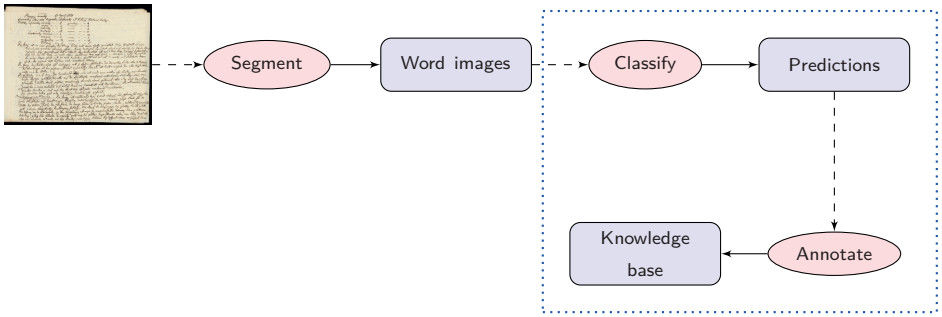


Figure 5.1: The full pipeline: automated semantic annotation of scientific names

domain experts. It then uses these labels for HTR. In this work, the word images were manually annotated using four semantic concepts, or classes: `genus`, `species`, `author` and `other`. The classification of each word image to its corresponding semantic class is discussed in Subsection 5.4.1. In Subsection 5.4.2, we discuss the semantic annotation of the classified word images using the NHC-Ontology[1] for species observations.

5.4.1 Classification of Word Images

To classify the word images to one of four classes, we use three distinct features; *visual structural features, position and context*. We chose to create one single neural architecture, built with help of Keras (125), that could be trained end-to-end, so that the classification error is only propagated once, in contrast to using predictions from multiple classifiers and combining them after training to form a single prediction. The final architecture is explained visually in Figure 5.2, and will be discussed below.

Visual Structural Features. The feature detector that was used in this work for the detection of visual structural features is a CNN (126). It has been shown that CNNs outperform other ANNs on image recognition tasks (127), see Section 2.2.1. The basic network used here is a deep CNN for object recognition developed and trained by Oxford's Visual Geometry Group (VGG) and called the VGG network (127). We use their configuration, with 16 convolutional layers, and import weights pre-trained on the ImageNet task by the VGG (128). Previous work (129) has demonstrated that transferring image representations with CNNs overcomes the problem of training with limited training data, e.g., less than a few thousand training images, despite differences in image statistics between the *source* dataset and *target* dataset. By, for instance, training on the ImageNet task, the VGG model learns filters on various different scales, which can be used as feature extractors for

[1] `http://www.makingsense.liacs.nl/rdf/nhc/`,`https://github.com/lisestork/nhc-ontology/`

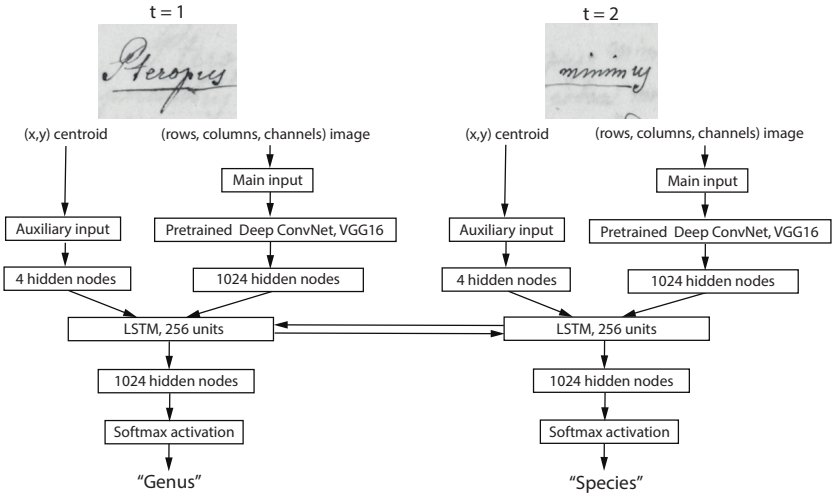


Figure 5.2: The CNN–MLP–BLSTM architecture, "unrolled" for both time steps t.

other types of images. These features, extracted from handwritten documents with help of the convolutional part of the VGG network, are used for training a simple MLP on our task.

Position. In addition to visual features, the position of a word in a document, especially (semi)-structured ones such as field observation records, often provides a good descriptive feature for the recognition of a named entity. The position is therefore often used as a feature in the field of NERC, however, it has been used more often in digital text, e.g., (130) than in digital images, e.g., (85; 84; 86; 83). In this work, we use the relative centroid of the word images, $c = (x, y)$, relative to the image borders, as input features to a simple MLP with two inputs, x and y, and one hidden layer of size 4. To train the entire model end-to-end, we concatenated the last hidden layers of both models. The merged hidden layer therefore has a size of $1024 + 4 = 1028$.

Context. As a third feature type, we introduce context: the characteristics of adjacent word images, specifically *bi-grams*. Figure 5.3 shows frequencies for word image bi-grams. First, horizontal pairwise alignment was calculated per word $w^{(i)}$ and $w^{(j)}$. They were seen as horizontally aligned if $y1^{(i)} < yc^{(j)} < y2^{(i)}$, where i and j indicate the i-th and j-th word image, $y1^{(i)}$ the first y coordinate of $w^{(i)}$, $y2^{(i)}$ the second, and $yc^{(j)}$ the y coordinate of the centroid of $w^{(j)}$. Second, the right neighbouring word of $w^{(i)}$ was retrieved by calculating all pairwise vertical distances for the horizontally aligned words: $dist_{ij} = cx^i - cx^j$, where cx^i refers to the x coordinate of the centroid of $w^{(i)}$.

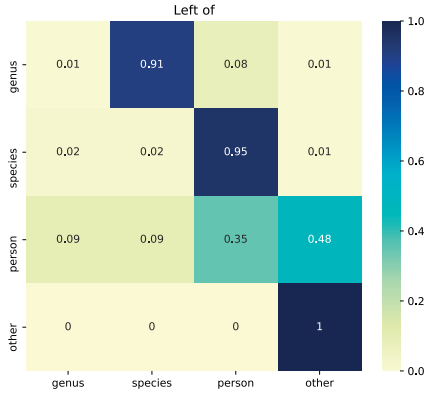


Figure 5.3: Adjacency matrix that shows frequencies for word bi-grams (sequences of two adjacent words). E.g., 'genus' was **left of** 'species' 91% of the time 'genus' was encountered.

The smallest negative distance, within a certain bound, indicated right adjacency. The adjacency matrix only takes into account instances that actually have an adjacent word, as it could be that a word is surrounded by white space on every side.

As expected, the different classes have strong co-occurrence dependencies. Therefore, we converted the dataset to sequences of size two (bi-grams), and added a last layer to the model architecture for sequence prediction. For an adequate prediction we used a BLSTM neural network (discussed in Subsection 2.2.1) that is able to learn long-term dependencies between features. By using the bidirectional variant of the LSTM, dependencies can be learned in both horizontal orientations, see Figure 5.2.

5.4.2 Semantic Annotation of Word Images

The NHC-Ontology[1] is an ontology for species observations, based on the DSW ontology, and written in OWL.[2] The ontology is centered around the description of meta-data relating to the observation of an organism, and allows a researcher to describe as which various taxon groups an organism has been identified. The model uses the Web Annotation Vocabulary[3] to link bounding boxes of word images to their semantic labels. In the example listing below, Listing 5.1, two images refer to a genus and a species, which together constitute one taxonomic name `ex:taxon1` of rank `ex:species`. They are linked to the publisher of the name with the *nhc:scientificNameAuthorship* property.

[1] `http://www.makingsense.liacs.nl/rdf/nhc/`,`https://github.com/lisestork/nhc-ontology/`
[2] `https://www.w3.org/OWL/`
[3] `https://www.w3.org/TR/annotation-vocab/`

```
@prefix nhc: <http://makingsense.liacs.nl/rdf/nhc/> .
@prefix ex: <http://example.org/terms/> .
@prefix rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#> .
@prefix rdfs: <http://www.w3.org/2000/01/rdf-schema#> .
@prefix oa: <http://www.w3.org/ns/oa#> .
@prefix dwc: <http://rs.tdwg.org/dwc/terms/> .
@prefix foaf: <http://xmlns.com/foaf/0.1/> .

ex:taxon1 rdf:type dwc:Taxon ;
          nhc:scientificNameAuthorship ex:author1 ;
          nhc:taxonRank ex:species .

ex:author1 rdf:type foaf:Person .

ex:anno1 rdf:type oa:Annotation ;
         oa:hasBody ex:taxon1 ;
         oa:hasTarget ex:image1.jpg#xywh=x,y,h,w ;
         oa:hasTarget ex:image1.jpg#xywh=x,y,h,w .

ex:anno2 rdf:type oa:Annotation ;
         oa:hasBody ex:author1 ;
         oa:hasTarget ex:image1.jpg#xywh=x,y,h,w .
```

Listing 5.1: Example of a semantically annotated species name

5.5 Experiments and Results

To analyse the influence of the three features on the predictive performance of the model, we conducted multiple experiments where we tested the performance of the pre-trained CNN, CNN–MLP and CNN–MLP–BLSTM.

5.5.1 Experimental Methodology

Before training, the images were scaled by dividing them by 255 so that they would fall within the range [0-1]. All images were re-sized to the average image dimensions: $y = 74$, $x = 139$. No data augmentation was used. Based on horizontal adjacency, as explained in Subsection 5.4.1, image bi-grams were constructed, sequences of $l = 2$, as input to the BLSTM.

The word images were shuffled, keeping together word images from the same page, and thereafter split into a train and test set. As one word image could occur in two bi-grams, we hereby avoid that word images from the test set were also in the training set, which would bias the classification results. However, by shuffling the pages, we still ensure that the model does not overfit to one writing style or structure. We used 80% of the word images for training and the remaining partition as test set, making sure that 20% of the scientific name elements were in the test set. As classes in the word bi-grams were

highly imbalanced, we used random minority oversampling with replacement, to increase the counts of samples from minority classes in the training data. When training a CNN, oversampling is thought to be the best method to deal with imbalanced datasets with few examples in minority classes, and appears to work best if the oversampling totally eliminates the imbalance (52).

However, as we are dealing with sequences rather than singular samples, we chose to oversample sequences, e.g., *species–author*. Converted back to singular images, this would result in a *step imbalance* with a small imbalance ratio $p = \pm 1.1$ rather than a large imbalance ratio of $p = \pm 16$ (52).

The networks were all trained using the Adam classifier with a learning rate of 10^{-4} and categorical cross-entropy loss. Each network was trained using early stopping with patience 2, meaning that training was stopped when, for two epochs, the validation error was increasing. Per epoch, the weights were only stored if the predictive performance had increased compared to the previous epoch. In the testing phase, thresholding was applied to the output of the networks to compensate for oversampling the data during training, as oversampling alters prior probability distributions. One way to perform thresholding is to simply correct for these prior probabilities, by dividing the output of the network for each class, then seen as posterior probabilities, by the estimated prior probabilities. In our case, the imbalance was not completely eliminated, so the thresholds were calculated as the ratio between the original class counts and those after oversampling.

As a final step, the output of the model that performed best was used to test the whole pipeline. Word images from the test set, that were classified as scientific names, were assigned IRI e.g., `ex:taxon1`. The names were linked and semantically enriched using terms from the ontology and transformed to the RDF format. The code can be found online.[1]

5.5.2 Results and Discussion

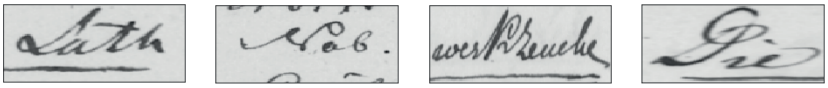
Table 5.2 summarises the final classification results for each network. Due to a large class imbalance, precision and recall were used to assess the predictive power of the classifier. Reporting accuracies would be misleading, as they would portray the underlying distribution, rather than the predictive power of the model (if the model would always predict "*other*", it would be a bad predictor for the task, but the accuracy would be 93%, as the "*other*" class accounts for 93% of the data).

[1] `https://github.com/lisestork/asa-species-names`

Table 5.2: Classification precision, recall and $F1$ results for each network. Support indicates the number of actual occurrences of that class in the given subset.

Method	Class	Precision	Recall	F1-score	Support
1. CNN	Genus	0.80	0.78	0.79	36
	Species	0.64	0.97	0.77	33
	Author	0.78	0.78	0.78	32
	Other	1.00	0.97	0.98	525
	avg / total	0.82	0.77	0.80	626
2. CNN–MLP	Genus	0.85	0.81	**0.83**	36
	Species	0.81	0.88	**0.84**	33
	Author	0.78	0.78	0.78	32
	Other	0.99	0.99	0.99	525
	avg / total	0.96	0.96	0.96	626
3. CNN–MLP–BLSTM	Genus	0.86	0.89	**0.88**	36
	Species	0.94	0.91	**0.92**	33
	Author	0.78	0.88	**0.82**	32
	Other	1.00	0.99	0.99	525
	avg / total	0.98	0.97	0.98	626

Bold $F1$ scores indicate statistical superiority over $F1$ scores for that same class within the cell of the preceding method. The table indicates that the BLSTM produced the highest average $F1$ scores for each class. The addition of the BLSTM layer specifically increases precision and recall scores for the author names. This makes sense; without context these appear similar to regular words. The input of centroid data to the network does not have an effect on the recall or precision of author names, but does increase precision for the retrieval of species names. Figure 5.4 shows 4 images from the test set that were misclassified. While both the CNN and CNN–MLP network misclassify most of the same word images, the output of the CNN–MLP–BLSTM is quite different. Image **(a)** and **(b)** were both misclassified by the networks without the BLSTM layer, but were correctly classified by the final model. Image **(a)** for example, was classified as "*species*", while actually being labelled as an author name. Visually, it resembles a species name; it is underlined and appears in a similar position on the page. Without context of other words it is challenging to correctly classify such images without proper historical knowledge of the domain. Image **(b)** was misclassified as "*other*", but correctly identified as an author name in the BLSTM model, most likely due to the visual characteristics of the word image that is left adjacent. On the other hand, image **(c)** and **(d)** are together misclassified as a species name and its author by the BLSTM network, while they were correctly classified by the other networks. Examining the images, we see that they are adjacent and visually resemble these classes (capitals, underlining).



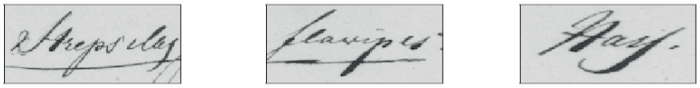
(a) $y = 2$, $\hat{y} = 1$ **(b)** $y = 2$, $\hat{y} = 0$ **(c)** $y = 3$, $\hat{y} = 1$ **(d)** $y = 3$, $\hat{y} = 2$

Figure 5.4: Four misclassified examples. Classlabels relate to those discussed in table 5.1

In Table 5.3, we present retrieval scores for the identification of complete scientific names from field book pages. A python script parsed the recognised species elements from the test set, and connected them together using the NHC-Ontology. A total of 27 out of 36 species names were retrieved, with an *F1* score of 0.86. Interestingly, there were no false-positives among the final predictions. Figure 5.5 shows one of the correctly classified scientific names. The final RDF dataset can be queried through our online SPARQL endpoint.[1]

Table 5.3: Final classification precision, recall and $F1$ results for the detection of scientific names.

Method	Class	Precision	Recall	F1-score	Support	Total
CNN–MLP–BLSTM	Scientific names	1.0	0.75	0.86	27	36



(a) $y = 0$, $\hat{y} = 0$ **(b)** $y = 1$, $\hat{y} = 1$ **(c)** $y = 2$, $\hat{y} = 2$

Figure 5.5: A correctly classified scientific species name: **(a)** *Genus* **(b)** *Species* **(c)** *Person*

5.6 Conclusions and Future Work

In this chapter we show that we can accurately identify and classify components of handwritten species observation records from different features: visual structural features, position and context. We show that our methods are applicable even though the dataset contains four authors with different handwriting styles and different processes of recording their species observations. A major challenge of working with handwritten text is its irregularity. Our results show that we can mitigate this challenge by building up multiple pieces of evidence for classification by learning from multiple features. Each of the different

[1] `http://makingsense.liacs.nl/rdf4j-server/repositories/SN`, can be queried through a query editor such as: `https://yasgui.org/`

features we examine in our model adds information and improves the overall results. In addition, as the results are extracted and structured in RDF format as part of the process, they are immediately available for search and comparison with other archives - historical or present day.

The dataset used for experiments in this chapter is part of the same expedition archive (the NC collection, see Subsection 2.3.2). Although we represent multiple authors and styles, the next step would be to demonstrate the generic nature of our results by analysing biodiversity records from other expeditions. Once we establish that, we will extend our methods to identify other common classes from biodiversity data, for example, locations, dates and anatomical entities.

It is our aim to integrate the new methods with established methods for automated handwriting recognition, using a fruitful dialogue between our system and a system for HTR, in which the hypotheses (highest confidence values) of both systems work together for the transcription and semantic annotation of named entities in manuscripts.

CHAPTER 6

Classification of Biological Illustrations

"Ludwig Wittgenstein once said that names are the only things that exist in the world. Maybe that's true, but the problem is that as time passes by, names do not remain the same—even if they don't change."

– Victor Pelevin

Historically, naturalists created hand-drawn scientific illustrations for the documentation of new species. These scientific illustrations often contain captions with handwritten *historical names*, as is demonstrated in Figure 2.7 and 6.1, which can be used to compare the illustrations with other collection objects and online resources. However, many names are unpublished or obsolete within today's taxonomy, and are therefore difficult to understand. By linking these illustrations to contemporary binomial names and taxonomies, they can be understood in their context.

In this chapter, we aim to answer research question **Q.4**: *How can we use automated methods for the extraction of knowledge from archives of NHCs?*, by researching how we can automatically classify—or help domain experts to classify—biological illustrations.

