Semantic Annotation of Natural History Collections

Lise Stork[a,*], Andreas Weber[b], Eulàlia Gassó Miracle[c], Fons Verbeek[a], Aske Plaat[a], Jaap van den Herik[a,d], Katherine Wolstencroft[a]

[a] *Leiden Institute of Advanced Computer Science, Niels Bohrweg 1, 2333 CA Leiden, the Netherlands*
[b] *University of Twente, Enschede, the Netherlands*
[c] *Naturalis Biodiversity Center, Leiden, the Netherlands*
[d] *The Leiden Centre of Data Science, Leiden, the Netherlands*

Abstract

Large collections of historical biodiversity expeditions are housed in natural history museums throughout the world. Potentially they can serve as rich sources of data for cultural historical and biodiversity research. However, they exist as only partially catalogued specimen repositories and images of unstructured, non-standardised, hand-written text and drawings. Although many archival collections have been digitised, disclosing their content is challenging. They refer to historical place names and outdated taxonomic classifications and are written in multiple languages. Efforts to transcribe the hand-written text can make the content accessible, but semantically describing and interlinking the content would further facilitate research. We propose a semantic model that serves to structure the named entities in natural history archival collections. In addition, we present an approach for the semantic annotation of these collections whilst documenting their provenance. This approach serves as an initial step for an adaptive learning approach for semi-automated extraction of named entities from natural history archival collections. The applicability of the semantic model and the annotation approach is demonstrated using image scans from a collection of 8,000 field book pages gathered by the Committee for Natural History of the Netherlands Indies between 1820 and 1850, and evaluated together with domain experts from the field of natural and cultural history.

Keywords: Linked Data, Biodiversity, Natural History Collections, Ontologies, Semantic Annotation, History of Science

1. Introduction

Within the field of biodiversity, species research includes the observation and recording of species occurrences in particular geographical areas. Naturalists have been collecting such data for several hundred years and early records are typically housed in natural history museums as hand-written field books, drawings and specimens. However, due to a lack of standardised classification practices during historical biodiversity expeditions, multilingualism and historical terms, the disclosure of such collections proves challenging and time-consuming [24]. Ideas should be developed for the use of semi-automated processes to disclose these collections in order to make them accessible to biodiversity researchers as well as those studying natural and cultural history. In the tower of the Naturalis Biodiversity Center in Leiden, one of the collections, which includes archives recorded in Indonesia between 1820 and 1850, already contains roughly 8,000 field book pages and about 10,000 specimens. Such a collection would shed light upon the development and evolution of biodiversity research concerning insular Southeast Asia in the first half of the nineteenth century. But, as few methods exist to disclose such collections, they remain hidden from the general public as well as researchers.

Through the emergence of digitisation projects [5, 37], new possibilities arise to disclose hand-written manuscript collections with digital tools. Initiatives such as the *Field Book Project* [37], for example, use manual full-text transcription to make their collections available to the general public. In this paper we propose to disclose natural history archival collections through *semantic annotation* of the archive 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 [22]. Work has already been done linking *collections* and *items* using the principles of linked data, not only regarding biodiversity [16, 29], but cultural heritage collections in general [10, 8, 9, 7, 11, 12]. Fewer examples exist where the *content* of items in such collections are semantically linked [8]. Such an approach would serve to facilitate the use of structured queries and reasoning over the data, data aggregation and, through the use of Internationalised Resource Identifiers (IRIs), disambiguation of entities. This paper makes the following contributions:

1. We provide a semantic model, an application ontology

[*] Corresponding author
Email addresses: l.stork@liacs.leidenuniv.nl (Lise Stork), a.weber@utwente.nl (Andreas Weber), eulalia.gassomiracle@naturalis.nl (Eulàlia Gassó Miracle), f.j.verbeek@liacs.leidenuniv.nl (Fons Verbeek), a.plaat@liacs.leidenuniv.nl (Aske Plaat), h.j.vandenherik@law.leidenuniv.nl (Jaap van den Herik), k.j.wolstencroft@liacs.leidenuniv.nl (Katherine Wolstencroft)

written in OWL[1] to structure drawing captions and historical occurrence records in field books. For this we integrate ontologies describing biodiversity, geographic locations and annotation provenance.

2. We present a semantic annotation tool, the *Semantic Field Book Annotator*, which uses the application ontology to enable domain experts to produce structured annotations from digitised natural history archival collections. 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. These results will inform the development of an adaptive learning approach leading to semi-automated annotation.

We show the applicability of the ontology and annotation workflow on a use-case of roughly 8,000 image scans from a collection of field notes and drawings, gathered by the Committee for Natural History of the Netherlands Indies (Natuurkundige Commissie voor Nederlandsch-Indië).

The paper is structured as follows: in section **2** we provide some background information regarding natural history research and outline the requirements for the development of the semantic model. In section **3** we discuss the development method and process: we discuss requirements in section **3.1**, the related work regarding semantics for biodiversity in section **3.2**, elucidation of the content of natural history collections by domain experts in section **3.3** and description of the design choices and the final semantic model for the description of natural history archival collections in section **3.4**. Section **4** describes the annotation approach, a workflow and tool to produce structured annotations from natural history archival collections using the semantic model. In section **5** we evaluate the semantic annotation approach qualitatively and discuss the data acquired from the semantic annotation of a field book from our use-case. Lastly we discuss our results, describe limitations and outline future work in section **6**. This work is part of the Making Sense project.[2]

2. Background

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 research regarding species, genetics and ecology. 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 classify and order observations of organisms and develop methods that moderate systematic descriptions. Expeditions to biodiverse areas allow naturalists to record organism observations and classifications. Field books are the containers that preserve these observation records. 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 [24]. Besides field books, visual material is assembled during expeditions. Historically, collectors were accompanied by professional illustrators, who produced detailed drawings of organisms, as shown in figure 1.

