Knowledge extraction from archives of natural history collections

Stork, L.

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

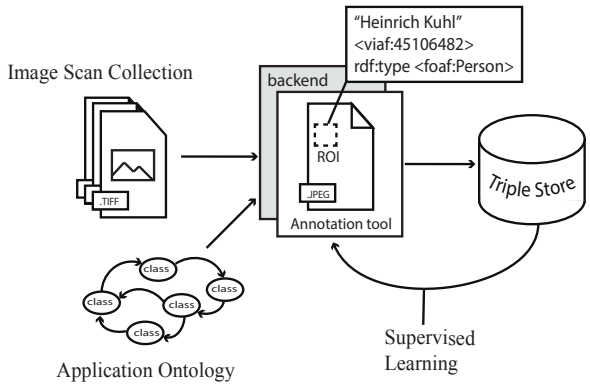

Figure 1.2: The proposed design pattern for the task of SNERC from text images of field notes.

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.