6.1 Introduction

Over the last 250 years, a large number of zoological species have been observed and documented through the use of scientific illustrations. Research into these scientific illustrations is complicated by several challenges. First, most illustrations are stored in museum repositories and archives that are not disclosed for generic use. Digitisation projects are currently ongoing worldwide to address this challenge, but as of now, most collections remain offline (7). Second, illustrations published as online digital collections can be used for research, but are often published with doubtful or no identifications (unique labels), which are required to study the illustrations. Finally, the identification of an organism from a photograph or illustration, using the system of biological classification,

is a complex and delicate task, even for domain experts (42). Automated methods can significantly reduce the time and effort required by scholars to identify and classify the images. Easy access to taxonomic classifications of illustrations facilitates research into the historical abundance, range and variation of species.

Automated Classification of Scientific Illustrations Automated species identification is a much researched problem within the computer vision and pattern recognition domain, but, to the best of our knowledge, no approaches have been described to deal with the wealth of detailed scientific illustrations (examples shown in Figure 2.7). Reasons could be that samples are small due to the nature of the data—many rare species have been depicted in small quantities—and because numerous institutions have yet to start with the digitisation of their collections (131).

Photographs and illustrations of species are quite distinct, As described in Subsection 2.1.2. To illustrate the differences between photographic and illustration data, three depictions and two photographs of the species *Lepas (Anatifa) anserifera Linnaeus, 1767* can be observed in Figure 6.1 and 6.2.

[1] https://bijzonderecollecties.uva.nl/gedeelde-content/beeldbanken/iconographia.html

(a) (b) (c)

Figure 6.1: Scientific illustrations from the Iconographia Zoologica[1] of *Lepas (Anatifa) anserifera Linnaeus, 1767*, with handwritten (historical) name *Anatifa laevis Bruguière, 1789* (best viewed in colour). (a) Species within shell, (b) shell of species, (c) species without shell. Images free of known restrictions under copyright law (Public Domain Mark 1.0).

The dissimilarity of the two modes demands training or fine-tuning a (pre-trained) classifier on the illustrations. However, this is a non-trivial task due to a couple of challenges, of which we name two:

1. for classifying zoological illustrations, only small samples from a subset of species described in modern taxonomy are available for training, and these samples are smaller for rarer species (see also **Chall.6** to **Chall.8**). Therefore, standard supervised classification models overfit the training data, and do not capture the totality of the problem.

2. testing a classification model on a test-set does not guarantee its value 'in the wild'. Due to various factors, there is always a divergence that affects performance: a change in distribution or differences in feature space (132). Illustrations, for instance, vary in use of materials, drawing style and method, and can portray zoological species unknown to the model.

(a) **(b)**

Figure 6.2: Photographs of the species *Lepas (Anatifa) anserifera Linnaeus, 1767 (Goose Barnacle)*, taken from iNaturalist.[1] (best viewed in colour). (a) Observation © David R.[2] (b) Observation © mervyngreening.[3] Images are licensed under CC BY-NC 4.0.

Approach Below we formulate a research approach that copes with the aforementioned challenges. To address the first challenge, our approach uses a non-standard learning strategy called zero-shot learning (ZSL). With ZSL, it is possible to exploit data from auxiliary data sources to form semantic descriptions of classes, which can help to classify images from *unseen* classes: classes that are not observed by the classifier during training, and hereby to push the boundaries of automated recognition for a specific problem. Such a classifier is also more flexible to deal with new definitions of classes, and therefore better

[1] https://www.inaturalist.org/
[2] https://www.inaturalist.org/observations/25983495
[3] https://www.inaturalist.org/observations/34793791

formulates real world conditions. This is especially useful for biological taxonomy, where the solution space is large, new class definitions can be introduced, and old ones can be revisited. To avoid overfitting, our approach additionally exploits image representations learned from another task—the recognition of zoological photographs—to extract meaningful features for our task (129). Moreover, we use a biological taxonomy as a label hierarchy for training (through Hierarchical Prototype Loss (HPL)), and hereby have access to a larger number of labelled examples for groups higher up the label hierarchy. We evaluate our approach on the ZICE dataset, that we introduce in this paper. The dataset consists of 14,502 zoological illustrations of 7973 species from the animal kingdom, and is formed by consolidating data used and managed by the biodiversity research community.[1]

To address the second challenge, we evaluate the trained model "in the wild", on a dataset collected under different conditions. To this end, our approach uses a second independent collection of illustrations without annotations, to analyse the final species embedding model.

Our contribution is threefold:

1. We introduce the Zoological Illustration and Class Embedding (ZICE) dataset constructed from real-world data. It consists of: (i) 14,502 biological illustrations of 7973 species from the animal kingdom, with labels organised hierarchically, and (ii) class embeddings from 3 different sources - a hierarchy (taxonomy), historical texts and photographs.

2. We introduce and evaluate a zero-shot learning (ZSL) approach for fine-grained hierarchical classification. We use the prototypical networks introduced by Snell et al. in (49) and introduce: Fused Prototype (FP), and HPL. Our approach is evaluated on the ZICE dataset.

3. We provide a qualitative analysis of the performance of our ZSL approach in a real-world scenario on an independent verification-set: a collection of 1,088 unlabelled zoological illustrations, collected during a historical biodiversity expedition (16).

The rest of this chapter is organised as follows. In Section 6.2 we discuss related work on automated species classification and ZSL. We discuss the data in Section 6.3, the methodology in Section 6.4, the experimental setting in Section 6.5 and the experiments in Section 6.6. We close the paper with an analysis and discussion of the results in Section 6.7, and our conclusions in Section 6.8.

[1] https://geheugen.delpher.nl/nl/geheugen/pages/collectie/Iconographia+Zoologica:+een+papieren+dierenrijk

6.2 Related Work

Below, we discuss datasets related to computer vision and biodiversity, where we briefly mention recent work that leverages contextual information for fine-grained classification, and provide a short survey of the field of ZSL.

Computer Vision and Biodiversity Recognising and identifying species in images is a well researched problem within the computer vision field. Most popular datasets contain classes of animals, (often birds), or plants (60; 133; 134; 135; 136; 137; 138). A citizen science project called iNaturalist (discussed in Subsection 2.3.1), allows users to upload photographs of organism encounters in the wild. Since 2017, a new dataset has been published every year as part of the iNaturalist Competition FGVC6 for fine-grained image classification.[1] Computer vision models trained on such datasets are much better prepared for the automatic identification of species in the wild. Nevertheless, much variation still exists among data captured for various tasks, such as between observation data from iNaturalist, and data collected from motion-triggered camera traps.[2] Recent datasets therefore combine data captured for distinct tasks to model the variation that exists among photographs of species observations (138).

To improve automated classification of species in images, recent work has demonstrated the usefulness of leveraging contextual data for the improvement of classification models, for instance the use of spatio-temporal data often accompanying observations to aid fine-grained classification (139; 140; 141). Moreover, zero-shot learning methodologies allow researchers to leverage contextual information from multimodal sources to calculate measures of similarity between classes (142; 143). Such contextual information can greatly aid a model to distinguish between visually similar classes where small samples are available for training.

In addition to photographs of species, there are examples of models trained for the automated classification of plants in herbaria (144). While a great deal of work is spent capturing often unclear images of species in the wild, a wealth of detailed zoological illustrations are under-utilised. Reasons could be that samples are small, many classes are under-represented, and numerous institutions have yet to start with the digitisation of their collections (131).

[1] https://www.kaggle.com/c/inaturalist-2019-fgvc6
[2] https://github.com/microsoft/CameraTraps

Zero-Shot Learning While standard supervised image classification methods learn to recognise images from classes observed during training, ZSL aims to recognise images from classes not observed during training, $y \in \mathcal{Y}^{ts}$, from examples of classes observed during training, $\mathcal{Y}^{tr}$, by using between-class feature transfer. With a training set $\mathcal{T} = \{(\mathbf{x}_1, y_1), ..., (\mathbf{x}_N, y_N)\}$ where $y \in \mathcal{Y}^{tr}$, and *embedding functions* $\varphi : \mathcal{Y} \rightarrow \tilde{\mathcal{Y}}$ and $\theta : \mathcal{X} \rightarrow \tilde{\mathcal{X}}$, the task is to learn a compatibility function $f : \tilde{\mathcal{X}} \rightarrow \tilde{\mathcal{Y}}$. At test time, the function is used to assign test images to classes from $\mathcal{Y}^{ts}$.

With θ, every image $\mathbf{x}_i \in \mathbb{R}^D$ from $\mathcal{Y}^{tr}$, is embedded in *visual feature space*, $\theta(\mathbf{x}_i) \in \mathbb{R}^M$, called an *image embedding*. Most commonly, θ is a CNN. After training the CNN, the top of the network - often just the softmax layer - is removed and an image embedding function remains.

With φ, every class $y_i \in \{1, ..., K\}$ is mapped to a vector in *semantic embedding space*, $\varphi(y_i) \in \mathbb{R}^M$, called a *class embedding*. The semantic embedding space is either (i) created manually, through class annotations or attributes (50; 51), or (ii) learned from auxiliary information such as taxonomies (145; 146) or texts (147; 148; 149). Attribute embeddings encode whether a certain attribute - from a set of predefined attributes - is present for a specific class. Attribute embeddings can be either binary or continuous, e.g., *{wing: 0.1, red: 0.4, tail: 0.7}* and fall within the interval [0,1]. Learned embeddings are continuous and represent similarities between classes more abstractly. Class embeddings from various sources can be used to complement one another; combining them often results in a higher accuracy (142; 143; 150). Combining class embeddings can be done in different ways, for instance by concatenating the class embeddings or combining compatibility scores. We refer to (143) for an extensive evaluation of class embeddings.

Most common ZSL methods learn either a linear (150; 151; 143; 152) or a non-linear (153; 154) compatibility function between the two feature spaces. Prototypical networks (49) belong to the latter group. They learn deep visual-semantic models, such as DeViSe (151) and Cross-modal transfer (CMT) (154), in which the visual object recognition network is trained to predict the class embedding vector in semantic embedding space, which is learned from auxiliary data. While all methods achieve impressive results on small- and medium-scale datasets, the more realistic variant generalised zero-shot learning (GZSL), that aims to classify both seen and unseen classes, performs poorly for unseen classes (154): the model overfits to seen classes and therefore favours seen over unseen classes at test time. Hence, ZSL models embedded in real world applications should include a method for dealing with this issue.

Datasets have been set up to facilitate progress in the field and demonstrate the possibilities and advantages of zero-shot learning (155; 156; 50). We argue that there is a need for research that analyses the performance of ZSL models on complex real-world data, collected to fulfill a need within a certain domain, e.g., such as for the identification of tree species from remote sensing images (142), for mapping the worlds' biodiversity (60), or for the estimation of species populations and richness (138). Specifically data from domains where the solution space is large and complex, and obtaining labels for training is costly or simply not feasible. When algorithms are evaluated on highly imbalanced large-scale datasets, results are often poor. Xian et al. show that experiments of state-of-the-art zero-shot learning algorithms achieve only $\sim 1.3\%$ top-1 per-class accuracy on the 5,000 least populated classes in ImageNet, and only $\sim 0.4\%$ top-1 accuracy for GZSL (157), where the classifier must choose the correct class from both seen *and* unseen classes.

For an extensive comparison of state-of-the-art of ZSL and GZSL methods and datasets, we point to the work of Xian et al. (157). In our work we use prototypical networks for zero-shot learning because they are state-of-the-art models within the few- and zero-shot learning domain (49).

6.3 The Data

In this section, we discuss the ZICE dataset (see Subsection 6.3.1), used for training, validating and testing our ZSL approach, and an independent verification-set (in Subsection 6.3.2) used to analyse the ZSL results in a real-world scenario (in Section 6.7).

6.3.1 The ZICE Dataset

The Zoological Illustration and Class Embedding (ZICE) dataset contains illustrations, from the Iconographia Zoologica online collection,[1] and class embeddings corresponding to the classes represented in the illustrations.

Illustrations The Iconographia Zoologica is a nineteenth century collection of biological illustrations from the Artis Library of the University of Amsterdam. The collection was formed by three collectors: the well-known collector and naturalist Th. G. van Lidth de Jeude, the zoologist R.T. Maitland and the curator of the shell collection at the Amsterdam Zoo, Abraham Oltman, together with the Amsterdam society *Natura Artis Magistra*. In the twenty-first century, the collection was digitised and labelled with either complete

[1] https://bijzonderecollecties.uva.nl/gedeelde-content/beeldbanken/iconographia.html

binomial species names (genus and specific epithet) or corresponding genera. The full online collection contains over 26,500 pages of zoological illustrations.

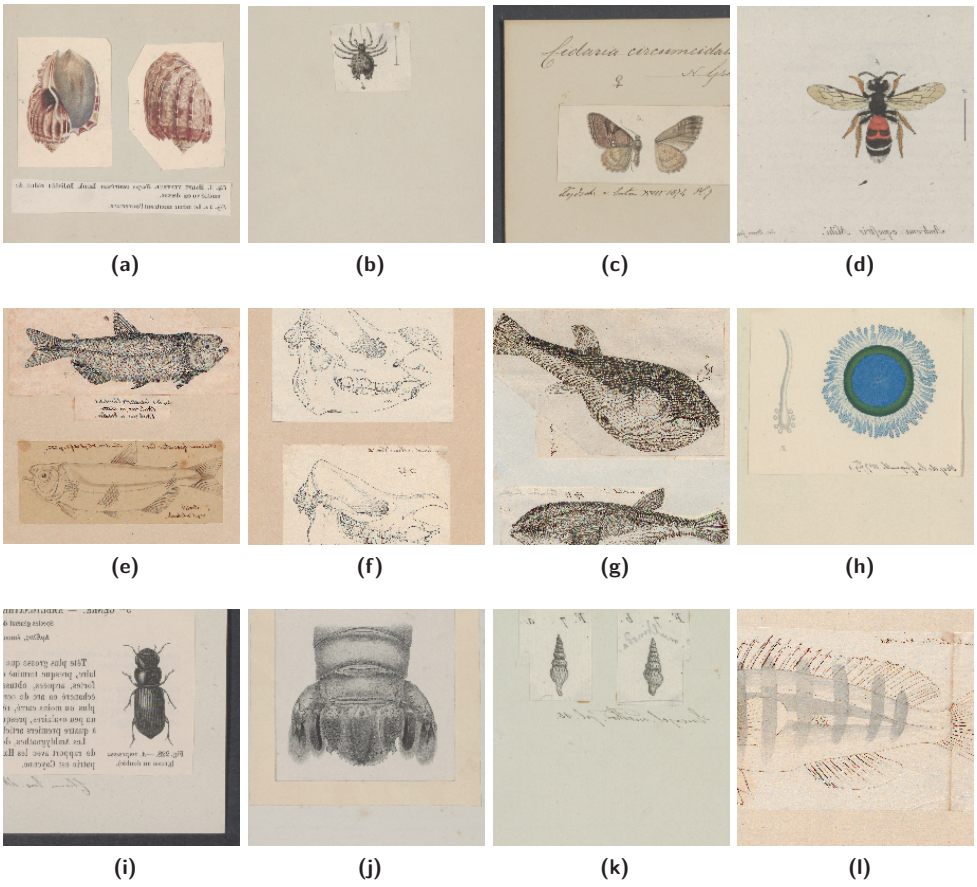


Figure 6.3: Cropped example illustrations from the ZICE train-set (best viewed in colour). Image (f), depicts the skull of a *Rhinosceros unicornis* and image (j) the tail of a *Squilla hoevenii*. Images free of known restrictions under copyright law (Public Domain Mark 1.0)

We have cross-referenced the illustration labels with the June 2018 backbone taxonomy (59) of the GBIF (discussed in Subsection 2.3.1),[1] a central repository for biodiversity occurrence data. For 14,502 illustrations of 7973 species, labels could be cross-referenced directly with GBIF without extra domain expert curation. Matches were only accepted when the names had the status "accepted" in the GBIF taxonomy, as using labels with the status "unaccepted" or "synonym" to train a ZSL model could prove problematic. Some

[1] https://www.gbif.org/

synonyms, for example, refer to both a plant and an animal. As a result, visual features would map to incorrect semantic representations. By the automated matching process, all classes in the ZICE dataset are organised according to a taxonomy. Figure 6.3 shows twelve example illustrations.

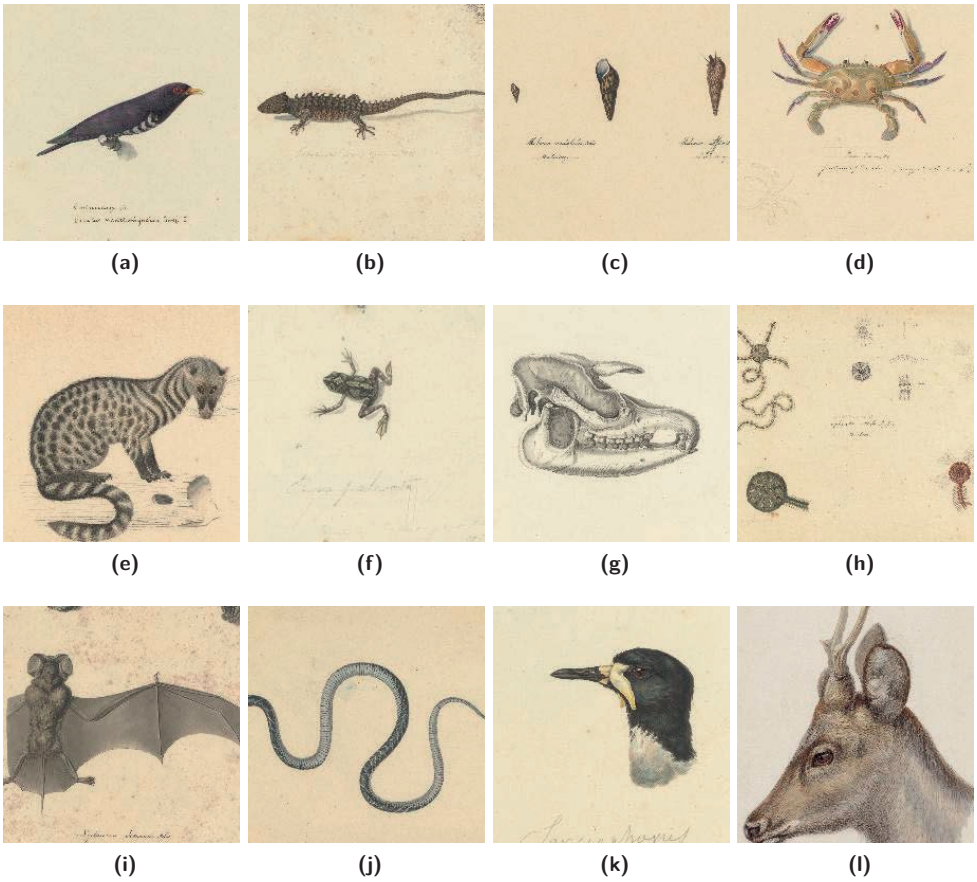


Figure 6.4: Cropped example illustrations from the verification-set (best viewed in colour). Labels are unknown. Images free of known restrictions under copyright law (Public Domain Mark 1.0)

Notation A biological taxonomy can be seen as a tree data structure, in which species are represented as leaf nodes, and parent classes represent their higher classifications based on features shared with other species. In the rest of this paper, we refer to the biological taxonomy by the term *label hierarchy*, and we refer to the various ranks (depths of the tree) by *levels*. The hierarchy consists of seven levels: kingdom, phylum, class, order, family, genus, species (genus + specific epithet). We use $\mathcal{D} = \{(\mathbf{x}_1, y_1, \mathbf{t}_1), ..., (\mathbf{x}_N, y_N, \mathbf{t}_N)\}$ to

refer to the ZICE dataset, where each $\mathbf{x}_i \in \mathbb{R}^D$ represents a D-dimensional feature vector of an image, each $y_i \in \{1, ..., K\}$ represents its species label, where K thus indicates the number of leaf nodes of the label hierarchy, and $\mathbf{t}_i = \begin{bmatrix} t_1, & \ldots, & t_L \end{bmatrix}$ represents its full path of labels, one from each level and ordered from fine-grained to course-grained such that $\mathbf{t}_i[1] = y_i$, and where L indicates the number of levels in the label hierarchy.