During the development of biodiversity research, methods of species classification were continuously subject to intense discussion [27]. Multiple theories emerged regarding collection practices and species classifications. In particular in the early nineteenth century and before, naturalists were struggling to find and agree upon one ‘true’ natural system [27].

Natural history collections embody this search for a terminological structure which could be used to order, describe and classify nature. The lack of consensus during historical biodiversity expeditions resulted in species descriptions that are challenging to analyse within the present scientific paradigm, and also within collections themselves: (i) biological classification systems implied in field books cannot be directly mapped to present taxonomies (ii) taxa have synonyms within collections and (iii) scientific names shift between genera and species [27, 21, 3], as shown in figure 2. Matching organisms based on metadata recorded in field books can potentially remove ambiguity concerning classifications. Manually structuring and comparing the data would, however, be a time consuming process, as natural history collections often contain thousands of manuscripts and specimens. Moreover, records are written in hard-to-read handwriting and multiple languages interspersed with historical terms. Making sense of the data without the use of automated processes becomes an intractable problem.

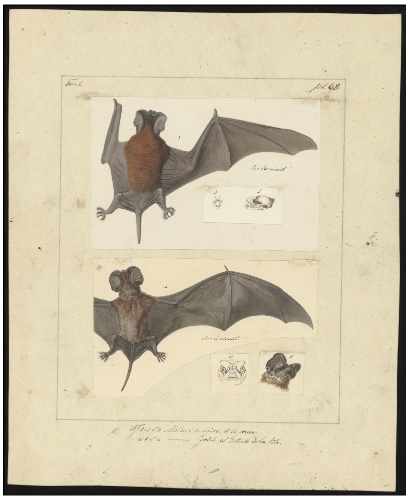


Figure 1: A manuscript taken from the collection of the Committee for Natural History of the Netherlands Indies. Collection Naturalis Biodiversity Center, MMNAT01_AF_NNM001000415. Captions say: *Fig.1-2 et 3. Molosse mégère e le crane. Fig.4-5 et 6. Molosse grêle et details de la tête. Pl.68.* Illustrator unknown. Image free of known restrictions under copyright law (Public Domain Mark 1.0).

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

Figure 2: Synonyms of the current taxon Scotophilus kuhlii temminckii

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

[2] http://makingsenseproject.org/

3. Development of a semantic model

Although data standards, such as the Darwin Core [40], exist for present-day biodiversity research, it became clear through interviews with cultural and natural historians that some tailoring would be required for the semantic annotation of historical biodiversity collections. The development process was set up taking into account the ontology development process described by Fernández et al [13]. The emphasis in the development process of our model is on the re-use and re-engineering of existing semantic models. We thus follow the ontology development process as outlined in scenario 4 of the NeOn methodology for ontology engineering [34]. 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 [15], where questions of domain experts become requirements for the design and evaluation of the system.

3.1. Requirements for a semantic model

The requirements for the semantic model describe user requirements for elucidating content, and requirements for adhering to the principles of sharing data in the semantic web.

1. Elucidating Content

   R1 The model should formalise the general semantics of species observations described in field books and drawings.

   (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 standard language, OWL.

   R2 The model should be able to deal with name variants, such as, historical terms, abbreviations, scientific and vernacular terms, and their context.

   (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 it can be used by domain experts as well as automated reasoners.

2. Serving Structured Annotations to the Semantic Web

   R3 The model should re-use existing ontologies and vocabularies to facilitate data aggregation on the web.

   R4 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 store metadata regarding the annotation process; annotator, date/time, interpretation, to track the provenance of an interpretation.

   (b) The annotations should store metadata regarding their span in the image collection: multiple pages, single pages or fragments from pages, to keep track of the provenance of annotations in relation to the collection. As we will use these fragments in further research for named entity extraction, linking the annotations and their metadata to these fragments facilitates repetition of experiments by other researchers.

3.2. Semantics for biodiversity

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

3.2.1. The Darwin Core

The biodiversity data standard that is most commonly used to model species occurrences is the Darwin Core standard (DwC) [40]. 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 **R1a**, and thus proves to be a suitable baseline for our model.

For the purpose of semantically annotating natural history archival collections, however, the DwC alone does not suffice. Firstly, the DwC does not satisfy **R1b**. Although the terms from the DwC were converted to be used with RDF [2] in 2012, the standard does not allow all properties to be used within its `dwciri:` namespace, adopted to refer to IRIs [2]. 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 RDFSchema. 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.,:`<taxon1> dwc:order "Chiroptera"`. Finally, the DwC 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 [21]. In these respects, the DwC does non satisfy **R2a** and **R2b**.

3.2.2. The Darwin Core Semantic Web

The Darwin Core Semantic Web (Darwin-SW)[3] ontology extends the DwC by providing properties to link the main classes

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

of the DwC [1]. It hereby addresses the limitations of the DwC regarding **R1b**. The Darwin-SW 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. This creates the possibility to match observation records to specimens and drawings, based on their metadata. However, the ontology still does not allow biological taxonomies to be graphically modelled, something that is also included in **R1b**. Finally, to the extent of our knowledge, the applicability of the Darwin-SW ontology has not yet been demonstrated on large datasets.