Class embeddings To train our ZSL model, we have generated class embeddings whose classes match those from the illustrations. They come from three different sources: (i) the GBIF backbone taxonomy (59), (ii) literature from the BHL (13) and (iii) photographs from the iNaturalist 2018 challenge dataset (60). Information on how these embeddings are produced is given in Section 6.4.

6.3.2 The Verification-Set

For the verification-set, we use 1,088 illustrations from the collection of the Natural Comittee (discussed in Subsection 2.3.2) to evaluate the model in a realistic setting. Example illustrations are presented in Figure 6.4.

6.4 Methodology

In this section, we describe the mathematical formulation of our approach: the ZSL (in Subsection 6.4.1), image embeddings (in Subsection 6.4.2), class embeddings (in Subsection 6.4.3), our method for (i) combining class embeddings: FP (in Subsection 6.4.4), and (ii) for calculating HPL based on the label hierarchy (in Subsection 6.4.5).

6.4.1 Zero-Shot Learning Model

Prototypical networks for few-shot learning, as described in (49), compute M-dimensional class representations $\mathbf{c}_k \in \mathbb{R}^M$ called *class prototypes*. They do so by embedding N_s *support points* $\{(\mathbf{x}_1, y_1), ..., (\mathbf{x}_N, y_N)\} \in \mathcal{S}$ from N_c classes with an embedding function $f_\phi : \mathbb{R}^D \to \mathbb{R}^M$, and taking the per-class average of the resulting embedded support points, see Equation 6.1. In Equation 6.1, $\mathcal{S}_k$ refers to the set of support points for class k, and $\mathbf{c}_k$ refers to its calculated prototype. We further refer to the space $\mathbb{R}^m$ by the term *prototype space*.

$$\mathbf{c}_k = \frac{1}{|\mathcal{S}_k|} \sum_{(\mathbf{x}_i, y_i) \in \mathcal{S}_k} f_\phi(\mathbf{x}_i) \tag{6.1}$$

To train the network, Prototypical Network Loss (PNL) is calculated by mapping a set of N_q *query points*: $\{(\mathbf{x}_1, y_1), ..., (\mathbf{x}_N, y_N)\} \in \mathcal{Q}$ from the same N_c classes to prototype space. In prototype space, distances from the query points to the class prototypes are computed so that, based on a softmax over these distances, a distribution over classes is obtained. Parameters ϕ are learned by minimising the negative log-probability of the true class k via Stochastic Gradient Descent (SGD). The network is trained with mini-batches. Each mini-batch consists of N_c classes, N_q query points and N_s support points, and is called an *episode*.

For ZSL, Snell et al. (49) mention that rather than embedding support points in prototype space, prototypes can be constructed by embedding auxiliary information, e.g., class embeddings in the form of attribute annotations, in prototype space. In their paper they use binary attribute vectors from the CUB-200-2011 dataset (156). They extract features from different crops of the images using a pre-trained model and map them to prototype space using a one-layer linear model. Similarly, they use a one-layer linear model to map the attributes to prototype space and prototypical training proceeds as in the few-shot setting. Rather than relying on one source (such as attributes), we rely on a combination of class embeddings from three distinct sources.

6.4.2 Image Embeddings

We embed images $\mathbf{x} \in \mathcal{X}$ of zoological illustrations in a lower dimensional feature space using a deep CNNs $\theta(x) : \mathcal{X} \rightarrow \tilde{\mathcal{X}}$. We will use θ to refer to the image embeddings. To make sure we don't learn features specific to our dataset (such as an illustrator's mark or a label). We transfer image representations learned from photographs (the *source* dataset) to illustrations (the *target* dataset) (129). We use the inception V3 model (158), and import weights learned on the iNaturalist 2018 competition dataset.[1] For zero-shot learning, image embeddings are often generated using CNNs pre-trained on a source task (e.g., the ImageNet task (128)). The choice of model is crucial as the quality of the image embeddings has a big impact on the performance of the ZSL model. Therefore, we have chosen to use a model that was trained on a task more similar to ours. Xian et al. (157) mention that class overlap between classes from the source and target dataset leads to an unwanted positively biased result. However, our goal is not to compare between various state-of-the-art ZSL methods, but rather to provide insights for training a model that is able to generalise to new data within the target domain.

[1] https://github.com/macaodha/inatcomp2018

6.4.3 Class Embeddings

Below we describe details concerning the embedding functions that map classes $y_i \in \mathcal{Y}$, the set of leaf nodes from the label hierarchy, to vectors $\varphi(y_i) \in \mathbb{R}^M$ in M-dimensional semantic embedding space: $\varphi : \mathcal{Y} \rightarrow \tilde{\mathcal{Y}}$. As each embedding comes from a different domain, all embeddings are l_2-normalised. For brevity, we use φ_k^i to refer to the class embeddings of source i for class k.

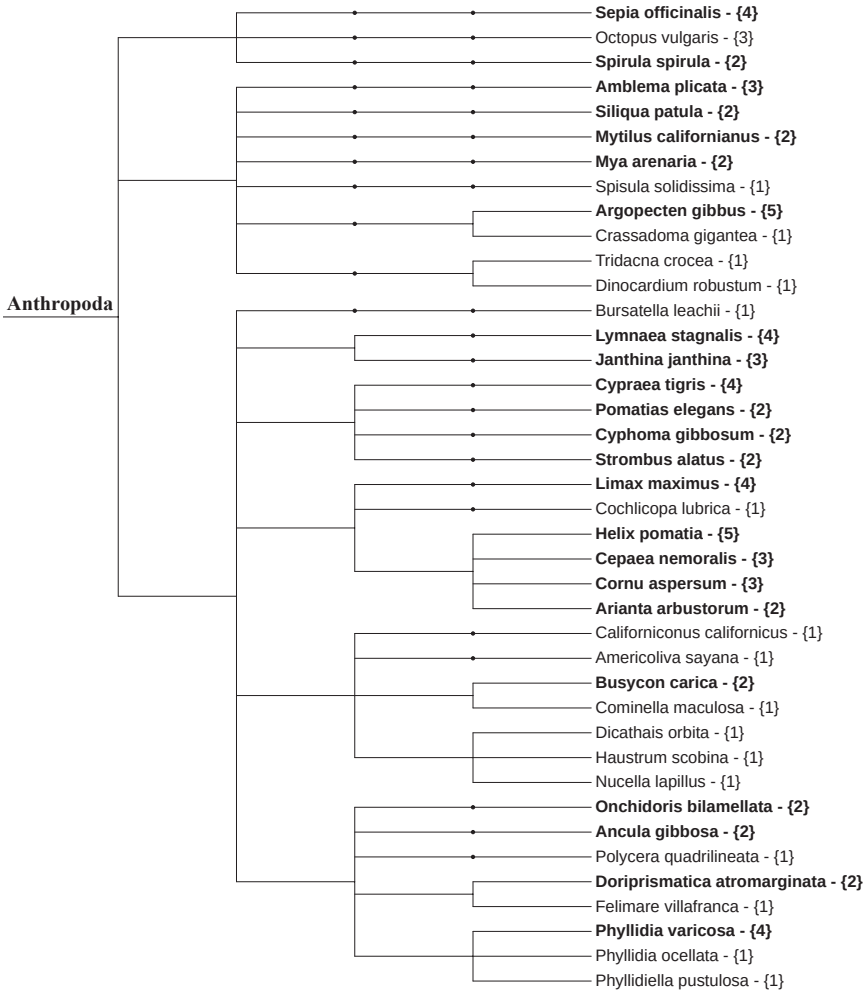


Figure 6.5: A subset of $\mathcal{Y}$ from the ZICE dataset, covering the phylum *Anthropoda*, with the corresponding label hierarchy (from left to right: phylum to species). **Bold** names indicate classes from $\mathcal{Y}^{tr}$, and numbers indicate number of instances within that class.

A hierarchy (φ^h) Through the GBIF backbone taxonomy, we had access to the ground truth list of higher taxon labels for nearly all classes (see Table 6.1 for class statistics). For 53 classes, no (or an incomplete) higher classification was available. Using the deterministic algorithm from Barz et al. (145), we have projected all 7920 classes onto a unit sphere of dimensionality N - where N is the number of classes. The negated dot product between classes on the sphere represents their semantic similarity. This similarity is based on the ratio of overlap between their ground truth list of higher taxon labels—nodes in the hierarchy. Part of the label hierarchy is given in Figure 6.5.

Texts (φ^t) To facilitate semantic search over large textual biodiversity archives, Nguyen et al. have constructed an inventory of name variants and synonyms from a large textual biodiversity corpus (BHL) (159). For this task, they have computed word embeddings from multi-word terms–"chipping sparrows" becomes "chipping_sparrows"–mentioned in the corpus. They compared multiple methods for computing word embeddings: *continuous-bag-of-words* (CBOW) (160), *count-based* (161) and *Global Vectors* (GloVe) (148). From these three, we rely on the 300 dimensional multi-word GloVe embeddings.

Photographs (φ^p) Features in photographs are quite distinct from those in illustrations, but their features capture the semantic similarity of the different classes they represent in a similar way. Hence, we have extracted 2048 dimensional features from the iNaturalist 2018 dataset photographs, using the inception V3 model trained on the corresponding dataset (previously mentioned in Subsection 6.4.2).

6.4.4 Combining Class Embeddings

Below we describe two methods for generating singular class prototypes for prototypical learning (see Subsection 6.4.1) from three distinct embeddings, each with a different dimensionality.

Concatenated Embeddings (CEs) One method that is often employed to combine the different embeddings is concatenation, in which the dimensions of each class embedding (from distinct sources) are concatenated together. This results in one sparse matrix with a large dimensionality. Similarly to Snell et al. (49), we learn a one-layer linear model on top of the concatenated class embeddings φ and on top of the image embeddings θ, mapping them to prototype space.

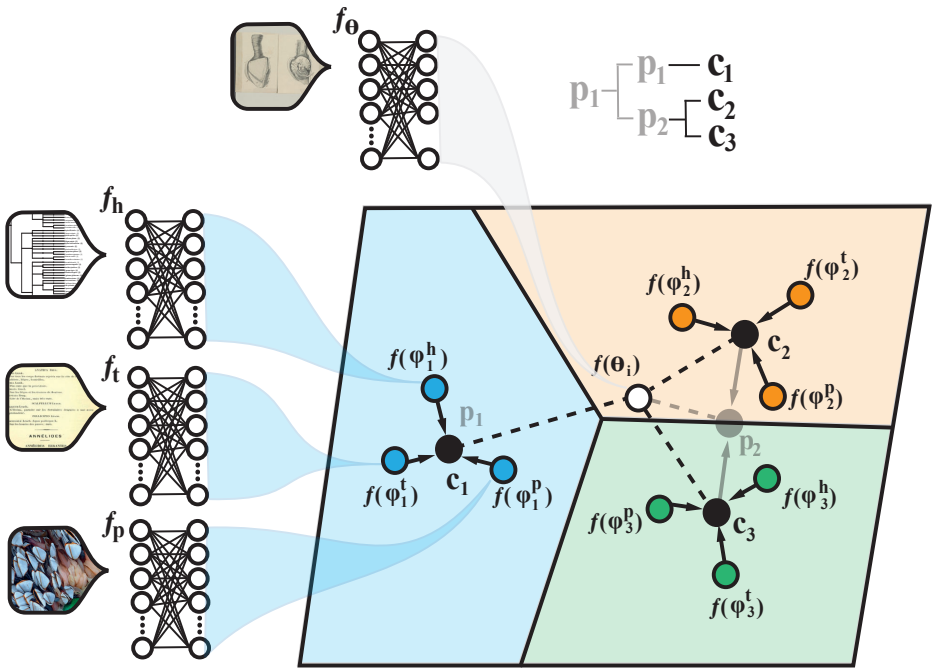


Figure 6.6: FP (best viewed in colour). Figure derived from (49). Features from φ^i (here i is replaced by: a hierarchy (h), texts (t), and photographs (p)) are mapped into prototype space using separate one-layer linear models f_{ϕ^i}, and fused into one prototype per class $\mathbf{c}_k$. To illustrate HPL, example temporary parent-class prototypes $\mathbf{p_k}$ are depicted in transparent grey.

Fused Prototypes (FPs) We implement FPs, see Figure 6.6. Essentially, FPs fuse prototypes from a variable number of multimodal sources into a single prototype per class. We derive our implementation from the prototypical FSL approach. Instead of using support points $s \in \mathcal{S}$, we use $\varphi^i \in \Phi$, the set of class embeddings from distinct sources $\{\varphi^1, ..., \varphi^N\}$. A simple one-layer linear model is learned on top of the feature space of each of the distinct φ^i's as well as the image embeddings θ, mapping both to prototype space. In prototype space, the embedded φ^i's are fused together, similarly to the way support points are fused to form class prototypes for FSL, see Equation 6.2.

$$\mathbf{c}_k = \frac{1}{|\Phi|} \sum_{(\varphi_k^i, y_k) \in \Phi} f_{\phi^i}(\varphi_k^i) \tag{6.2}$$

In that equation, $\mathbf{c}_k$ refers to the class prototype for class k, where N is the number of sources, and $f_i\phi$ refers to the linear model that maps the individual class embeddings from

φ^i to prototype space. We hypothesise that fused prototypes will perform better than concatenated embeddings, as the latter introduce one large sparse input space whereas fused prototypes are optimised from multiple dense input spaces.

6.4.5 Hierarchical Prototype Loss

HPL extends PNL, and is defined as the sum of the losses for each level of the label hierarchy (see Figure 6.5). The loss for a specific level l is calculated by first computing temporary parent-class prototypes $\mathbf{p}_k \in \mathbb{R}^M$ for that level from the set of class prototypes $\mathcal{C} = \{(\mathbf{c}_1, y_1, \mathbf{t}_1), ..., (\mathbf{c}_K, y_K, \mathbf{t}_K)\}$, see Figure 6.6 and Equation 6.3. In the Equation, $\mathcal{C}_k$ refers to the subset of $\mathcal{C}$ containing all prototypes $(\mathbf{c}_i, y_i, \mathbf{t}_i)$ such that $\mathbf{t}_i[l] = k$. As described in Subsection 6.4.1, distances of the query points to the temporary parent-class prototypes are then computed and the loss is calculated over these distances. HPL is calculated by summing the losses for all L levels.

$$\mathbf{p}_k = \frac{1}{|\mathcal{C}_k|} \sum_{(\mathbf{c}_i, y_i, \mathbf{t}_i) \in \mathcal{C}_k} \mathbf{c}_i \tag{6.3}$$

By implementing HPL, we take a multi-granularity approach: we enforce a clearer separation of classes not only for the finest grain, but also for coarser taxonomic groups. As more labels are available for each level higher up in the label hierarchy, this intuitively supports the discovery of more robust features for the classification of coarser classes.

6.5 Experimental Setting

In this section we discuss details regarding the settings of the experiment: the dataset splits (in Subsection 6.5.1), data augmentation (in Subsection 6.5.2), and evaluation criteria (in Subsection 6.5.3).

6.5.1 Dataset Splits

As recommended by (157), we split the classes $\mathcal{Y}$ for training and evaluation based on the number of instances each of them contain. Since our dataset contains so few instances per class, ($n_k \in [1, 283]$, μ: 1.79, σ: 3.93). We have used all classes with n $\geq$ 2 per class for the training set $\mathcal{Y}^{tr}$. Two examples per class is not sufficient to learn a good class representation, but the features of these illustrations are useful for between super-class feature sharing. Moreover, we exploit them for learning representations of classes on a

Table 6.1: Dataset statistics per super-class (phylum) $\mathcal{Y}$: total number of leaf node classes $\mathcal{Y}$ and instances N, the number of leaf node classes per split $\mathcal{Y}^s$, $s \in \{tr, v, ts\}$, the number of instances per split N^s, and the number of leaf node classes per embedding $\mathcal{Y}^{\phi^e}$, $e \in \{h, t, p\}$.

Super-class (phylum)	$\mathcal{Y}^{tot}$	N^{tot}	$\mathcal{Y}^{tr}$	$\mathcal{Y}^{v}$	$\mathcal{Y}^{ts}$	N^{tr}	N^{v}	N^{ts}	$\mathcal{Y}^{\phi^h}$	$\mathcal{Y}^{\phi^t}$	$\mathcal{Y}^{\phi^p}$
Arthropoda	2977	3740	620	1106	1251	1383	1112	1245	2977	218	14
Chordata	2903	7358	1281	870	752	5736	878	744	2901	2050	475
Mollusca	1423	2384	488	464	471	1449	464	471	1385	475	40
Cnidaria	179	299	58	47	74	178	48	73	179	88	5
Echinodermata	111	180	36	33	42	105	33	42	111	62	10
Annelida	109	171	32	44	33	94	44	33	106	61	3
Porifera	59	79	17	17	25	37	17	25	59	11	-
Platyhelminthes	56	75	9	38	9	28	38	9	55	9	-
Bryozoa	45	67	10	12	23	32	12	23	45	23	-
Brachiopoda	37	38	1	23	13	2	23	13	37	2	-
Nematoda	18	24	4	9	5	10	9	5	18	8	-
Rotifera	17	20	2	7	8	5	7	8	17	12	-
Ctenophora	14	33	5	2	7	24	3	6	14	6	-
Nemertea	6	8	2	3	1	4	3	1	4	4	-
Sipuncula	5	6	1	3	1	2	3	1	5	-	-
Acanthocephala	4	5	1	1	2	2	1	2	4	2	-
Nematomorpha	2	6	1	1	-	5	1	-	2	1	-
Onychophora	2	2	-	2	-	-	2	-	0	2	-
Cephalorhyncha	1	1	-	1	-	-	1	-	0	1	-
Chaetognatha	1	2	1	-	-	2	-	-	1	1	-
Entoprocta	1	1	-	1	-	-	1	-	0	1	-
Animalia	3	3	-	2	1	-	2	1	0	3	-
Total	7973	14502	2569	2684	2717	9098	2702	2702	7920	3040	547

higher taxonomic level, since a larger number of instances are available higher up the label hierarchy. All remaining classes with n = 1 were equally distributed over the validation set $\mathcal{Y}^v$, and the test set $\mathcal{Y}^{ts}$. Table 6.1 shows dataset statistics per super-class. Since not all of the classes were represented in each source (GBIF, BHL and iNaturalist), each embedding (φ^h, φ^t, and φ^p respectively) represents a subset of $\mathcal{Y}$. However, together they span the totality of classes $\mathcal{Y}$. The super-class *Animalia* is used for classes that are not assigned to a phylum.

6.5.2 Data Augmentation

For training, we used image embeddings extracted from augmented versions of all images, in order to increase the ability of the classifier to generalise the classification with respect to the data. Before cropping all images, the largest side of each image was first resized to 300. During resizing, we kept the aspect ratio identical to the original image. 2048-dimensional features were extracted by applying the pre-trained Inception V3 model to crops (middle, upper left, upper right, lower left and lower right) of each resized original illustration and its horizontally flipped version. Crops containing only white space or text were manually discarded.

6.5.3 Evaluation Criteria

In our experimental ZSL results (Subsection 6.6.2) we report two accuracy metrics: top-k accuracy and hierarchical accuracy@k.

Top-k accuracy Flat top-1 accuracy does not always sufficiently portray the classifier's capabilities. When the solution space is large, it is valuable for domain experts to obtain top-k predictions, as exemplified later in Figure 6.8. We therefore report top-k accuracy, $k \in \{1, 2, 5, 10\}$. This metric is computed by the percentage of images for which the correct label is among the top k predictions.

Hierarchical accuracy@k For our task, classifying an illustration of a *Boiga nigriceps* as a *Boiga dendrophila* - both tree snakes - is less problematic than classifying it as a *Procyon lotor*, a common raccoon. In the former case, the classifier has learnt important coarse features specific to tree snakes, and has provided researchers with a partially incorrect, but valuable classification nonetheless. For each illustration, we would therefore like to shed light on the accuracy of the entire label path from the label hierarchy. Hence, we additionally report *hierarchical accuracy*. Hierarchical @k precision is sometimes used as a metric for hierarchical datasets (151). We report a new metric that we deem more informative in our context: *average per-level accuracy*, or *hierarchical accuracy*. It is

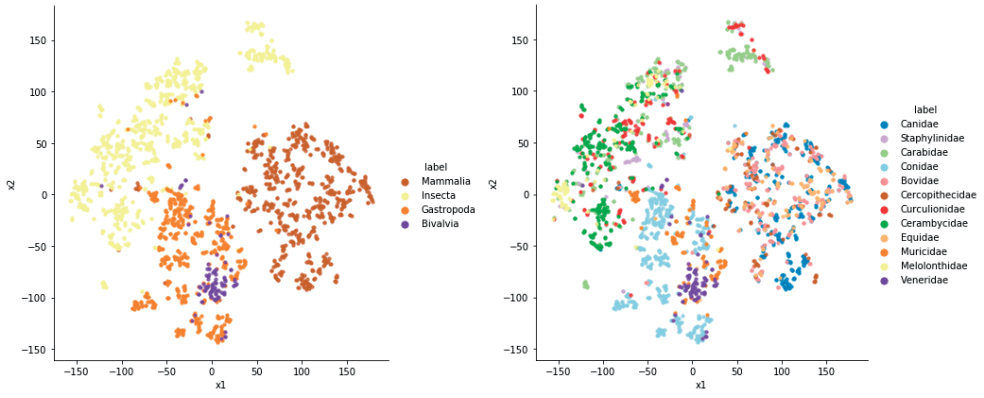
computed by calculating the accuracy for each level in the label hierarchy and averaging over these, see formula 6.4. In formula 6.4, L refers to the number of levels for which we have labels and l to a specific level l.

$$\textit{Hierarchical acc} = \sum_{l=1}^{L} \frac{n \text{ correct preds in } l}{n \text{ samples in } l} \tag{6.4}$$

Additionally, we report accuracies for labels k levels up the label hierarchy, where $k \in \{1, 2, 3\}$, thus referring to the accuracy for labels one, two and three levels up the label hierarchy respectively.

6.6 Experimental Results

The following section is divided as follows: first we evaluate the image embeddings (**Task 1**) in a supervised classification setting (Subsection 6.6.1), after which we evaluate each of the elements of our zero-shot learning approach (Subsection 6.6.2): the class embeddings (**Task 2**), combining class embeddings (**Task 3**), hierarchical prototypical loss (**Task 4**), and an analysis of the final network, which incorporates results from Task 2-4 (**Task 5**).



(a) Class labels **(b)** Family labels

Figure 6.7: t-Distributed Stochastic Neighbour Embedding (t-SNE) plots showing image embeddings of images from the ZICE dataset (should be viewed in colour). Plot (a) shows class level labels and (b) family level labels. Family labels come from a selection of 12 families of which the binomial name was not present in the iNaturalist 2018 dataset. The t-SNE algorithm was run for 5,000 iterations with perplexity 20.

6.6.1 Supervised Classification and Visualisation

For Task 1, we selected image embeddings from the set of species that is disjoint from the set of species represented in the iNaturalist 2018 dataset (on which the embedding function was trained), so as to obtain a deeper insight into the ability of image embedding function to find generic features. From this selection, we again selected a subset for classification and visualisation purposes: the 12 most populated classes from the *family* level (two levels up the label hierarchy).

We show per-class, micro, macro and weighted average precision and recall results for a Support Vector Machine (SVM) trained on the subset, see Table 6.2. Additional to family labels (Table 6.2, 2nd column), we show higher-taxon labels from the class level (Table 6.2, 1st column). The weighted average alters the macro metric to account for label imbalance. The support column indicates the number of actual occurrences of that class in the given subset.

The SVM was trained using a stratified 80%-20% split for the train and test-set, respectively. Note that the classification results serve to provide an insight into the quality of the features rather than the difficulty of our task. For visualisation, we show a t-SNE (162) visualisation of the subset with family labels (see Figure 6.7 (b)). Also here, we present higher-taxon labels from the class level (see Figure 6.7 (a)).

Looking at Figure 6.7, we see that same-class image embeddings are visibly clustered. However, classes within certain taxon groups overlap, for instance, families within the class *Mammalia*, see the classes of Figure 6.7 (b) that are coloured brown in Figure 6.7 (a). This effect is reflected in Table 6.2 (see **bold** text): the image embeddings from only one of four families subsumed under the class *Mammalia* can be classified correctly (*Canidae*, with 100% recall). From the classifications and the precision value (48%) we find that image embeddings from other classes subsumed under the class Mammalia are also classified as Canidae, and thus a large part of the brown cluster from Figure 6.7 is classified as the family Canidae (dog-like carnivores).

The results of Task 1 show us that the features learned from the iNaturalist 2018 task are not sufficiently specific to properly classify all fine-grained classes in our task well. Therefore, further improving the image embeddings would improve zero-shot learning results, although the inter-class variation of species within certain taxon groups can be quite small. Some species within the order *Coleoptera* (beatles), for instance, can only be accurately identified after a close inspection of their genitalia (163). Visualisation of the features can give an indication up to which grain the features within specific taxon groups are sufficiently informative for proper classification.

Table 6.2: Classification precision, recall and $F1$ results for **Task 1** in % (rounded off to whole integers) for a SVM trained on the image embeddings belonging to 12 families (also visualised in Figure 6.7 (b)). Support indicates the number of actual occurrences of that class in the given subset. The top-1 per-class average accuracy is 43.58%.

Class	Family	Precision	Recall	$F1$	Support
Mammalia	**Bovidae**	**0**	**0**	**0**	**19**
Mammalia	**Canidae**	**48**	**100**	**65**	**33**
Insecta	Carabidae	44	74	56	27
Insecta	Cerambycidae	56	85	68	26
Mammalia	**Cercopithecidae**	**0**	**0**	**0**	**9**
Gastropoda	Conidae	87	98	92	41
Insecta	Curculionidae	0	0	0	14
Mammalia	**Equidae**	**0**	**0**	**0**	**12**
Insecta	Melolonthinae	100	22	36	9
Gastropoda	Muricidae	67	55	60	11
Insecta	Staphylinidae	0	0	0	10
Bivalvia	Veneridae	82	90	86	10
	micro avg	60	60	60	221
	macro avg	40	44	38	221
	weighted avg	46	60	50	221

6.6.2 Fine-Grained Zero-Shot Learning

All prototypical networks were trained using the Adam optimisation algorithm from pytorch.[1] Episodes for training were comprised of $N_c = 50$, $N_q = 1$ and $N_s = 0$, similar to a balanced mini-batch of size 50. The validation loss was monitored during training and if, for 10 iterations, the loss did not decrease, the learning rate was decreased with a factor of 0.5. We tuned hyper-parameters using hyper-parameter optimisation–tree-structured parzen estimators–and ended up with a learning rate of 10^{-4} and a weight decay of 10^{-5}. Early stopping on the validation loss was used to determine the optimal number of epochs for training. For each model, five different networks were trained. As a statistical test for comparing classifiers we used the McNemar test (164) for each classifier pair for all predictions of 5 runs accumulated. It is a test that works well for testing statistical significance when dealing with paired nominal data for comparing classifiers trained, validated and tested multiple times on the same splits of a dataset. **Bold** numbers indicate statistical superiority over other values within that column and cell (which separates tasks). A final model was trained, again 5 times, with the configuration that we found to work best. The last row of Table 6.3 indicates accuracy values for the majority guess, where the model simply always predicts the majority class.

[1] `https://pytorch.org/docs/stable/optim.html`

Table 6.3: ZSL classification results in % for **Task 2, 3, 4** and **5**. The 50-way classification accuracy for the final model was 35.53%, calculated by averaging results over 6,000 randomly drawn episodes.

					top-k acc $\mathcal{Y}^{ts}$				**Hierarchical acc@k $\mathcal{Y}^{ts}$**			
Task	**Method**	φ^h	φ^t	φ^p	*1*	*2*	*5*	*10*	*1*	*2*	*3*	*avg*
Task 2	N/A	✓	✗	✗	**2.29**	**4.12**	**8.9**	**15.34**	**5.93**	13.23	43.74	36.38
		✗	✓	✗	0.41	0.66	1.14	1.72	0.72	1.22	7.33	12.53
		✗	✗	✓	0.55	0.85	1.47	2.15	1.03	2.81	15.29	18.26
	FP	✓	✓	✗	2.13	3.89	**8.79**	**15.11**	5.51	13.56	43.21	35.96
		✓	✗	✓	**2.50**	**4.26**	**8.91**	**15.26**	**6.05**	**14.24**	**45.69**	**36.85**
		✗	✓	✓	0.53	0.84	1.45	2.06	1.04	2.02	9.41	13.50
		✓	✓	✓	**2.42**	**4.29**	**9.10**	**15.37**	**5.98**	**14.22**	**45.09**	**36.70**
Task 3	CE (baseline)	✓	✓	✓	2.09	**4.05**	**8.96**	**15.54**	5.45	13.42	**44.76**	36.41
	FP	✓	✓	✓	**2.42**	**4.29**	**9.10**	**15.37**	**5.98**	**14.23**	**45.09**	**36.70**
Task 4	FP + PNL	✓	✓	✓	**2.42**	**4.29**	**9.10**	**15.37**	**5.98**	14.23	45.09	36.70
	FP + HPL	✓	✓	✓	2.12	3.88	**8.88**	**15.03**	**6.23**	**15.71**	**51.10**	**39.35**
Task 5	Final model	✓	✗	✓	2.77	4.74	9.64	16.02	6.94	16.65	50.71	39.67
	Majority guess	-	-	-	0.04	0.07	0.19	0.37	2.85	3.26	21.87	18.66

Evaluation (Task 2, 3 and 4) First, Table 6.4 presents results for Task 2, which show the performance of the networks trained, validated and tested with embeddings from each unique source separately, and additionally each combination of the three distinct embeddings E. In order for the results to be comparable between all combinations, we used the totality of $\mathcal{Y}$ to train, validate and test the networks, despite the fact that each φ^i spans only a subset of classes from $\mathcal{Y}$ (see the last row of Table 6.1). In case a class k was not represented in φ^i, the dimensions for φ^i_k were set to zero. In this context, the results inform us, first and foremost, about the contribution of each embedding to the overall accuracy (Table 6.3, Task 2, last row). We discuss each embedding separately.

φ^h is the most complete and informative embedding. φ^t spans many classes (3040 out of 7973), but appears less informative. The prototypical network trained with φ^t performs better than the majority guess for the top-k acc metric, but φ^t seems to harm the learning ability of the network when used in combination with other embeddings. This could be due to a myriad of factors. We believe the two most likely factors are that (i) the embedding is better suited for finding synonyms between taxon terms - as similar species are described similarly, and, (ii) that some names in the BHL are ambiguous: referring to one species in the historical texts, while they refer to another in modern taxonomy. Particularly, any historical unpublished name could have been published today as a different species. Matching them with sources from a modern taxonomy could therefore be problematic. Finally, the network trained with φ^p shows improvement over the majority guess, and φ^p complements φ^h, as the network trained with $\{\varphi^h, \varphi^p\}$ improves over the accuracy of the model trained with just $\{\varphi^h$ (see Table 6.3, Task 2, row 1 and 5), specifically the hierarchical acc@2 (13.23% to 14.24%) and @3 (43.74% to 45.69%). We hypothesise that if we increase the number of instances and fine-grained classes used to generate φ^p, results could be improved further.

Second, Table 6.3 presents results for Task 3: combining class embeddings. CE represents the baseline model: it is comparable to the method used by Snell et al. (49) for zero-shot learning. Results for Task 3 show us that by using our fused prototypes (FP) formulation, we can increase the top-1 accuracy from 2.09% to 2.42% (see Table 6.3, Task 3). Such an increase is non-trivial. As the test-set contains an instance per class, with a total of 2702 classes (on the finest grain), an increase of 0.33% for the top-1 accuracy equals the capability of the classifier to correctly classify illustrations from an *additional* 9 unseen classes from different parts of the biological taxonomy. Fused prototypes also induce a higher hierarchical accuracy @1 and @2 (from 5.45% to 5.98% and 13.42% to 14.23%, respectively). When class embeddings from additional (informative) sources are used, we

Table 6.4: ZSL classification results in % for **Task 5** on the test-set per super-class (phylum).

			top-k acc $\mathcal{Y}^{ts}$				**Hierarchical acc@k** $\mathcal{Y}^{ts}$			
Super-class (phylum)	N_{tr}	N_{ts}	*1*	*2*	*5*	*10*	*1*	*2*	*3*	*avg*
Chordata	5736	744	4.7	7.39	14.65	24.06	14.65	53.36	81.05	50.22
Mollusca	1449	471	3.4	6.16	11.89	20.59	29.3	47.56	73.25	47.77
Arthropoda	1383	1245	1.61	2.97	6.59	10.6	15.74	60.88	80.0	50.1
Cnidaria	178	73	8.22	9.59	16.44	30.14	19.18	31.51	41.1	29.86
Echinodermata	105	42	4.76	7.14	9.52	21.43	9.52	11.9	33.33	19.05
Annelida	94	33	0.0	0.0	0.0	0.0	0.0	0.0	3.03	1.21
Porifera	37	25	0.0	8.0	8.0	16.0	4.0	8.0	44.0	20.0
Bryozoa	32	23	0.0	0.0	0.0	8.7	0.0	4.35	4.35	3.48
Platyhelminthes	28	9	0.0	0.0	0.0	0.0	0.0	0.0	11.11	4.44
Ctenophora	24	6	0.0	0.0	33.33	33.33	0.0	0.0	0.0	3.33
Nematoda	10	5	20.0	20.0	40.0	40.0	20.0	40.0	40.0	32.0
Rotifera	5	8	0.0	0.0	0.0	12.5	0.0	0.0	0.0	0.0
Nemertea	4	1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Sipuncula	2	1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Brachiopoda	2	13	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Acanthocephala	2	2	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Animalia	0	1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Per super-class average	534.76	158.94	2.51	3.6	8.26	12.79	6.61	15.15	24.19	15.38
Per leaf-node average	9098	2702	2.96	4.96	9.96	16.65	7.11	17.10	52.26	40.05

anticipate that this effect which we discuss in Subsection 6.4.4 will become more evident: the value of using fused prototypes over concatenated embeddings will increase.