3.2.3. TaxMeOn

The TaxMeOn[4] Meta-Ontology of Biological Names is an ontology that models biological taxonomies [36]. The ontology uses IRIs for taxa and introduces hierarchy by connecting the taxa to each other using the transitive `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 IRIs, avoiding name ambiguity. Its conceptualisation, however, is subtly different than the Darwin-SW ontology: TaxMeOn models taxa as instances of a rank class such as `genus` whereas the Darwin-SW vocabulary 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 Darwin-SW, but some additions need to be considered for the description of natural history collections. Domain experts' interests were explored to complement the existing vocabularies to satisfy (**R1a**) and to address **R1b**, the darwin-SW ontology was re-structured so that the biological taxonomies can 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 (**R2a**), the inclusion of context metadata (**R2b**) and enables data aggregation on the web (**R3**). Finally, the storage of provenance metadata of annotations (**R4**) was addressed. The process is explained in the coming subsections.

3.3. 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 **R1a**. 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.

3.3.1. Knowledge acquisition

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 thus to think in concepts, they were asked to use these concepts to annotate the field book pages and drawings, allowing the addition of other concepts discovered during the annotation process.

Table 1: Observation record elements organised by topic. Similar concepts were merged, e.g., *Linnean Name* and *Species Name*.

Topic	**Annotated Concepts**	*c*, (**n-7**)
Classification	1. Linnean Name: *30*, (**7-7**) 2. Vernacular Name: *2*, (**2-7**) 3. Literature used: *2*, (**2-7**) 4. Synonyms: *6*, (**4-7**)	5. New namings: *3*, (**2-7**) 6. Additional class.: *6*, (**4-7**)
Species	1. Rarity: *5*, (**2-7**) 2. Use by Locals: *0*	3. Range: *5*, (**2-7**)
Expedition	1. Person: *23*, (**7-7**) (a) Collector: *2*, (**1-7**) (b) Author: *6*, (**2-7**) (c) Companion: *0* (d) Local person: *0* (e) Illustrator: *5*, (**3-7**) 2. Role of Indigenous Population in Knowledge Retrieval: *0*	3. Collection Practice: *2*, (**2-7**) 4. Drawing property: *5*, (**3-7**) 5. Language peculiarity: *0* 6. Date of Observation: *10*, (**7-7**) 7. Place of Observation: *22*, (**7-7**) 8. Publication field book: *0*
Organism	1. Corresponding specimen: *1*, (**1-7**) 2. Corresponding drawing: *2*, (**1-7**) 3. Condition: *0* (a) Living: *0* (b) Dead: *0* 4. Quality: *14*, (**7-7**) (a) Morphology: *5*, (**5-7**) (b) Colour: *2*, (**2-7**) (c) Behaviour: *8*, (**2-7**) 5. Preservation *0*	6. Drawing *17*, (**7-7**) (a) parts *7*, (**2-7**) (b) views *4*, (**3-7**) 7. Anatomy: *40*, (**7-7**) 8. Measurement: *5*, (**5-7**) 9. Count: *1*, (**1-7**) (a) Specimen *0* (b) Anatomical entity: *1*, (**1-7**) 10. Gender: *1*, (**1-7**)

3.3.2. Results

Table 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

[4] http://schema.onki.fi/taxmeon/

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 NHC-Ontology. The questions from domain experts were used to test the output of the annotated field book in section 5.

The most important named entities from table 1 which were extensively annotated by the experts in the field books, but which are not included in the Darwin-SW model, are dates, additional classifications - synonyms and later classifications, additional occurrences - species range and rarity - and structured organism descriptions such as the anatomical parts, qualities and measurements. We thus adopt these in the final model.

3.4. The core model: the NHC-Ontology

In this section we explain further design choices for the Natural History Collection-Ontology (NHC-Ontology) and describe the adoption and application of the classes and properties. The ontology extends the Darwin-SW ontology with two classes and seven properties in order to address the remaining limitations mentioned in section 3.2. Figure 3 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.

3.4.1. Classifications and taxonomies

The class `nhc:TaxonRank` connects to the Darwin-SW 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 3. 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 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 binomial nomenclature, species are named using two names: a genus and a specific epithet or species name. Furthermore, an abbreviated publisher name is included to avoid name ambiguity, e.g., *Pteropus minimus Geoff*, where Geoff refers to *Étienne Geoffroy-Saint-Hilaire*, a french zoologist. Similarly in our model *Genus+species* is seen as a unit representing a species.[5] The name of the publisher is linked separately, as domain experts indicated to have special interest in some authors and would like to be able to retrieve all taxonomical names from a specific scientific author. For instance to obtain knowledge concerning which species they named and their naming conventions. When a species is newly discovered and thus unpublished, authors sometimes use *'Nobis'*, latin for *'by us'*, or some other place holder for the name of the scientific publisher. 'Nobis' in this case still refers to a scientific author name, namely the writers of the field book. Annotating the term as the scientific author of the scientific name is useful as, in combination with the author name of the field book, the taxonomical names can be resolved. 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.

3.4.2. Evidence for identification

In the Darwin-SW model, 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. To achieve this granularity, 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 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 3. As the `dsw:derivedFrom` property is transitive, the measurement is also derived from the specific organism, beneficial for querying and reasoning. We use this measurement class to span measurement tables. Organism fact descriptions however 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 or property of the organism, such as *liver* or *colour*. This way, we can point to a free text description of an organism characteristic, by annotating the anatomical entity or property initiating the description. 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 one query e.g. *'Show me all observation records from person X that measure a liver'*.