Third, Table 6.5 gives results for Task 4, which show that using HPL improves the average hierarchical accuracy significantly - from 36.70% to 39.35%. However, a decrease is measured for the top-1 and top-2 accuracy: from 2.42% to 2.12% and 4.29% to 3.88% respectively. This effect demonstrates intra super-class variation of taxon groups, as it appears that learning better coarser features slightly complicates the classification of some fine-grained taxon groups.

Table 6.5: Generalised zero-shot learning (GZSL) classification results in % for final model

	top-k acc $\mathcal{Y}^{ts}$				**Hier. acc@k** $\mathcal{Y}^{ts}$		
Method	*1*	*2*	*5*	*10*	*1*	*2*	*avg*
GZSL	0.04	0.21	1.24	3.25	4.47	16.03	38.19
M. guess	0.01	0.03	0.06	0.13	2.85	3.26	18.66

Final results (Task 5) A final model was trained 5 times using the best configuration - $\{\varphi^t, \varphi^p\}$, PNL and HPL. Although implementing HPL decreases the top-1 and top-2 accuracy, a substantial increase of the average hierarchical accuracy was measured. We therefore chose to implement it in the final model.

Table 6.3 (Task 5) shows per-network averaged results for the final model on the test-set, and Table 6.4 gives results for the final model's best network, detailed per super-class. Table 6.4 serves to provide a deeper insight into the trained network. Evidently, illustrations from some super-classes were not recognised at all due to their limited contribution to the training of the network–visible from the column avg N_{ts}–as most feature sharing occurs within super-classes. For reason of comparison we add the results for the leaf node level (species).

On top of these results, Table 6.5 details results for GZSL. The top-k accuracies for GZSL are poor: during classification, a network trained for ZSL tends to favour seen classes over unseen classes (154). Logically, GZSL does not affect the average hierarchical accuracy by much, as seen and unseen classes share parent-classes (see Figure 6.8).

Finally, we present and discuss four example images from the test-set with their top-5 predictions (and corresponding confidence values), see Figure 6.8. Image (a) and (b) have good top-5 predictions: the top-1 prediction of image (a) is incorrect (the classifier is most confident about the label *Brachirus macrolepis*, while the correct label is *Brachirus panoides*), but the top-1 prediction is correct up to the fine-grained genus level: *Brachirus*.

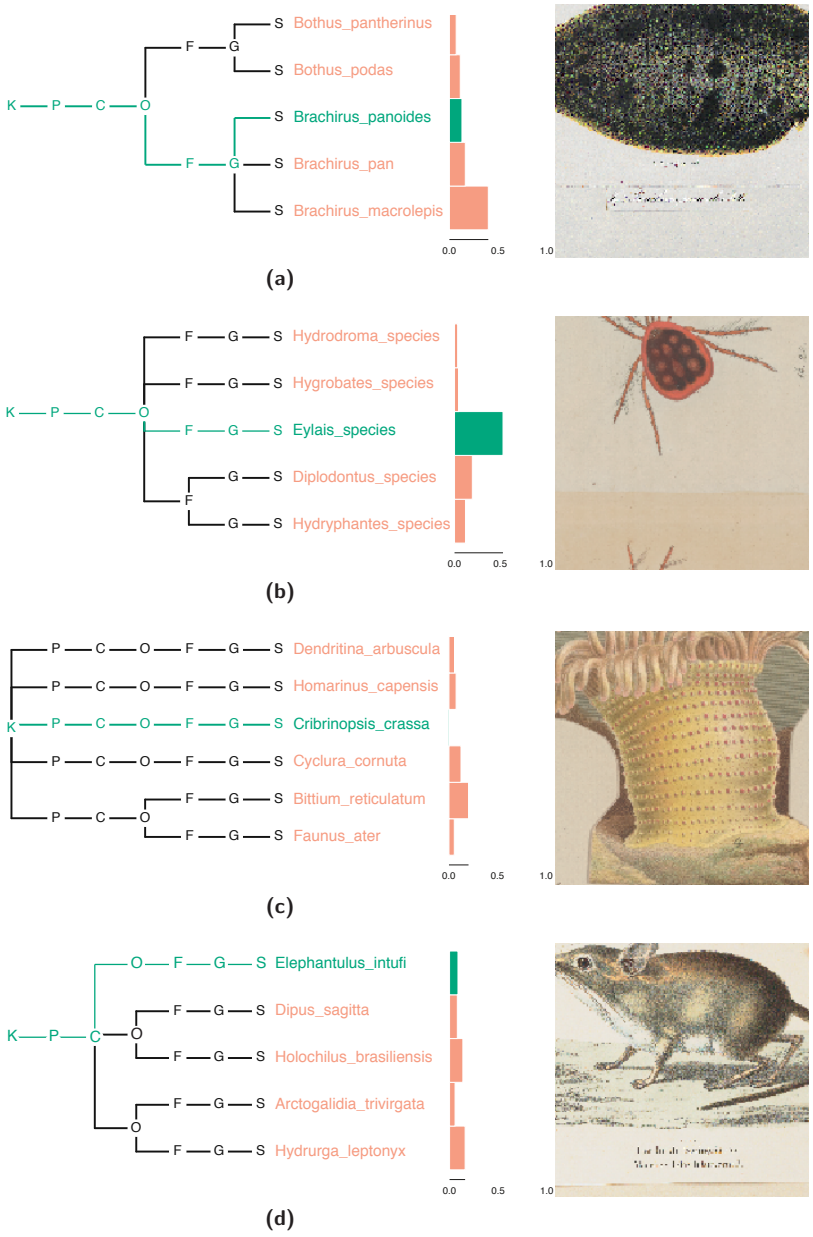


Figure 6.8: Top 5 predicted classes and their confidence values for two example test images (best viewed in colour). Labels are organised hierarchically (K: kingdom to S: species) to show the diversity of predictions and how close–in the label hierarchy–the classifier is to the real label. For image (c) the correct label was not among the top 5 predictions (therefore 6 are shown). Green paths, labels and confidence bars denote correct labels. Orange confidence bars indicate incorrect predictions.

Moreover, the top-3 predictions are all correct up to the genus level. For image (b), the top-1 prediction is correct, and the remaining predictions are from the same correct order.

The third image (c) has poor predictions, as (i) the correct label is not among the top 5 predictions, and (ii) almost all predictions are from a different phylum. Interestingly, however, the top-2 predictions (the *Bittium reticulatum* and *Cyclura cornuta*) have something in common with the correct species (*Cribrinopsis crassa*): they share its most salient feature - their skin is covered with small tubercles.

Lastly, for the fourth image (d) the correct label (*Elephantulus intufi*) belongs to the order *Macroscelidea* (Elephant shrew), and the other predictions belong to the orders (from top to bottom): *Rodentia* (Rodents) and *Carnivora* (Carnivores). The two predictions from the *Rodentia* order are two different *mice* species (*Dipus sagitta* and *Holochilus brasiliensis*. Elephant shrew visually resemble mice. Interestingly, the most salient feature that would allow a classifier to distinguish between a mouse and an elephant shrew, is cut off from the illustration: its long trunk-like nose, which resembles an elephant's trunk. It is therefore good to consider that cropping the image at its center in a standardised way can cause the loss of information that is vital for proper classification.

6.7 Analysis and Discussion

Standard supervised classification offers limited solutions to deal with the full scope of the problem presented above. ZSL models are better suited to deal with limited data (small samples for only a subset of classes from the domain). For instance, Table 6.5 shows that 20 *Anthropod* species could be correctly classified without any training examples, from their similarity to 620 other seen *Anthropod* species. We note that this shows an important gain: the labelling of these illustrations by domain experts is costly, and does not necessarily guarantee high-quality annotations, due to the complex nature of species classification (42). Especially prototypes optimised according to the label hierarchy can be exploited in an expert support system to guide experts in the identification process.

In practice, it can be a real challenge to transfer results to real-world scenarios. We provide two telling examples. First, Table 6.5 shows us that with GZSL, seen classes are favoured over unseen classes during classification. In real-world applications, methods are required that deal with this issue. If not, a classifier will often prefer classes from $\mathcal{Y}^{tr}$ over $\mathcal{Y}^{ts}$ for classification. Second, using a trained network in real-world applications can prove problematic due to a domain shift between datasets. Our verification-set, that we have presented in Subsection 6.3.2, serves to illustrate this issue. When using the final species

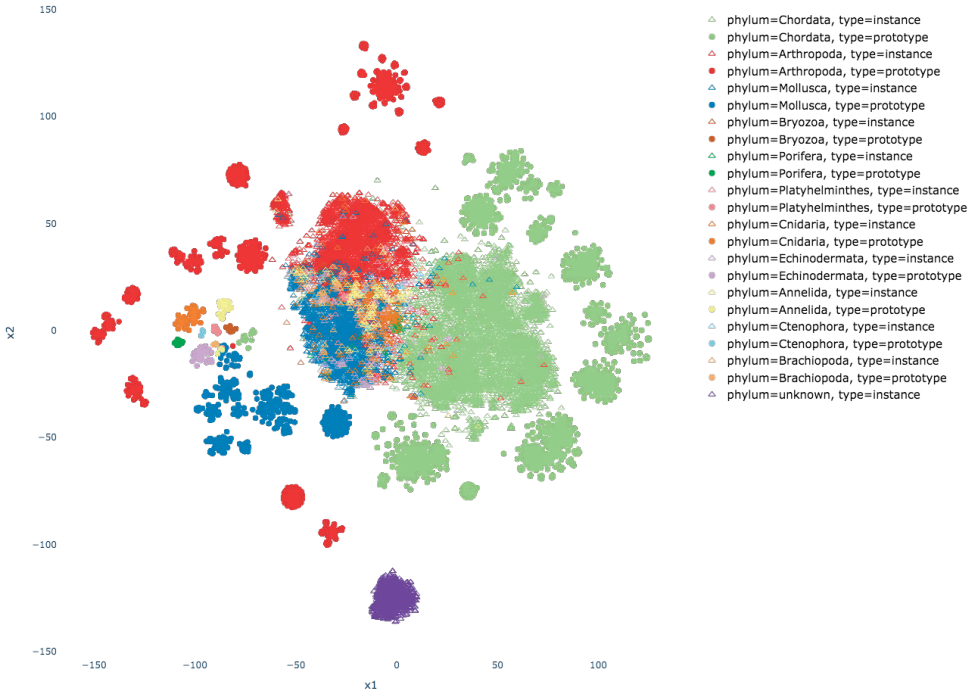


Figure 6.9: A t-SNE plot showing all prototypes (closed circles) and instances (open triangles), from the 12 most populated phyla, embedded by the final prototypical network (should be viewed in colour). Instances from the verification-set (bottom cluster) are indicated by the label 'unknown'. Note that t-SNE does not accurately preserve distances between clusters. The t-SNE algorithm was run for 5,000 iterations with perplexity 100.

embedding model for classification of the verification-set, all instances are classified as species of *Anthropods*, although it contains illustrations from a variety of phyla (among which Chordates and Annelids, see Figure 6.4). The t-SNE visualisation, see Figure 6.9, allows us to hypothesise about the results. The visualisation shows instances from the verification-set (depicted as purple triangles, see bottom cluster), as well as instances and prototypes from the ZICE dataset (all other open triangles and closed circles respectively), all embedded by the species embedding model. The species embedding model appears to have mapped instances from the verification-set to a different manifold than those from the ZICE dataset. Consequently, instances from the verification-set manifold are classified as *Anthropods*, as its prototypes are closest (see the red prototype clusters in Figure 6.9). We hypothesise that both datasets must come from a distinct marginal probability distribution. Most likely, this domain shift is the result of differences in paper

types, sketching techniques and materials.

Overcoming the aforementioned issues is key, but we argue that ZSL and hierarchical learning methods (methods that exploit the label hierarchy) are fundamental for problem domains such as the one described here: where labelling of images is expensive, but where, at the same time, auxiliary data sources contain a wealth of domain knowledge maintained by a community of experts.

6.8 Conclusions

In this chapter we have analysed the problem of classifying species in zoological illustrations. For this purpose, we have introduced a dataset, with many classes and few samples, and an independant (unlabelled) verification-set, both representative of the problem domain.

From the experimental results, we conclude that auxiliary data sources have allowed us to push the boundaries of automated recognition for this specific problem: illustrations from 80 classes, that contained zero example instances for training, could be classified correctly. We furthermore conclude that our model improves over the baseline classifier. Compared with the baseline, our FP implementation allowed us to classify instances from an additional 9 unseen fine-grained classes. Moreover, implementing HPL increased the average hierarchical accuracy substantially (from 36.41% to 39.35%). Finally, from the results of the analysis of the verification-set in Section 6.7, we show the complexity of our task. Aside from the depicted illustrations, there are other differences between the digital images that impact the predictive capabilities of the model. The illustrators' technique, the physical drawing materials and the chosen perspectives change significantly between illustrators. In order for our zero-shot learning model to function well in an application, domain adaptation methods should be employed to align domain marginal probability distributions (132) between datasets, and therefore make the model illustrator-invariant.

Coming back to our main problem description, we conclude that biodiversity datasets, storing domain knowledge and auxiliary data, can be exploited to develop models for classification (especially when small samples are available for training). These models can then serve as decision support systems for biodiversity researchers to help classify the historical and present-day scientific illustrations from various species of living organisms, which reside underutilised in natural history museums globally.

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.

Bibliography

[1] A. MacGregor, *Naturalists in the field: collecting, recording and preserving the natural world from the fifteenth to the twenty-first century*. Brill, 2018.

[2] P. L. Farber, *Finding order in nature: the naturalist tradition from Linnaeus to E.O. Wilson*. JHU Press, 2000.

[3] M. W. Holmes, T. T. Hammond, G. O. Wogan, R. E. Walsh, K. LaBarbera, E. A. Wommack, F. M. Martins, J. C. Crawford, K. L. Mack, L. M. Bloch, *et al.*, "Natural history collections as windows on evolutionary processes," *Molecular Ecology*, vol. 25, no. 4, pp. 864–881, 2016.

[4] M. Schilthuizen and F. Vonk, *Wie Wat Bewaart, die Heeft Wat*. Spectrum Uitgeverij Unieboek, 2020.

[5] T. Monquil-Broersen and E. Gassó, *Van onschatbare waarde: 200 jaar Naturalis*. Amsterdam University Press, 2021.

[6] A. H. Ariño, "Approaches to estimating the universe of natural history collections data," *Biodiversity Informatics*, vol. 7, no. 2, pp. 81–92, 2010.

[7] B. P. Hedrick, J. M. Heberling, E. K. Meineke, K. G. Turner, C. J. Grassa, D. S. Park, J. Kennedy, J. A. Clarke, J. A. Cook, D. C. Blackburn, S. V. Edwards, and C. C. Davis, "Digitization and the future of natural history collections," *BioScience*, vol. 70, no. 3, pp. 243–251, 2020.

[8] V. Blagoderov, I. J. Kitching, L. Livermore, T. J. Simonsen, and V. S. Smith, "No specimen left behind: industrial scale digitization of natural history collections," *ZooKeys*, vol. 209, pp. 133–146, July 2012.

[9] 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," *Journal on Computing and Cultural Heritage (JOCCH)*, vol. 8, no. 1, pp. 1–11, 2015.

[10] R. C. Baird, "Leveraging the fullest potential of scientific collections through digitisation.," *Biodiversity Informatics*, vol. 7, no. 2, 2010.

[11] E. Hyvönen, "Publishing and using cultural heritage linked data on the semantic web," *Synthesis Lectures on the Semantic Web: Theory and Technology*, vol. 2, no. 1, pp. 1–159, 2012.

[12] V. Petras, T. Hill, J. Stiller, and M. Gäde, "Europeana–a search engine for digitised cultural heritage material," *Datenbank-Spektrum*, vol. 17, no. 1, pp. 41–46, 2017.

[13] N. E. Gwinn and C. Rinaldo, "The biodiversity heritage library: sharing biodiversity literature with the world," *IFLA journal*, vol. 35, pp. 25–34, March 2009.

[14] TDWG Interest Group, "Collection descriptions." `https://www.tdwg.org/community/cd/`, 2017. last accessed: 26-11-2020.

[15] Smithsonian Institution Archives, "The field book project." `https://siarchives.si.edu/about/field-book-project`, 2010. last accessed: 30-12-2020.

[16] A. Weber, "Collecting colonial nature: European naturalists and the netherlands indies in the early nineteenth century," *BMGN-Low Countries Historical Review*, vol. 134, no. 3, 2019.

[17] M. Moyle, J. Tonra, and V. Wallace, "Manuscript transcription by crowdsourcing: Transcribe bentham," *Liber Quarterly*, vol. 20, no. 3-4, 2011.

[18] K. A. Mika, J. De Veer, and C. Rinaldo, "Crowdsourcing natural history archives: Tools for extracting transcriptions and data," *Biodiversity Informatics*, vol. 12, pp. 58–75, 2017.

[19] A. Weber, M. Ameryan, K. Wolstencorft, L. Stork, M. Heerlien, and L. Schomaker, "Towards a digital infrastructure for illustrated handwritten archives," in *Digital Cultural Heritage* (M. Loannides, ed.), vol. 10605 of *Information Systems and Applications, incl. Internet/Web, and HCI*, pp. 155–166, Springer International Publishing, April 2018.

[20] R. E. Drinkwater, R. W. Cubey, and E. M. Haston, "The use of optical character recognition (ocr) in the digitisation of herbarium specimen labels," *PhytoKeys*, no. 38, p. 15, 2014.

[21] P. Kahle, S. Colutto, G. Hackl, and G. Mühlberger, "Transkribus-a service platform for transcription, recognition and retrieval of historical documents," in *2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR)*, vol. 4, pp. 19–24, IEEE, 2017.

[22] J. A. Sánchez, V. Bosch, V. Romero, K. Depuydt, and J. De Does, "Handwritten text recognition for historical documents in the transcriptorium project," in *Proceedings of the First International Conference on Digital Access to Textual Cultural Heritage*, pp. 111–117, 2014.

[23] L. Schomaker, "Design considerations for a large-scale image-based text search engine in historical manuscript collections," *It - Information Technology*, vol. 58, pp. 80–88, April 2016.