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

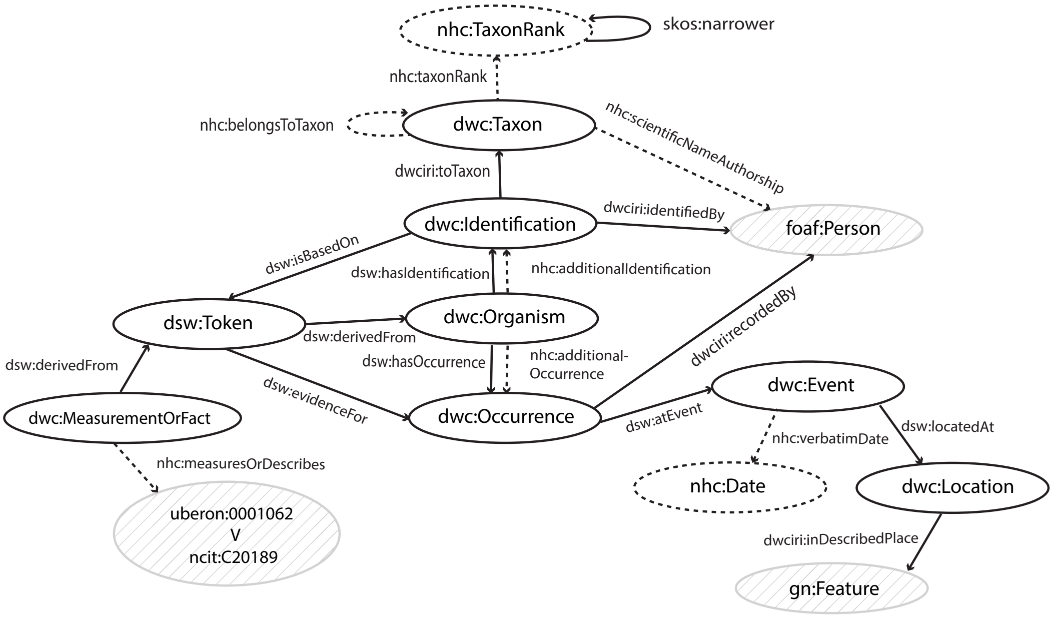


Figure 3: The NHC-Ontology, an extension of the Darwin-SW graph model for annotating natural history collections.

3.4.3. 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.

3.4.4. Written annotations

In field books, we often see 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. It has to be transparent for them and other researchers 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.

3.4.5. Linking to external ontologies and datasets

The ontology connects to classes from other ontologies and thesauri such as Uberon[6] for anatomical entities [28] and NCIT[7] for species attributes [14], both used for the identification of a taxon, the Geonames Database[8] for geographical locations [39] and VIAF[9] for referring to persons [25] as instances of the class `foaf:Person`. These classes are indicated by a striped fill in figure 3. Linking to these vocabularies provides us with three benefits. First, the entities can be resolved. Second, queries can utilise the structures of these ontologies, when available, for querying and reasoning purposes. Third, these 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. 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 the ontology to biological taxon IRIs from different namespaces. As mentioned in section 3.2.1, The same species name can sometimes refer to different organisms.

[6] http://uberon.github.io
[7] https://ncit.nci.nih.gov/
[8] http://sws.geonames.org/
[9] http://viaf.org/viaf/

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 the provenance of multiple mappings to present taxonomies, should this be required to represent different theories.

3.4.6. 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 digital objects on the web, the Web Annotation Data Model,[10] initially the Open Annotation Model (OA) [18], was used.[11] Reasons for its adoption in our 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 well established in the linked data community. Using this data model and its ontology, we link instances of the classes from the ontology depicted in figure 3 to the image scans. Figure 4 shows an example annotation. The instance node of te class `oa:Annotation` refers to the annotation object itself to which metadata relating to the annotation process is added. The instances of the classes `oa:TextualTag` and `oa:SemanticTag` are the bodies of the annotation. They indicate the semantic interpretation of the annotation, and the verbatim transcription. A semantic body is always an instance of the class `oa:SemanticTag`, but it is also an instance of a class from the NHC-Ontology, in this case `dwc:Taxon`. Each annotation *always* has a textual body, containing its verbatim transcription. This way, the text is transcribed and semantically annotated simultaneously. At the same time, this allows for different name variants of entities that exist within the field books. When an annotation is linked to the IRI of a naturalist such as `<http://viaf.org/viaf/69703180/>` which refers to the dutch naturalist *Coenraad Jacob Temminck*, the textual body will contain the verbatim label that is used in the field book such as the abbreviation *Tem*. Both the full name and the abbreviation from the field book will point to the part of the field book page where Temminck is referenced. The instance of the class `dcmitype:StillImage` from figure 4 refers to the annotated field book page and the instance of the class `oa:Target` to the selected fragment within the page.

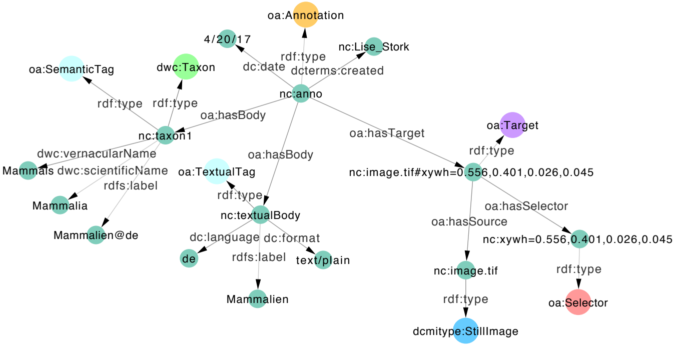


Figure 4: Example of an annotation of the taxon *Mammals* written in a field book, using the Web Annotation Data Model. This annotation contains both a textual and a semantic body. The namespace `nc:` refers to the collection from the Natural Committee for Natural History of the Netherlands Indies.

The resulting application ontology, a combination of the NHC-Ontology and the Web Annotation Data Model, provides a framework for annotating important named entities in the data. It is made accessible to users through a semantic annotation tool, the *Semantic Field Book Annotator* (SFB-Annotator), that enables the semantic annotation of digitised images of handwritten text and illustrations. The tool is discussed in the next section.

4. Semantic annotation of natural history collections

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* [37], for instance, 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 natural history collections and make them available for the general public.

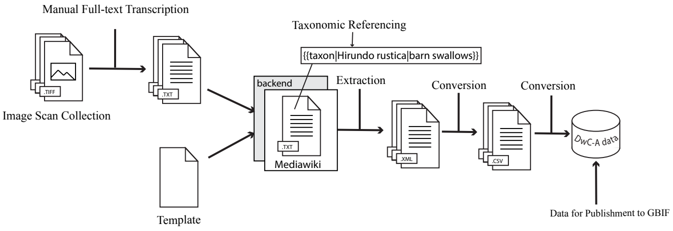


Figure 5: *From Documents to Datasets*[35] workflow

The Field Book Project makes use of the Natural Collections Description (NCD)[12] standard for storing metadata on a *collection* level. Further, the project uses the Metadata Object Description Schema (MODS)[13] to create *item* level metadata[29]. The Biodiversity Heritage Library (BHL)[14] describe their data using XML and MODS or Dublin Core (DC).[15] None of the above mentioned projects, however, aims to annotate the *content* from items within natural history collections. Responding to this need, the project *From Documents to Datasets* [35] provides

[10] https://www.w3.org/TR/annotation-model/

[11] https://www.w3.org/annotation/

[12] https://terms.tdwg.org/wiki/Natural_Collections_Description

[13] http://www.loc.gov/standards/mods/

[14] http://www.biodiversitylibrary.org/

[15] http://dublincore.org/

a workflow for the conversion from digitised handwritten field books to flat data files, see figure 5, structured according to the terms from the Darwin Core standard. They propose first to fully transcribe the texts together with experts, then upload those texts together with the image scans to a MediaWiki[16] server. Via templates, the *taxa*, *locations* and *dates*, are annotated by researchers through a crowd-sourcing initiative. *Taxonomic referencing*, the process of resolving a historical taxon to a current one, occurs *within* the semantic annotation process through interpretation by the annotators. The annotations are then extracted and converted manually to Darwin Core terms, in order to publish them in the Global Biodiversity Information Facility (GBIF)[17] data server [31]. 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.

4.1. Workflow

Similar to the projects mentioned at the beginning of section 4, we use the Natural Collection Description standard and the Dublin Core to enrich natural history collections on a collection and item level. On an item level, the methodological workflow approach in this project differs from the approach in figure 5 as it does not merely structure the entities semantically, it also links all the entities to form a connected graph. The data become readable and interpretable by machines and can be interlinked and aggregated with other biodiversity data on the web. To link the named entities together we use the NHC-ontology, which also enables rich querying and reasoning. Our workflow is shown in figure 6. In our approach, we omit full-text transcription. 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. We also 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 annotation process itself. We furthermore 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.

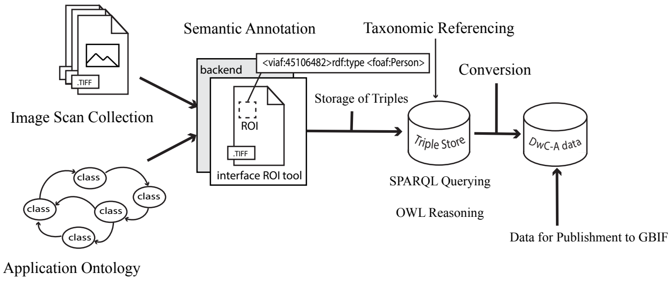
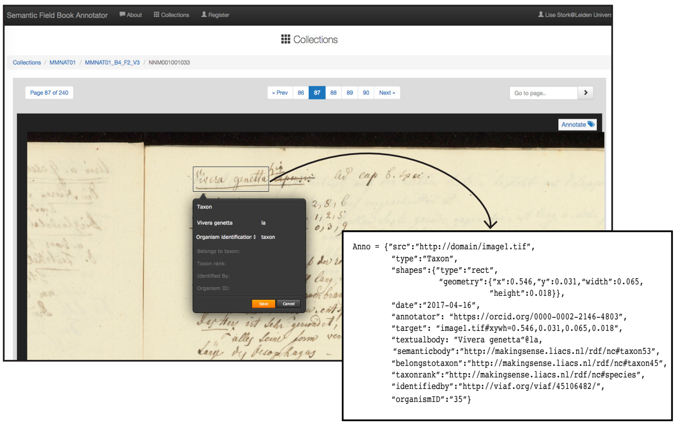


Figure 6: The proposed workflow for semantically annotating natural history collections.

If necessary, 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.

4.2. The Semantic Field Book Annotator

The Semantic Field Book Annotator is a web application, developed for domain experts, to harvest structured annotations from field books using the NHC-Ontology and proposed workflow. With some practice, the tool can also be used to crowdsource annotations, as long as these are validated by an expert curator.