[24] T. M. Rath, R. Manmatha, and V. Lavrenko, "A search engine for historical manuscript images," in *Proceedings of the 27th annual international ACM SIGIR conference on Research and development in information retrieval*, pp. 369–376, 2004.

[25] H. S. Baird, V. Govindaraju, and D. P. Lopresti, "Document analysis systems for digital libraries: Challenges and opportunities," in *International Workshop on Document Analysis Systems*, pp. 1–16, Springer, 2004.

[26] M. Ameryan and L. Schomaker, "A high-performance word recognition system for the biological fieldnotes of the natuurkundige commissie.," in *Proceedings of the International Conference Collect and Connect (COLCO): Archives and Collections in a Digital Age*, pp. 92–103, 2020.

[27] L. Stork, A. Weber, J. van den Herik, A. Plaat, F. Verbeek, and K. Wolstencroft, "Large-scale zero-shot learning in the wild: Classifying zoological illustrations," *Ecological Informatics*, vol. 62, p. 101222, 2021.

[28] L. Stork, A. Weber, E. G. Miracle, F. Verbeek, A. Plaat, J. van den Herik, and K. Wolstencroft, "Semantic annotation of natural history collections," *Journal of Web Semantics*, vol. 59, 2019. 100462.

[29] J. B. Kennedy, R. Kukla, and T. Paterson, "Scientific names are ambiguous as identifiers for biological taxa: their context and definition are required for accurate data integration," in *International Workshop on Data Integration in the Life Sciences* (B. Ludäscher and L. Raschid, eds.), vol. 3615 of *Lecture Notes in Computer Science*, (Berlin, Heidelberg), pp. 80–95, Springer, 2005.

[30] D. Nadeau and S. Sekine, "A survey of named entity recognition and classification," *Lingvisticae Investigationes*, vol. 30, no. 1, pp. 3–26, 2007.

[31] A. Kiryakov, B. Popov, I. Terziev, D. Manov, and D. Ognyanoff, "Semantic annotation, indexing, and retrieval," *Web Semantics: Science, Services and Agents on the World Wide Web*, vol. 2, pp. 49–79, December 2004.

[32] M. D. Wilkinson, M. Dumontier, I. J. Aalbersberg, G. Appleton, M. Axton, A. Baak, N. Blomberg, J.-W. Boiten, L. B. da Silva Santos, P. E. Bourne, J. Bouwman, A. J. Brookes, T. Clark, M. Crosas, I. Dillo, O. Dumon, S. Edmunds, C. T. Evelo, R. Finkers, A. Gonzalez-Beltran, A. J. Gray, P. Groth, C. Goble, J. S. Grethe, J. Heringa, P. A. 't Hoen, R. Hooft, T. Kuhn, R. Kok, J. Kok, S. J. Lusher, M. E. Martone, A. Mons, A. L. Packer, B. Persson, P. Rocca-Serra, M. Roos, R. van Schaik, S.-A. Sansone, E. Schultes, T. Sengstag, T. Slater, G. Strawn, M. A. Swertz, M. Thompson, J. van der Lei, E. van Mulligen, J. Velterop, A. Waagmeester, P. Wittenburg, K. Wolstencroft, J. Zhao, and B. Mons, "The fair guiding principles for scientific data management and stewardship," *Scientific data*, vol. 3, no. 1, pp. 1–9, 2016.

[33] A. M. Lister and C. C. R. Group, "Natural history collections as sources of long-term datasets," *Trends in Ecology & Evolution*, vol. 26, pp. 153–154, January 2011.

[34] B. J. Cardinale, J. E. Duffy, A. Gonzalez, D. U. Hooper, C. Perrings, P. Venail, A. Narwani, G. M. Mace, D. Tilman, D. A. Wardle, *et al.*, "Biodiversity loss and its impact on humanity," *Nature*, vol. 486, no. 7401, pp. 59–67, 2012.

[35] A. Thomer, G. Vaidya, R. Guralnick, D. Bloom, and L. Russell, "From documents to datasets: A mediawiki-based metod of annotating and extracting species observations in century-old field notebooks," *ZooKeys*, vol. 209, pp. 235–253, July 2012.

[36] J. Wieczorek, D. Bloom, R. Guralnick, S. Blum, M. Döring, R. Giovanni, T. Robertson, and D. Vieglais, "Darwin core: an evolving community-developed biodiversity data standard," *PloS one*, vol. 7, p. e29715, January 2012.

[37] E. G. Miracle, L. Stork, A. Weber, M. Ameryan, and K. Wolstencroft, "Natuurkundige commissie archives online." doi:10.1163/isbn.9789004336865, 2020. Leiden, the Netherlands: Brill.

[38] S. Müller-Wille, "Names and numbers:"data" in classical natural history, 1758–1859," *Osiris*, vol. 32, no. 1, pp. 109–128, 2017.

[39] E. G. Miracle, "On whose authority? temminck's debates on zoological classification and nomenclature: 1820–1850," *Journal of the History of Biology*, vol. 44, pp. 445–481, January 2011.

[40] W. G. Berendsohn, "The concept of "potential taxa" in databases," *Taxon*, vol. 44, pp. 207–212, May 1995.

[41] A. MacGregor, ed., *Naturalists in the Field*. Leiden, the Netherlands: Brill, 2018.

[42] G. E. Austen, M. Bindemann, R. A. Griffiths, and D. L. Roberts, "Species identification by experts and non-experts: comparing images from field guides," *Scientific Reports*, vol. 6, p. 33634, 2016.

[43] D. Maynard, K. Bontcheva, and I. Augenstein, "Natural language processing for the semantic web," *Synthesis Lectures on the Semantic Web: Theory and Technology*, vol. 6, no. 2, pp. 1–194, 2016.

[44] I. Goodfellow, Y. Bengio, A. Courville, and Y. Bengio, *Deep learning*, vol. 1. MIT press Cambridge, 2016.

[45] R. C. Gonzales and R. E. Woods, "Digital image processing," 2002.

[46] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *nature*, vol. 521, no. 7553, p. 436, 2015.

[47] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," *Communications of the ACM*, vol. 60, no. 6, pp. 84–90, 2017.

[48] D. E. Rumelhart, G. E. Hinton, and R. J. Williams, "Learning representations by back-propagating errors," *nature*, vol. 323, no. 6088, pp. 533–536, 1986.

[49] J. Snell, K. Swersky, and R. Zemel, "Prototypical networks for few-shot learning," *Advances in neural information processing systems*, vol. 30, pp. 4077–4087, 2017.

[50] C. H. Lampert, H. Nickisch, and S. Harmeling, "Attribute-based classification for zero-shot visual object categorization," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 36, no. 3, pp. 453–465, 2013.

[51] V. Ferrari and A. Zisserman, "Learning visual attributes," in *Advances in Neural Information Processing Systems*, no. 20, pp. 433–440, 2007.

[52] M. Buda, A. Maki, and M. A. Mazurowski, "A systematic study of the class imbalance problem in convolutional neural networks," *Neural Networks*, vol. 106, pp. 249–259, 2018.

[53] G. Antoniou, P. Groth, F. Van Harmelen, and R. Hoekstra, *A semantic web primer*. MIT press, 2004.

[54] D. K. Ahern, I. M. Braun, M. E. Cooley, and T. W. Bickmore, "Oncology informatics: behavioral and psychological sciences," in *Oncology informatics*, pp. 231–251, Elsevier, 2016.

[55] N. Guarino, D. Oberle, and S. Staab, "What is an ontology?," in *Handbook on ontologies*, pp. 1–17, Springer, 2009.

[56] T. R. Gruber, "Knowledge acquisition," *A translation approach to portable ontology specifications*, vol. 5, no. 199-220, pp. 10–1006, 1993.

[57] A. Hogan, E. Blomqvist, M. Cochez, C. d'Amato, G. de Melo, C. Gutierrez, J. E. L. Gayo, S. Kirrane, S. Neumaier, A. Polleres, R. Navigli, A.-C. Ngonga Ngomo, S. Rashid M., A. Rula, L. Schmelzeisen, J. Sequeda, S. Staab, and A. Zimmerman, "Knowledge graphs," *arXiv preprint arXiv:2003.02320*, 2020.

[58] The Global Biodiversity Information Facility, "Gbif: The global biodiversity information facility (year) what is gbif?." `https://www.gbif.org/what-is-gbif`, 2020.

[59] GBIF Secretariat, "Gbif backbone taxonomy." `https://hosted-datasets.gbif.org/datasets/backbone/2018-06-20/`, 2018.

[60] G. Van Horn, O. Mac Aodha, Y. Song, Y. Cui, C. Sun, A. Shepard, H. Adam, P. Perona, and S. Belongie, "The inaturalist species classification and detection dataset," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 8769–8778, IEEE, 2018.

[61] T. Causer and M. Terras, "'many hands make light work. many hands together make merry work': transcribe bentham and crowdsourcing manuscript collections," in *Crowdsourcing our cultural heritage*, pp. 57–88, Ashgate Farnham, 2014.

[62] V. De Boer, M. Hildebrand, L. Aroyo, P. De Leenheer, C. Dijkshoorn, B. Tesfa, and G. Schreiber, "Nichesourcing: harnessing the power of crowds of experts," in *International Conference on Knowledge Engineering and Knowledge Management*, pp. 16–20, Springer, 2012.

[63] C. Dijkshoorn, M. H. Leyssen, A. Nottamkandath, J. Oosterman, M. C. Traub, L. Aroyo, A. Bozzon, W. Fokkink, G.-J. Houben, H. Hovelmann, L. Jongma, J. van Ossenbruggen, G. Schreiber, and J. Wielemaker, "Personalized nichesourcing: Acquisition of qualitative annotations from niche communities.," in *UMAP Workshops*, 2013.

[64] M. Baechler, A. Fischer, N. Naji, R. Ingold, H. Bunke, and J. Savoy, "Hisdoc: historical document analysis, recognition, and retrieval," in *Digital humanities–international conference of the alliance of digital humanities organizations (ADHO)*, 2012.

[65] A. Fischer, A. Keller, V. Frinken, and H. Bunke, "Lexicon-free handwritten word spotting using character hmms," *Pattern Recognition Letters*, vol. 33, no. 7, pp. 934–942, 2012.

[66] J. P. Philips and N. Tabrizi, "Historical document processing: Historical document processing: A survey of techniques, tools, and trends," *arXiv preprint arXiv:2002.06300*, 2020.

[67] L. Parilla and J. Blase, "The value of flexibility on long-term value of grant funded projects," *D-Lib Magazine*, vol. 21, no. 9/10, 2015.

[68] S. Nakasone and C. Sheffield, "Descriptive metadata for field books: Methods and practices of the field book project," *D-Lib Magazine*, vol. 19, p. 1, December 2013.

[69] T. Robertson, M. Döring, R. Guralnick, D. Bloom, J. Wieczorek, K. Braak, J. Otegui, L. Russell, and P. Desmet, "The gbif integrated publishing toolkit: facilitating the efficient publishing of biodiversity data on the internet," *PloS one*, vol. 9, August 2014. e102623.

[70] A. P. Giotis, G. Sfikas, B. Gatos, and C. Nikou, "A survey of document image word spotting techniques," *Pattern Recognition*, vol. 68, pp. 310–332, 2017.

[71] T. M. Rath and R. Manmatha, "Word spotting for historical documents," *International Journal of Document Analysis and Recognition (IJDAR)*, vol. 9, no. 2-4, pp. 139–152, 2007.

[72] N. Naji and J. Savoy, "Etude comparative de l'efficacité du dépistage de l'information dans des manuscrits médiévaux," in *Actes 11e Journées internationales d'analyse statistique des données textuelles JADT 2012*, pp. 753–766, 2012.

[73] A. H. Toselli, L. A. Leiva, I. Bordes-Cabrera, C. Hernández-Tornero, V. Bosch, and E. Vidal, "Transcribing a 17th-century botanical manuscript: Longitudinal evaluation of document layout detection and interactive transcription," *Digital Scholarship in the Humanities*, vol. 33, no. 1, pp. 173–202, 2018.

[74] A. H. Toselli, V. Romero, M. Pastor, and E. Vidal, "Multimodal interactive transcription of text images," *Pattern Recognition*, vol. 43, no. 5, pp. 1814–1825, 2010.

[75] V. Romero, A. H. Toselli, and E. Vidal, *Multimodal interactive handwritten text transcription*, vol. 80. World Scientific, 2012.

[76] S. Colutto, P. Kahle, H. Guenter, and G. Muehlberger, "Transkribus. a platform for automated text recognition and searching of historical documents," in *2019 15th International Conference on eScience (eScience)*, pp. 463–466, IEEE, 2019.

[77] A. Caceres, A. Weber, and L. Schomaker, "Monk in practice: Indexing heterogeneous handwritten collections," in *7th Digital Humatities Benelux 2020*, (Leiden, The Netherlands), 2020.

[78] E. Hyvönen, E. Heino, P. Leskinen, E. Ikkala, M. Koho, M. Tamper, J. Tuominen, and E. Mäkelä, "Warsampo data service and semantic portal for publishing linked open data about the second world war history," in *The Semantic Web. Latest Advances and New Domains* (H. Sack, E. Blomqvist, M. d'Aquin, C. Ghidini, S. P. Ponzetto, and C. Lange, eds.), (Cham), pp. 758–773, Springer International Publishing, 2016.

[79] V. de Boer, M. van Rossum, J. Leinenga, and R. Hoekstra, "Dutch ships and sailors linked data," in *International Semantic Web Conference (ISWC 2014)* (P. Mika, T. Tudorache, A. Bernstein, C. Welty, C. Knoblock, D. Vrandečić, P. Groth, N. Noy, K. Janowicz, and C. Goble, eds.), vol. 8796 of *Lecture Notes in Computer Science*, (Cham), pp. 229–244, Springer International Publishing, October 2014.

[80] A. Meroño-Peñuela, A. Ashkpour, L. Rietveld, and R. Hoekstra, "Linked humanities data: The next frontier? a case-study in historical census data," in *Proceedings of the Second International Workshop on Linked Science 2012 - Tackling Big Data: in conjunction with the International Semantic Web Conference (ISWC2012)*, vol. 951, (Boston, MA), 2012.

[81] J. Kahan, M.-R. Koivunen, E. Prud'Hommeaux, and R. R. Swick, "Annotea: an open rdf infrastructure for shared web annotations," *Computer Networks*, vol. 39, no. 5, pp. 589–608, 2002.

[82] C. Dijkshoorn, V. De Boer, L. Aroyo, and G. Schreiber, "Accurator: Nichesourcing for cultural heritage," *Computing Research Repository*, 2017. abs/1709.09249.

[83] S. Ebert, M. Liwicki, and A. Dengel, "Ontology-based information extraction from handwritten documents," in *2010 12th International Conference on Frontiers in Handwriting Recognition*, pp. 483–488, IEEE, 2010.

[84] C. Adak, B. B. Chaudhuri, and M. Blumenstein, "Named entity recognition from unstructured handwritten document images," in *12th IAPR Workshop on Document Analysis Systems (DAS), 2016*, pp. 375–380, IEEE, 2016.

[85] J. I. Toledo, S. Sudholt, A. Fornés, J. Cucurull, G. A. Fink, and J. Lladós, "Handwritten word image categorization with convolutional neural networks and spatial pyramid pooling," in *Joint IAPR International Workshops on Statistical Techniques in Pattern Recognition (SPR) and Structural and Syntactic Pattern Recognition (SSPR)*, pp. 543–552, Springer, 2016.

[86] M. Carbonell, M. Villegas, A. Fornés, and J. Lladós, "Joint recognition of handwritten text and named entities with a neural end-to-end model," in *2018 13th IAPR International Workshop on Document Analysis Systems (DAS)*, pp. 399–404, IEEE, 2018.

[87] M. Greene and D. Meissner, "More product, less process: Revamping traditional archival processing," *The American Archivist*, vol. 68, no. 2, pp. 208–263, 2005.

[88] M. F. Desnoyers, "When is a collection processed?," *The Midwestern Archivist*, vol. 7, no. 1, pp. 5–23, 1982.

[89] TDWG Interest Group, "Minimum information about a digital specimen." `https://www.tdwg.org/community/cd/mids/`, 2017. last accessed: 26-11-2020.

[90] Y. L. Simmhan, B. Plale, and D. Gannon, "A survey of data provenance in e-science," *ACM Sigmod Record*, vol. 34, no. 3, pp. 31–36, 2005.

[91] K. Eckert, "Provenance and annotations for linked data," in *International Conference on Dublin Core and Metadata Applications*, pp. 9–18, 2013.

[92] P. Groth, Y. Gil, J. Cheney, and S. Miles, "Requirements for provenance on the web," *International Journal of Digital Curation*, vol. 7, no. 1, pp. 39–56, 2012.

[93] D. Lewis, "General semantics," in *Montague grammar*, pp. 1–50, Elsevier, 1976.

[94] M. Doerr, S. Gradmann, S. Hennicke, A. Isaac, C. Meghini, and H. van de Sompel, "The europeana data model (edm)," in *World Library and Information Congress: 76th IFLA general conference and assembly*, pp. 10–15, 2010.

[95] V. De Boer, J. Wielemaker, J. Van Gent, M. Hildebrand, A. Isaac, J. Van Ossenbruggen, and G. Schreiber, "Supporting linked data production for cultural heritage institutes: The amsterdam museum case study," in *The Semantic Web: Research and Applications. ESWC 2012.* (E. Simperl, P. Cimiano, A. Polleres, O. Corcho,

and V. Presutti, eds.), vol. 7295 of *Lecture Notes in Computer Science*, (Berlin, Heidelberg), pp. 733–747, Springer, 2012.

[96] C. Dijkshoorn, L. Aroyo, G. Schreiber, J. Wielemaker, and L. Jongma, "Using linked data to diversify search results a case study in cultural heritage," in *International Conference on Knowledge Engineering and Knowledge Management* (K. Janowicz, S. Schlobach, P. Lambrix, and E. Hyvönen, eds.), vol. 8876 of *Lecture Notes in Computer Science*, (Cham), pp. 109–120, Springer International Publishing, 2014.