Semantic Field Book Annotator
About
Collections
Register
Lise Stork@Leiden Univers
Collections
Collections / MMNAT01 / MMNAT01_B4_F2_V3 / NNM001001033
Page 87 of 240
« Prev
86
87
88
89
90
Next »
Go to page..
Annotate
Save
Cancel

Taxon

Vivera genetta la

Organism identification taxon

Belongs to taxon:

Taxon rank:

Identified By:

Organism ID:

```
Anno = {"src":"http://domain/image1.tif",
        "type":"Taxon",
        "shapes":{"type":"rect",
                  "geometry":{"x":0.546,"y":0.031,"width":0.065,
                              "height":0.018}},
        "date":"2017-04-16",
        "annotator": "https://orcid.org/0000-0002-2146-4803",
        "target": "image1.tif#xywh=0.546,0.031,0.065,0.018",
        "textualbody: "Vivera genetta"@la,
        "semanticbody":"http://makingsense.liacs.nl/rdf/nc#taxon53",
        "belongstotaxon":"http://makingsense.liacs.nl/rdf/nc#taxon45",
        "taxonrank":"http://makingsense.liacs.nl/rdf/nc#species",
        "identifiedby":"http://viaf.org/viaf/45106482/",
        "organismID":"35"}
```

Figure 7: The annotation process using the *Semantic Field Book Annotator*

As shown in figure 6 and 7, users can draw bounding boxes, or Regions Of Interest (ROIs), over the image scans to which annotations can be attached. The ROI tool makes use of the *Annotorious* annotation API[18] to select a ROI and create an annotation object, see figure 7. The annotation object is connected with its metadata and: a target - a page or a ROI -, a textual body and a semantic body. The *shapes* variable is used to store the geometry of the ROI relative to the image borders. In RDF, these coordinates are stored with the `oa:Selector` class to specify part of the source image, see figure 4. In order to make the manuscript images zoomable, Annotorious is used together with the OpenSeaDragon API.[19]

For storage, we use a servlet that pushes the annotation to an annotation server. In the servlet, annotation objects written in JSON are converted to RDF triples using the RDF4J API, an open source Java framework for processing RDF data. For

[16] https://wikisource.org/
[17] https://www.gbif.org/
[18] https://annotorious.github.io/
[19] https://openseadragon.github.io/

storage of annotations we use the Virtuoso quad store as it is a well evaluated store for data-intensive server applications[17]. Moreover, it can be accessed via the RDF4J API.

In the annotation process, a distinction is made between explicit and implicit classes, where explicit classes, in comparison to implicit classes, refer to the group of named entities that are easily observed in the field books. These are: the *taxonomical 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'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 section 4.1. Another reason for this adoption is that salient named entities can be pulled out of the text more easily by annotators, and finally by automated processes.

During the annotation process, a user first links a ROI to a class c from the set of *explicit* classes $C = \{c_1, c_2, ..., c_n\}$ of the application ontology. In figure 7 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 7, or for instance *Additional occurrence recorded at:* `location`. When a class and predicate are specified, optional metadata fields appear such as the `uberon:` IRI in case of an anatomical entity.

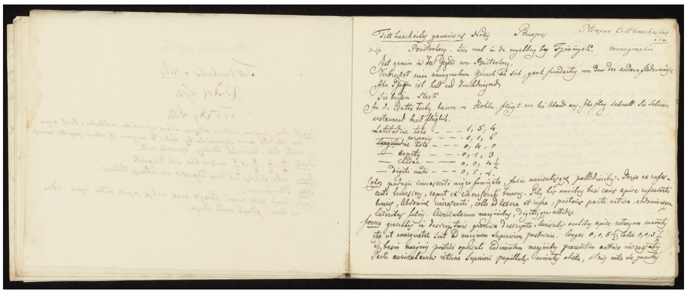
To create connections between all entities from the model that belong to one occurrence record, every time an instance with a `dwc:Taxon` type is annotated, the entire base model, excluding the measurements, is instantiated together with their semantic connections as visible in figure 3. As instances of these classes, unique identifiers are created such as `nc:identification1` or `nc:date1`. 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.[20]

4.3. Towards semi-automated annotation

As a first step towards semi-automated annotation, we prepopulated the 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 a candidate instance $d \in X$, where X is the *instance space*. If no instance yet exists or if it is an implicit instance such as one from the organism class, a random IRI is created.

5. Qualitative Evaluation

In concordance with a domain expert from the field of natural history, one of the field books from the collection of the Natural Committee, named *'Manuscripten van de leden der Natuurkundige commissie: Mammalien, van Kuhl'*, was semantically annotated using the Semantic Field Book (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* and 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, the resulting data and possibilities for querying using the concepts and questions composed by the domain experts mentioned in section 3.3.



Titthaecheilos javanicus Nobis Pteropus Pteropus titthaecheilus Tem

n.sp. Buitenzorg. Een veel in de vogelbey by Tjiminjak. Monographie

Ist gemein in der Gegend von Buitenzorg.

Verbreitet einen unangenehmen Geruch an sich, ganz fremdartig von dem der andern Fledermäuse

Ihr Pfeifen ist hell und durchdringend

Sie beissen stark.

In d. Batta toely baum in Höhlen. fliegt aus bei Abend aus. Ihr flug schnell. Sie scheinen erstaunend breit flüglich.

Latitudine tota — — — 1, 5, 4
corporis — — . 0, 1, 6
~~Longitudine~~ tota — — — 0, 2, 0
— capitis — — 0, 1, 3
— caudae — — 0, 0, 2 1/2
— digiti medii — — 0, 5, 4.