[97] M. Dragoni, E. Cabrio, S. Tonelli, and S. Villata, "Enriching a small artwork collection through semantic linking," in *The Semantic Web. Latest Advances and New Domains. ESWC 2016.* (H. Sack, E. Blomqvist, M. d'Aquin, C. Ghidini, S. P. Ponzetto, and C. Lange, eds.), vol. 9678 of *Lecture Notes in Computer Science*, (Cham), pp. 724–740, Springer International Publishing, 2016.

[98] M. Dragoni, S. Tonelli, and G. Moretti, "A knowledge management architecture for digital cultural heritage," *Journal on Computing and Cultural Heritage (JOCCH)*, vol. 10, pp. 1–18, August 2017.

[99] E. Hyvönen, E. Mäkelä, T. Kauppinen, O. Alm, J. Kurki, T. Ruotsalo, K. Seppälä, J. Takala, K. Puputti, H. Kuittinen, *et al.*, "Culturesampo: A national publication system of cultural heritage on the semantic web 2.0," in *European Semantic Web Conference*, pp. 851–856, Springer, 2009.

[100] M. Fernández-López, A. Gómez-Pérez, and N. Juristo, "Methondology: from ontological art towards ontological engineering," in *Proceedings of the AAAI97 Spring Symposium*, pp. 33–40, March 1997.

[101] M. C. Suárez-Figueroa, A. Gómez-Pérez, and M. Fernández-López, "The neon methodology for ontology engineering," in *Ontology engineering in a networked world*, pp. 9–34, Springer, Berlin, Heidelberg, 2012.

[102] J. Gray and A. S. Szalay, "Where the rubber meets the sky: Bridging the gap between databases and science," *CoRR abs/cs/0502011*, 2005.

[103] S. J. Baskauf, J. Wieczorek, J. Deck, and C. O. Webb, "Lessons learned from adapting the darwin core vocabulary standard for use in rdf," *Semantic Web*, vol. 7, pp. 617–627, October 2016.

[104] S. J. Baskauf and C. O. Webb, "Darwin-sw: Darwin core-based terms for expressing biodiversity data as rdf," *Semantic Web*, vol. 7, pp. 629–643, October 2016.

[105] J. Tuominen, N. Laurenne, and E. Hyvönen, "Biological names and taxonomies on the semantic web–managing the change in scientific conception," in *The Semantic Web: Research and Applications. ESWC 2011.* (G. Antoniou, M. Grobelnik, E. Simperl, B. Parsia, D. Plexousakis, P. De Leenheer, and J. Pan, eds.), vol. 6644 of *Lecture Notes in Computer Science*, (Berlin, Heidelberg), pp. 255–269, Springer, 2011.

[106] C. J. Mungall, C. Torniai, G. V. Gkoutos, S. E. Lewis, and M. A. Haendel, "Uberon, an integrative multi-species anatomy ontology," *Genome biology*, vol. 13, p. R5, January 2012.

[107] G. Fragoso, S. de Coronado, M. Haber, F. Hartel, and L. Wright, "Overview and utilization of the nci thesaurus," *Comparative and Functional Genomics*, vol. 5, no. 8, pp. 648–654, 2004.

[108] M. Wick and B. Vatant, "The geonames geographical database." `http://www.geonames.org/`, 2012. last accessed: 30-03-2019.

[109] M. F. Loesch, "Viaf (the virtual international authority file)–http://viaf.org," *Technical Services Quarterly*, vol. 28, pp. 255–256, March 2011.

[110] B. Haslhofer, E. Momeni Roochi, B. Schandl, and S. Zander, "Europeana rdf store report," tech. rep., University of Vienna, 2011.

[111] E. Minack, W. Siberski, and W. Nejdl, "Benchmarking fulltext search performance of rdf stores," in *The Semantic Web: Research and Applications. ESWC 2009.* (L. Aroyo, P. Traverso, F. Ciravegna, P. Cimiano, T. Heath, E. Hyvönen, R. Mizoguchi, E. Oren, M. Sabou, and E. Simperl, eds.), vol. 5554 of *Lecture Notes in Computer Science*, (Berlin, Heidelberg), pp. 81–95, Springer, 2009.

[112] R. Bhagdev, S. Chapman, F. Ciravegna, V. Lanfranchi, and D. Petrelli, "Hybrid search: Effectively combining keywords and semantic searches," in *The Semantic Web: Research and Applications* (S. Bechhofer, M. Hauswirth, J. Hoffmann, and M. Koubarakis, eds.), vol. 5021 of *Lecture Notes in Computer Science*, (Berlin, Heidelberg), pp. 554–568, Springer, 2008.

[113] E. Kaufmann, "Talking to the semantic web – query interfaces to ontologies for the casual user," in *The Semantic Web - ISWC 2006* (I. Cruz, S. Decker, D. Allemang, C. Preist, D. Schwabe, P. Mika, M. Uschold, and L. M. Aroyo, eds.), vol. 4273 of *Lecture Notes on Computer Science*, (Berlin, Heidelberg), pp. 980–981, Springer, November 2006.

[114] E. Kaufmann and A. Bernstein, "Evaluating the usability of natural language query languages and interfaces to semantic web knowledge bases," *Web Semantics: Science, Services and Agents on the World Wide Web*, vol. 8, pp. 377–393, November 2010.

[115] D. A. Koutsomitropoulos, R. B. Domenech, and G. D. Solomou, "A structured semantic query interface for reasoning-based search and retrieval," in *The Semantic Web: Research and Applications. ESWC 2011.* (G. Antoniou, M. Grobelnik, E. Simperl, B. Parsia, D. Plexousakis, P. De Leenheer, and J. Pan, eds.), vol. 6643 of *Lecture Notes in Computer Science*, (Berlin, Heidelberg), pp. 17–31, Springer, 2011.

[116] L. Stork and A. Kuzniar, "Sfb-annotator (version 0.1.1) zenodo." `https://doi.org/10.5281/zenodo.4602263`, 2021.

[117] A. Meroño-Peñuela and R. Hoekstra, "grlc Makes GitHub Taste Like Linked Data APIs," in *The Semantic Web: ESWC 2016 Satellite Events, Heraklion, Crete, Greece, May 29 – June 2, 2016*, pp. 342–353, Springer, 2016.

[118] M. Ritsema van Eck and L. Schomaker, "Formal semantic modeling for human and machine-based decoding of medieval manuscripts," in *Proceedings of Digital Humanities*, pp. 336–338, University of Hamburg, July 2012.

[119] Z. Shi, "Datefinder: detecting date regions on handwritten document images based on positional expectancy," Master's thesis, University of Groningen, Groningen, the Netherlands, 2016.

[120] D. Koning, I. N. Sarkar, and T. Moritz, "Taxongrab: Extracting taxonomic names from text," *Biodiversity Informatics*, vol. 2, pp. 79–82, 2005.

[121] P. B. Heidorn and Q. Wei, "Automatic metadata extraction from museum specimen labels," in *International Conference on Dublin Core and Metadata Applications*, pp. 57–68, 2008.

[122] M. Bulacu, R. van Koert, L. Schomaker, and T. van der Zant, "Layout analysis of handwritten historical documents for searching the archive of the cabinet of the dutch queen," in *Document Analysis and Recognition, 2007. ICDAR 2007. Ninth International Conference on*, vol. 1, pp. 357–361, IEEE, 2007.

[123] T. Van der Zant, L. Schomaker, and K. Haak, "Handwritten-word spotting using biologically inspired features," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 30, no. 11, pp. 1945–1957, 2008.

[124] J.-P. van Oosten and L. Schomaker, "Separability versus prototypicality in handwritten word-image retrieval," *Pattern Recognition*, vol. 47, no. 3, pp. 1031–1038, 2014.

[125] F. Chollet *et al.*, "Keras." `https://keras.io`, 2015.

[126] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proceedings of the IEEE*, vol. 86, no. 11, pp. 2278–2324, 1998.

[127] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *CoRR*, vol. abs/1409.1556, 2014.

[128] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pp. 248–255, IEEE, 2009.

[129] M. Oquab, L. Bottou, I. Laptev, and J. Sivic, "Learning and transferring mid-level image representations using convolutional neural networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1717–1724, IEEE, 2014.

[130] A. McCallum and W. Li, "Early results for named entity recognition with conditional random fields, feature induction and web-enhanced lexicons," in *Proceedings of the seventh conference on Natural language learning at HLT-NAACL 2003-Volume 4*, pp. 188–191, Association for Computational Linguistics, 2003.

[131] J. A. Drew, C. S. Moreau, and M. L. Stiassny, "Digitization of museum collections holds the potential to enhance researcher diversity," *Nature ecology & evolution*, vol. 1, no. 12, p. 1789, 2017.

[132] M. Wang and W. Deng, "Deep visual domain adaptation: A survey," *Neurocomputing*, vol. 312, pp. 135–153, 2018.

[133] G. Van Horn, S. Branson, R. Farrell, S. Haber, J. Barry, P. Ipeirotis, P. Perona, and S. Belongie, "Building a bird recognition app and large scale dataset with citizen scientists: The fine print in fine-grained dataset collection," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 595–604, IEEE, 2015.

[134] T. Berg, J. Liu, S. Woo Lee, M. L. Alexander, D. W. Jacobs, and P. N. Belhumeur, "Birdsnap: Large-scale fine-grained visual categorization of birds," in *Proceedings of*

the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2011–2018, IEEE, 2014.

[135] C. H. Lampert, H. Nickisch, and S. Harmeling, "Learning to detect unseen object classes by between-class attribute transfer," in *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pp. 951–958, IEEE, 2009.

[136] N. Kumar, P. N. Belhumeur, A. Biswas, D. W. Jacobs, W. J. Kress, I. C. Lopez, and J. V. B. Soares, "Leafsnap: A computer vision system for automatic plant species identification," in *Proceedings of the European Conference on Computer Vision*, pp. 502–516, Springer, 2012.

[137] M. E. Nilsback and A. Zisserman, "A visual vocabulary for flower classification," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, vol. 2, pp. 1447–1454, IEEE, 2006.

[138] S. Beery, E. Cole, and A. Gjoka, "The iwildcam 2020 competition dataset," *arXiv preprint arXiv:2004.10340*, 2020.

[139] O. Mac Aodha, E. Cole, and P. Perona, "Presence-only geographical priors for fine-grained image classification," in *Proceedings of the IEEE International Conference on Computer Vision*, (Seoul, Korea (South)), pp. 9595–9605, 2019.

[140] G. Chu, B. Potetz, W. Wang, A. Howard, Y. Song, F. Brucher, T. Leung, and H. Adam, "Geo-aware networks for fine-grained recognition," in *Proceedings of the IEEE International Conference on Computer Vision Workshops*, (Seoul, Korea (South)), pp. 247–254, 2019.

[141] S. Beery, G. Wu, V. Rathod, R. Votel, and J. Huang, "Context r-cnn: Long term temporal context for per-camera object detection," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, (Seattle, Washington), pp. 13072–13082, 2020.

[142] G. Sumbul, R. G. Cinbis, and S. Aksoy, "Fine-grained object recognition and zero-shot learning in remote sensing imagery," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 56, no. 2, pp. 770–779, 2018.

[143] Z. Akata, S. Reed, D. Walter, H. Lee, and B. Schiele, "Evaluation of output embeddings for fine-grained image classification," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2927–2936, IEEE, 2015.

[144] P. N. Belhumeur, D. Chen, S. Feiner, D. W. Jacobs, W. J. Kress, H. Ling, I. Lopez, R. Ramamoorthi, S. Sheorey, S. White, and L. Zhang, "Searching the world's herbaria: A system for visual identification of plant species," in *Proceedings of the European Conference on Computer Vision*, pp. 116–129, Springer, 2008.

[145] B. Barz and J. Denzler, "Hierarchy-based image embeddings for semantic image retrieval," in *Proceedings of the IEEE Winter Conference on Applications of Computer Vision (WACV)*, pp. 638–647, IEEE, 2019.

[146] I. Tsochantaridis, T. Joachims, T. Hofmann, and Y. Altun, "Large margin methods for structured and interdependent output variables," *Journal of Machine Learning Research*, vol. 6, no. Sep, pp. 1453–1484, 2005.

[147] T. Mikolov, I. Sutskever, K. Chen, G. Corrado, and J. Dean, "Distributed representations of words and phrases and their compositionality," in *Advances in Neural Information Processing Systems*, vol. 2, pp. 3111–3119, 2013.

[148] J. Pennington, R. Socher, and C. Manning, "Glove: Global vectors for word representation," in *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pp. 1532–1543, Association for Computational Linguistics, 2014.

[149] Z. S. Harris, "Distributional structure," *Word*, vol. 10, no. 2-3, pp. 146–162, 1954.

[150] Z. Akata, F. Perronnin, Z. Harchaoui, and C. Schmid, "Label-embedding for image classification," *IEEE transactions on pattern analysis and machine intelligence*, vol. 38, no. 7, pp. 1425–1438, 2015.

[151] A. Frome, G. S. Corrado, J. Shlens, S. Bengio, J. Dean, M. Ranzato, and T. Mikolov, "Devise: A deep visual-semantic embedding model," in *Advances in Neural Information Processing Systems*, pp. 2121–2129, 2013.

[152] B. Romera-Paredes and P. H. S. Torr, "An embarrassingly simple approach to zero-shot learning," in *Visual Attributes*, pp. 11–30, Springer, 2017.

[153] Y. Xian, Z. Akata, G. Sharma, Q. Nguyen, M. Hein, and B. Schiele, "Latent embeddings for zero-shot classification," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 69–77, 2016.

[154] R. Socher, M. Ganjoo, C. D. Manning, and A. Ng, "Zero-shot learning through cross-modal transfer," in *Advances in Neural Information Processing Systems*, pp. 935–943, 2013.

[155] G. Patterson and J. Hays, "Sun attribute database: Discovering, annotating, and recognizing scene attributes," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2751–2758, IEEE, 2012.

[156] C. Wah, S. Branson, P. Welinder, P. Perona, and S. Belongie, "The caltech-ucsd birds-200-2011 dataset," 2011.

[157] Y. Xian, C. H. Lampert, B. Schiele, and Z. Akata, "Zero-shot learning–a comprehensive evaluation of the good, the bad and the ugly," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 41, no. 9, pp. 2251–2265, 2019.

[158] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2818–2826, IEEE, 2016.

[159] N. T. H. Nguyen, A. J. Soto, G. Kontonatsios, R. Batista-Navarro, and S. Ananiadou, "Constructing a biodiversity terminological inventory," *PLoS ONE*, vol. 12, no. 4, p. e0175277, 2017.

[160] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient estimation of word representations in vector space," in *Workshop proceedings of the International Conference on Learning Representations*, arXiv preprint arXiv:1301.3781, 2013.

[161] P. D. Turney and P. Pantel, "From frequency to meaning: Vector space models of semantics," *Journal of Artificial Intelligence Research*, vol. 37, no. 1, pp. 141–188, 2010.

[162] L. v. d. Maaten and G. Hinton, "Visualizing data using t-sne," *Journal of machine learning research*, vol. 9, no. Nov, pp. 2579–2605, 2008.

[163] P. M. Choate, "Introduction to the identification of beetles (coleoptera)," *Dichotomous keys to some Families of Florida Coleoptera*, pp. 23–33, 1999.

[164] T. G. Dietterich, "Approximate statistical tests for comparing supervised classification learning algorithms," *Neural computation*, vol. 10, no. 7, pp. 1895–1923, 1998.

[165] M. Ameryan and L. Schomaker, "Improving the robustness of lstms for word classification using stressed word endings in dual-state word-beam search," in *2020 17th International Conference on Frontiers in Handwriting Recognition (ICFHR)*, pp. 13–18, IEEE, 2020.

[166] S. Chanda, J. Baas, D. Haitink, S. Hamel, D. Stutzmann, and L. Schomaker, "Zero-shot learning based approach for medieval word recognition using deep-learned

features," in *2018 16th International Conference on Frontiers in Handwriting Recognition (ICFHR)*, pp. 345–350, IEEE, 2018.