Color podagii cinerascenti nigro fumigato, facie auriculisque pallidioribus. Dorso ex rufescenti brunescens, caput ex cinerascenti bruneus. Pili his omnibus basi canis apice rufescenti bruneis, Abdomine cinerascenti, collo ad latera et infra, pectoris parte antica, abdominisque lateribus fulvis. Auricularum marginibus, digitisque albidis.

Forma generalis in descriptione generica descripta. Auriculis ovalibus apice retrorsum curvatis ita ut emarginatae sint ad marginem superiorem posticum. Longae 0, 0, 8 1/2 latae 0, 0, 5 1/2 ad basin marginis postici operculi rudimentum marginibus praesertim antico incrassatis. Parte auricularum interna superiori papillulis minutis obsita, striis inter se junctis

Figure 8: A page from the annotated field book describing the species Titthaecheilos javanicus *Nobis*. *Pteropus titthaecheilus* Tem (upper right corner) is believed to be 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, resulting in the triple: `nc:organism1` `nhc:additionalIdentification` `nc:taxon2`. Collection Naturalis Biodiversity Center, MMNAT01_AF_NNM001001033_013. Image free of known restrictions under copyright law (Public Domain Mark 1.0)

.

5.1. The annotation process

Annotating a page from the field book using the *Semantic Field Book Annotator* took approximately 1 to 10 minutes, depending upon the amount of named entities on the page and the difficulty of interpreting a named entity. Taxonomical names such as the one in figure 8, *Titthaecheilos javanicus* can be difficult to read and sometimes the order of pages is shuffled, hampering the correct interpretation of links between entities. Other times however, a page only contains one or two easy to read named entities of which the relation is clearly defined. Also, the layout of the document hints to the location of the named entities. Taxonomical names, scientific publishers of names and locations are likely to appear on the top of the 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 5 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, this results in faster processing of the field books into a knowledge base.

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

5.2. The data

From the annotated field book, 98 single pages[21] were semantically annotated and their annotations validated by a natural history expert. Table 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.

Table 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

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

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 the data quality. When a named entity is not annotated, for instance the location of the organism spotting, the IRI exists, as mentioned at the end of section 4.2, but remains without a label and link to an annotation object and a ROI. 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.

5.3. Semantic Queries

The evaluation in section 3.3 resulted in a list containing 53 research questions. 18 questions were from biologists, 28 from cultural historians and 7 from information specialists. Here 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 the NHC-Ontology and demonstrate the gain in comparison to full-text search.

5.3.1. Domain experts' questions

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 IRIs 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 3.4.2. The class `dwc:PreservedSpecimen` is then used instead of `dwc:HumanObservation`.

For clarification a distinction is made between six types of queries, see table 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 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 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

5.3.2. 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

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

Table 4: Example queries for cultural history and biology research

<table>
<tr><th>Cultural History</th><th>Biology</th></tr>
<tr><td>

[Q1] How were species collected by Heinrich Kuhl, viaf:45106482?

```
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 DISTINCT ?label ?page WHERE {
        ?organism dsw:hasOccurrence ?occurrence .
        ?occurrence dwciri:recordedBy viaf:45106482 .
        ?occurrence dsw:hasEvidence ?observationRecord .
        ?organism dsw:hasIdentification ?identification .
        ?identification dwciri:toTaxon ?taxon .
        ?taxon rdfs:label ?label .
        ?anno oa:hasBody ?observationRecord .
        ?anno oa:hasTarget ?page
        }
```

</td><td>

[Q3] Which chiroptera species were collected by Heinrich Kuhl, viaf:45106482, on Java?

```
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>
PREFIX nhc: <http://makingsense.liacs.nl/rdf/nhc/>
PREFIX dwciri: <http://rs.tdwg.org/dwc/iri/>
PREFIX dsw: <http://purl.org/dsw/>
PREFIX viaf: <http://viaf.org/viaf/>
PREFIX gn: <http://www.geonames.org/ontology#>
SELECT DISTINCT ?taxonlabel ?loclabel ?parentlabel WHERE {
        ?taxon rdfs:label ?taxonlabel .
        ?taxon nhc:belongsToTaxon ?order .
        ?order rdfs:label ?Chiropterae
        FILTER regex(?Chiropterae, "Chiropterae") .
        ?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 rdfs:label ?loclabel .
        ?location dwciri:inDescribedPlace ?place .
        ?place gn:parentFeature ?parent .
        ?parent gn:alternateName ?parentlabel
        FILTER regex(str(?parentlabel), "Java", "i")
        FILTER (langMatches(lang(?parentlabel), "en"))
        }
```

</td></tr>
<tr><td>

[Q2] How were habitats described in the collection between *1820* and *1821*?

```
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 DISTINCT ?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
        }
```

</td><td>

[Q4] Which anatomical entities were used for the classification of the genus Pteropus?

```
PREFIX dwciri: <http://rs.tdwg.org/dwc/iri/>
PREFIX dsw: <http://purl.org/dsw/>
PREFIX uberon: <http://purl.obolibrary.org/obo/>
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
         FILTER(?uberon NOT IN (uberon:UBERON_0001062))
         }
```

</td></tr>
</table>

better at querying the *content* [26]. Here, we demonstrate that in the case of field books, structured or hybrid queries[4] using the NHC-Ontology are able to provide more relevant query results than full-text queries.

It is notable from table 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.