Acronyms

AI artificial intelligence. 15, 118, 122, 153, 157

ALICE Artificial Intelligence and Cognitive Engineering. iv, 13

ANN artificial neural network. 25, 26, 27, 81

API Application Programming Interface. 35, 63, 64, 74

BHL Biodiversity Heritage Library. 35, 61, 78, 79, 98, 101, 105, 110, 117, 118, 121

BLSTM bilateral long short-term memory network. 28, 43, 45, 82, 83, 84, 86, 87

CE Concatenated Embedding. 101

CH cultural heritage. 50

CNN convolutional neural networks. 27, 28, 81, 82, 84, 85, 86, 87, 94, 99

COL Catalogue of Life. 117

DC Dublin Core. 61, 62

DIAR Document Image Analysis and Recognition. 77, 79

DNN deep (artificial) neural network. 25, 26, 28

DSW Darwin Core Semantic Web. 53, 54, 56, 57, 58, 83

DwC Darwin Core. 31, 41, 42, 53, 54, 56, 57, 58, 61, 62, 63, 64, 79

EOL Encyclopedia of Life. 35

FAIR Findable, Accessible, Interoperable, and Reusable. 8, 10, 31, 48, 49, 120, 121, 122, 148, 150

FOAF Friend Of A Friend. 59

FP Fused Prototype. 92, 98, 102, 116

FSL few-shot learning. 28, 29, 102

GBIF Global Biodiversity Data Facility. 35, 42, 62, 96, 98, 101, 105, 117, 121

GLAMs Galleries, Libraries, Archives and Museums. 37, 40, 42, 45

GZSL generalised zero-shot learning. 94, 95, 112, 114

HMM Hidden Markov Model. 43

HPL Hierarchical Prototype Loss. 92, 98, 102, 103, 112, 116

HTR Handwritten Text Recognition. 6, 13, 38, 39, 41, 43, 44, 45, 46, 77, 78, 79, 81, 88, 121, 122

IIIF International Image Interoperability Framework. 72, 74

IRI Internationalised Resource Identifier. 8, 33, 39, 48, 50, 52, 53, 54, 60, 62, 63, 65, 67, 68, 85

IUCN International Union for Conservation of Nature and Natural Resources. 35

JSON JavaScript Object Notation. 64

KRR knowledge representation and reasoning. 22, 30, 119

LCDS Leiden Centre of Data Science. iv

LD Linked Data. 30, 48

LIACS Leiden Institute of Advanced Computer Science. iv, 13, 153, 157, 158

LINNAE LInking Notes of NAturE. 11, 72, 159

LOD Linked Open Data. 11, 120

LSTM long short-term memory network. 28, 83

MIDS Minimum Information about a Digital Specimen. 47

MLP multi-layer perceptron. 26, 27, 28, 82, 84, 86, 87

MODS Metadata Object Description Schema. 61

NBC Naturalis Biodiversity Center. iv, 1, 2, 3, 13, 18, 36, 54

NC Committee for Natural History of the Netherlands Indies ("Natuurkundige Commissie voor Nederlands-Indië"). 1, 2, 13, 20, 35, 51, 66, 72, 88

NCD Natural Collections Description. 41, 61, 62

NCO Natural Committee Online. 13

NER named entity recognition. 45

NERC named entity recognition and classification. 7, 11, 22, 24, 44, 46, 48, 75, 78, 79, 82, 120, 121

NHC natural history collection. i, 3, 5, 6, 7, 9, 10, 11, 12, 13, 17, 34, 47, 49, 54, 56, 57, 60, 61, 62, 63, 68, 69, 77, 78, 81, 83, 87, 89, 117, 118, 119, 120, 121, 122, 147

NLP natural language processing. 38, 44, 75

NMNH Smithsonian National Museum of Natural History. 61

OBIE ontology-based information extraction. 45

OCR Optical Character Recognition. 6, 35, 39, 43, 79

OOV out-of-vocabulary. 6

OWL Web Ontology Language. 33, 50, 52, 83

PNL Prototypical Network Loss. 99, 103, 112

RDF Resource Description Framework. 8, 33, 34, 45, 53, 59, 64, 74, 85, 87, 88

RNN recurrent neural network. 28

ROI region of interest. 39, 40, 43, 46, 63, 64, 67

SFB-Annotator Semantic Field Book Annotator. 11, 50, 61, 63, 65, 66, 71, 72, 74, 117, 120, 122, 148, 150

SGD Stochastic Gradient Descent. 99

SIA Smithsonian Institution Archives. 61

SNERC salient named entity recognition and classification. 11, 12, 52, 65, 78, 120, 122

SPARQL SPARQL Protocol and RDF Query Language. 8, 34, 68, 69, 71, 72, 74, 87

STePS Department of Science, Technology, and Policy Studies. iv, 13

SVM Support Vector Machine. 107, 108

t-SNE t-Distributed Stochastic Neighbour Embedding. 106, 107, 115

TDWG Biodiversity Information Standards. 35, 47

TEI Text Encoding Initiative. 41

TNS Taxonomic Name Server. 79

URI Uniform Resource Identifier. 8, 21, 33, 34, 46

URL Uniform Resource Locator. 8, 33

VGG Visual Geometry Group. 81, 82

VIAF Virtual International Authority File. 48, 59, 65

W3C World Wide Web Consortium. 33, 52, 60

XML Extensible Markup Language. 31, 32, 41, 42, 61

ZICE Zoological Illustration and Class Embedding. 12, 92, 95, 97, 98, 106, 115

ZSL zero-shot learning. 6, 12, 25, 28, 29, 91, 92, 93, 94, 95, 96, 98, 99, 105, 109, 111, 112, 114, 116, 121, 122, 151

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.

Samenvatting

Beschrijvende kennis over de natuurlijke wereld vormt begrip over de verschillende soorten entiteiten die erin leven, hoe ze invloed uitoefenen op en beïnvloed worden door hun veranderende omgeving, en de processen die hun diversiteit teweegbrengen. Dergelijke kennis is cruciaal als het erom gaat beter geïnformeerde beslissingen te nemen voor beleid dat van invloed is op de natuurlijke diversiteit van de wereld, van organismen tot aan ecosystemen. Om kennis over biodiversiteit te vergaren, worden al sinds eeuwen expedities opgezet naar biodiverse plekken om daar levende organismen te beschrijven, illustreren, en verzamelen. Wereldwijd bestaan er vele verzamelde collectieobjecten zoals specimens, veldnotities, soortillustraties, en andere bronnen, opgeslagen in musea en andere instituten. Helaas blijven deze collectieobjecten vaak onderbelicht, vooral vanwege hun complexe contextafhankelijke karakter, het feit dat ze kennis vaak slechts impliciet overdragen, en dat objecten uit verzamelingen vaak verspreid zijn over verschillende collecties en instituten, waardoor hereniging bemoeilijkt wordt en gegevens om deze reden soms incompleet zijn.

Door de hoeveelheid aan historische en hedendaagse collecties samen te brengen op het wereldwijde web (the World Wide Web), als één wereldwijde natuurhistorische collectie, wordt het mogelijk gedetailleerde spatio-temporele analyses te maken van de natuurlijke wereld en van veranderende praktijken in de natuurhistorie. Het bundelen en destilleren van kennis uit grote collecties digitale en fysieke natuurhistorische objecten en het opslaan van het resultaat als gestructureerde, wereldwijd herbruikbare, en toegankelijke kennis, bevordert de samenwerking binnen de onderzoeksgemeenschap en bevordert daarmee het onderzoek en de ontdekking van nieuwe kennis. Het semantisch web (the Semantic Web) biedt een raamwerk om kennis op een zodanige manier op te slaan: *informatie op het wereldwijde web een goed gedefinieerde betekenis geven, waardoor computers en mensen beter kunnen samenwerken aan deze informatie.*[1]

[1] Een geparafraseerd fragment uit een artikel welke gepubliceerd is in de Scientific American van mei 2001, genaamd "Het Semantische Web": "Het Semantische Web is een uitbreiding van het huidige web waarin informatie een duidelijk gedefinieerde betekenis krijgt, waardoor computers en mensen beter kunnen samenwerken (vertaling vanuit het Engels)"

SAMENVATTING

In dit proefschrift analyseren we verschillende methoden om (i) rijke kennis te extraheren die opgeschreven staat in verschillende bronnen van natuurhistorische collecties en (ii) het resultaat op het semantische web te publiceren als machinaal leesbare kennis, zodat anderen deze kennis kunnen hergebruiken voor eigen onderzoek of integratie met eigen collectiegegevens.

Eerst analyseren we systeemontwerpen voor het extraheren van informatie uit documenten met daarin soortobservaties, zoals veldnotities en wetenschappelijke illustraties. Voor de extractie van informatie en kennis uit veldnotities, pleiten we voor een benadering die kwaliteit verkiest boven kwantiteit, rijke semantische annotatie boven volledige teksttranscriptie. Veldnotities zijn een uitdaging om mee te werken vanwege een verscheidenheid aan factoren, zoals de evoluerende visuele stijl van een enkel alfabet en historisch, moeilijk leesbaar handschrift. Semantische annotatie spoort domeinexperts aan om hoge kwaliteit data te produceren en, door de kennis-intensieve aard van de taak zorgt voor intrinsieke motivatie. Verder wordt hierbij ook de minimale hoeveelheid informatie geformaliseerd die nodig is voor voldoende digitalisering.

Allereerst analyseren we welk type semantische informatie domeinexperts gebruiken om collecties te doorzoeken en welke metadata ze nodig hebben om hun collectiegegevens te integreren. Vervolgens analyseren we hoe dergelijke kennis opgeslagen kan worden aan de hand van de principes van Findable, Accessible, Interoperable, and Reusable (FAIR) data, om verder wetenschappelijk discours en ontdekking van nieuwe kennis aan te moedigen. In het Nederlands vertaalt FAIR naar data die vindbaar, toegankelijk, interoperabel, en herbruikbaar opgeslagen moeten worden. Als resultaat van het hierboven beschreven proces beschrijven we een webapplicatie voor de semantische annotatie van natuurhistorische archiefcollecties: de SFB-Annotator. De webapplicatie, met het onderliggende semantische model, wordt verder ontwikkeld binnen het LInking Notes of NAturE (LINNAE)–project.

Daarnaast stellen we een methode voor om het semantische annotatie proces te automatiseren. Het handmatig extraheren van kennis uit archieven is een tijdrovend en arbeidsintensief proces. We laten zien dat we wetenschappelijke namen kunnen identificeren en classificeren in handgeschreven veldnotities, met behulp van sterke aannames gebaseerd op de kennis van domeinexperts over de structuur en inhoud van observatierecords.

Ten slotte analyseren we de extractie van kennis uit wetenschappelijke illustraties. Geautomatiseerde identificatie van soorten is een lastige taak vanwege het feit dat het grootste deel van de kansmassa van soorten zich in de staart van de distributie bevindt en er miljoenen klassen in huidige soortentaxonomiën bestaan. Hierdoor is het lastig om modellen te

creëren die zowel veelvoorkomende als zeldzame soorten kunnen herkennen. We stellen voor om het probleem aan te pakken met zero-shot learning (ZSL). Hoewel openstaande kwesties blijven bestaan—bijv., een verschuiving in distributie tussen collecties van wetenschappelijke illustraties, voortkomend uit verschillen in papiersoorten, illustratiestijl, en granulariteit van afgebeelde objecten—faciliteert ZSL het leren van een model met behulp van achtergrondkennis, welke naar onze mening cruciaal is voor het leren van modellen die automatisch kennis uit kleine datasets met heterogene gegevens kunnen extraheren.

Curriculum Vitae

Lise Stork was born on the 8th of April, 1990, in Eindhoven, the Netherlands. She graduated from the Strabrecht College in 2008. In 2012, she moved to Utrecht, where she obtained a BA degree in Communication and Information Studies from the University of Utrecht. Considering a domain change, she moved her studies to the Leiden Institute of Advanced Computer Science (LIACS) of Leiden University, where she eventually, in 2016, graduated cum laude with a MSc degree in Media Technology.

In her pursuit of a second Master's degree in the field of artificial intelligence (AI), she came across another alternative: a PhD research position in the field of AI and digital cultural heritage at the LIACS institute, on the project *Making Sense of Illustrated Handwritten Archives*. Hence, in 2016, immediately after having obtained her Master's degree, Lise joined the consortium as a PhD candidate, conducting her research at the LIACS institute. Since December 2020, she works as a Postdoc in the Knowledge Representation and Reasoning group (KRR) of the Vrije Universiteit Amsterdam (VUA), on the project Meaning and Understanding in Human Centered Artificial Intelligence (MUHAI), with a focus on the development of systems that reason in a human-understandable way.

1st of July, 2021

List of Publications

- **Journals**
  - Stork, L., Weber, A., Gassó Miracle, E., Verbeek, F., Plaat, A., Herik, J. van den, and Wolstencroft, K. Semantic annotation of natural history collections. Web Semantics: Science, Services and Agents on the World Wide Web (2018), https://doi.org/10.1016/j.websem.2018.06.002
  - Stork, L., Weber, A., van den Herik, J., Plaat, A., Verbeek, F., & Wolstencroft, K. (2021). Large-scale zero-shot learning in the wild: Classifying zoological illustrations. Ecological Informatics, 62, 101222. https://doi.org/10.1016/j.ecoinf.2021.101222
- **Conferences**
  - Weber, A., Ameryan, M., Wolstencroft, K., Stork, L., Heerlien, M., and Schomaker, L. Towards a digital infrastructure for illustrated handwritten archives. In M. Loannides, editor, Digital Cultural Heritage, volume 10605 of Lecture Notes in Computer Science, pages 155–166. Springer International Publishing, April 2018. https://doi.org/10.1007/978-3-319-75826-8_13
  - Stork, L., Weber, A., Herik, J. van den, Plaat, A., Verbeek, F., and Wolstencroft, K. From handwritten manuscripts to linked data, In: Méndez E., Crestani F., Ribeiro C., David G., Lopes J. (eds) Digital Libraries for Open Knowledge. TPDL 2018. volume 11057 of Lecture Notes in Computer Science, Springer, Cham https://doi.org/10.1007/978-3-030-00066-0_34
  - Stork, L., Weber, A., Van den Herik, J., Plaat, A., Verbeek, F., Wolstencroft, K., Automated semantic annotation of species names in handwritten texts, In: Fuhr, N., Azzopardi, L., Stein, B., Hauff, C., Mayr, P. & Hiemstra, D. (eds.) Advances in Information Retrieval: 41st European Conference on Information Retrieval Research (ECIR), 2019. vol. 11437 of Lecture Notes in Computer Science, Springer, Cham. 667-680 14 p. https://doi.org/10.1007/978-3-030-15712-8_43

- **Other**
  - Stork, L., Weber, A., Gassó Miracle, E., and Wolstencroft, K., A workflow for the semantic annotation of field Books and specimen labels, in Biodiversity Information Science and Standards 2: e25839 (2018) https://doi.org/10.3897/biss.2.25839

Acknowledgements

Writing up your dissertation during a pandemic is a strange process. For me, not being able to close off such an exceptional time, from the beginning of my Master Media Technology till the end of my PhD, near those who have helped immensely during the whole process, is just plain awkard and unfortunate. Luckily it has also shown me that physical proximity is not a prerequisite for friendship and academic guidance.

I would first of all like to thank my supervisor Katy for her limitless enthusiasm, support, patience, and openness I believe to be quite uncommon. I am very much convinced that these characteristics are what make a great supervisor; at the very least they made my time at Leiden Institute of Advanced Computer Science (LIACS) and my PhD in general especially enjoyable.

I thank Fons and Maarten, who showed me how cool research and AI can be. It is in part because of them that I ventured to pursue a career in academia. I sincerely thank especially Fons for considering me for a PhD position at the time, when I had only just entered the field of computer science and AI.

Jaap I would like to thank for his wise words and critical and amazingly specialised eye for detail, improving my manuscripts every time he touched them. How he steers a consortium meeting, in such a pleasant and fruitful way, is not something just anyone can do.

My promotors, Aske and Fons, I thank for all the time they invested in me, and the good atmosphere they brought to the institute. I still remember the feedback I received from Aske on my first paper, through which I have mastered the art of *getting to the point*. What I have always appreciated in him, is the effort he spends connecting with PhD students and other colleagues on a personal level, not just an academic one, making sure they are mentally fit to work to the best of your abilities. Fons I thank for his great care and advice, how he cares for his very many PhD candidates is admirable.

I had the luck to work within a large, interdisciplinary consortium with an inspiring group of people. Andreas, even though he was, officially, no supervisor of mine, has been a role

model nonetheless, for his good advice and insightful histories of European naturalists exploring the Indonesian archipelago. He taught me the ins and outs of research amidst different disciplines and people, and most importantly: how to make them work together as one well-oiled machine. Maarten I sincerely thank for his kindness towards new researchers, and strong will to change things for the better. I thank Lissa, for her bright mind, strong character, and critical viewpoint, Eulàlia for her support, expert taxonomic knowledge, and warm welcome during my first week on the project: she and Andreas gave me the best introduction a PhD could wish for, by taking me on a tour through the beautiful collection facilities, the 62 meter high tower of the Naturalis Biodiversity Center. I am thankful for the team in Groningen, Mahya and Lambert for having introduced me to the challenging world of handwriting recognition, without their expertise I would not have been able to identify the various challenges and bottlenecks of manuscript digitisation.

From our industry partner, Brill publishers, I would like to thank Ernest, for his brilliantly philosophical take on things, Etienne for his enthusiasm, technical support, willingness to help and honesty, Marti for his great leadership and the interesting conversations we had in the train to Leiden, and Michiel, for his enthusiasm and great interdisciplinary take on things.

From the Vrije Universiteit Amsterdam (VUA), I would warmly like to thank Victor, Tobias, Chris and Frank as experts from the field of digital heritage and the Semantic Web, for warmly welcoming me in the Semantic Web group when I asked them for expert feedback and guidance. It taught me a great deal about collaboration in science, in a way that I would like to pursue.

During my time at LIACS, I have met many great researchers. Without all of their support, my research would not have been possible. I am grateful for the coffee club, the fish tank, the aquarium, whatever name we gave it in the end. Together, we spent many afternoons drinking (a lot of) coffee, discussing each other's interesting research topics, solving coding puzzles, and playing games. Those weeks when we played 'don't get got' were days to remember, even though they often resulted in socially awkward situations. I sincerely thank the Imaging and BioInformatics group, and the people in my office especially: it has really been great working among such wonderful people. I am also profoundly grateful to other colleagues at the institute that helped and supported me, some of whom I got to know very well during my time at LIACS. Aside from the colleagues at LIACS, I thank the great young researchers I have met during my many trips to conferences and other universities, for the valuable insights they gave me for my own work.

My sincere thanks to the promotion committee: Tinde van Andel, Melissa Terras, Jetty Kleijn, Michael Lew, Victor de Boer, and Andreas Weber, for investing their valuable time in the reading of my dissertation. I feel honoured that such great researchers accepted to do so.

I would like to thank the eScience center for their support in the form of technical support from one of their research software engineers. Arnold Kuzniar contributed greatly to the continuation of my PhD work, through the LINNAE project, and made working from home a bit more bearable. I hope we will encounter each other in the three-dimensional world at least once.

I am profoundly grateful to my dear family and friends: to my parents, Anja and Paul, for their love and support, and for having given me the qualities needed to finish a PhD, to Thijs for the great twin bond we have, and to his amazing girlfriend Mirian, to my parents in law, Jan en Jitske, for treating me like a daughter, to my best friend Viola, for being my role model and motivator during those fifteen years that we know each other, and for our recurring Tuesday *viola-and-lise*-nights that have saved my weeks, thanks to Dirk for having listened to my PhD complaints and doubts, but also for having become such a good friend, and to him and Viola for agreeing to be my paranymphs. Kemilly, I thank for her helpful PhD advice and for having become such a close friend in such a short time. Thanks furthermore to all my other dear friends for their support, with an additional thanks to those that took me on cycling and rock climbing trips for keeping me sane.

Lastly, I would like to thank Serge. Throughout the many years we know each other, he has been so incredibly supportive. I want to thank him for the many adventures we had and inside jokes we share. But, I thank him most for his kindness, for always believing in me when I myself do not, and for making me laugh so much, especially during times when I was so immersed in work I had trouble coming down from my ivory tower.