Table 4 presents 4 of the 46 questions in SPARQL form. Q1 and Q2 are examples of SPARQL queries that provide an indirect answer to the question, whereas Q3 and Q4 provide a direct answer. More example queries can be found online.[22]

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

We finally argue that, as Virtuoso is equipped with full-text indices that can be queried via SPARQL [17], 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 [19, 20, 23]. In our specific case, a query engine will be developed by partners at *Brill publishers*, collaborators within the Making Sense project.[2]

Although beneficial, the formulation of rich semantic queries is not the main reason for the use of a semantic model for the annotation of natural history collections. Most interesting is the semantic linking of named entities within and between resources, as well as within and across collections. 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.[22] The semantic annotations can be accessed through a SPARQL endpoint[23] which can be queried using a SPARQL query editor.[24] The code for the SFB-Annotator and annotation guidelines can also be found online,[20] and will be updated once newer versions are available.

6. Discussion and Future Work

In this paper, 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 collection of the Committee for Natural History of the Netherlands Indies, 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.

In our next steps, the usability of the SFB-Annotator will be further improved; we will thus 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. After that, we will develop methods for semi-automated semantic annotation of field book records. With fully transcribed texts, language processing is used for semi-automated semantic annotation. As we use pixel data instead of text, we require alternative, image processing methods for salient named entity extraction. Using the output of the annotation process, the system can learn which information is important and where this important information resides in the images [30, 33, 6].

Our final goal within the Making Sense project[2] [38] is to assist a handwriting recognition system MONK [32], with the enrichment of natural history collections. MONK is an adaptive learning system achieving good results on the recognition of text from handwritten collections. Exploiting domain knowledge and the structure of text in natural history collections can potentially aid the recognition process, especially when words have few instances in the archives.

Using automated processes will facilitate efficient enrichment of natural history collections and provide a framework to make sense of complex data that would aid researchers within the field of natural and cultural history research.

Acknowledgement

This work is supported by the Netherlands Organisation for Scientific Research (NWO), grant 652.001.001, and Brill publishers.

Literature

[1] S. J. Baskauf and C. O. Webb. Darwin-sw: Darwin core-based terms for expressing biodiversity data as rdf. *Semantic Web*, 7(6):629–643, October 2016.

[2] 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*, 7(6):617–627, October 2016.

[3] W. G. Berendsohn. The concept of "potential taxa" in databases. *Taxon*, 44(2):207–212, May 1995.

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

[5] 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*, 209:133–146, July 2012.

[6] M. Carbonell, M. Villegas, A. Fornés, and J. Lladós. Joint recognition of handwritten text and named entities with a neural end-to-end model. *arXiv preprint arXiv:1803.06252*, 2018.

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

[23] http://makingsense.liacs.nl/rdf4j-server/repositories/NC

[24] An example query editor is the Yasgui editor: http://yasgui.org/, accessed: 30-03-2018

[8] 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 E. Simperl, P. Cimiano, A. Polleres, O. Corcho, and V. Presutti, editors, *The Semantic Web: Research and Applications. ESWC 2012.*, volume 7295 of *Lecture Notes in Computer Science*, pages 733–747, Berlin, Heidelberg, 2012. Springer.

[9] 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 K. Janowicz, S. Schlobach, P. Lambrix, and E. Hyvönen, editors, *International Conference on Knowledge Engineering and Knowledge Management*, volume 8876 of *Lecture Notes in Computer Science*, pages 109–120, Cham, 2014. Springer International Publishing.

[10] 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*, pages 10–15, 2010.

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

[12] M. Dragoni, S. Tonelli, and G. Moretti. A knowledge management architecture for digital cultural heritage. *Journal on Computing and Cultural Heritage (JOCCH)*, 10(3):1–18, August 2017.

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

[14] G. Fragoso, S. de Coronado, M. Haber, F. Hartel, and L. Wright. Overview and utilization of the nci thesaurus. *Comparative and Functional Genomics*, 5(8):648–654, 2004.

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

[16] N. E. Gwinn and C. Rinaldo. The biodiversity heritage library: sharing biodiversity literature with the world. *IFLA journal*, 35(1):25–34, March 2009.

[17] B. Haslhofer, E. Momeni Roochi, B. Schandl, and S. Zander. Europeana rdf store report. Technical report, University of Vienna, 2011.

[18] B. Haslhofer, R. Simon, R. Sanderson, and H. Van de Sompel. The open annotation collaboration (oac) model. In *Workshop on Multimedia on the Web (MMWeb 2011)*, pages 5–9, Washington, DC, USA, September 2011. IEEE Computer Society.

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

[20] 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*, 8(4):377–393, November 2010.

[21] 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 B. Ludäscher and L. Raschid, editors, *International Workshop on Data Integration in the Life Sciences*, volume 3615 of *Lecture Notes in Computer Science*, pages 80–95, Berlin, Heidelberg, 2005. Springer.

[22] 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*, 2(1):49–79, December 2004.

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

[24] A. M. Lister and C. C. R. Group. Natural history collections as sources of long-term datasets. *Trends in Ecology & Evolution*, 26(4):153–154, January 2011.

[25] M. F. Loesch. Viaf (the virtual international authority file)–http://viaf.org. *Technical Services Quarterly*, 28(2):255–256, March 2011.

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

[27] E. G. Miracle. On whose authority? temminck's debates on zoological classification and nomenclature: 1820–1850. *Journal of the History of Biology*, 44(3):445–481, January 2011.

[28] C. J. Mungall, C. Torniai, G. V. Gkoutos, S. E. Lewis, and M. A. Haendel. Uberon, an integrative multi-species anatomy ontology. *Genome biology*, 13(1):R5, January 2012.

[29] S. Nakasone and C. Sheffield. Descriptive metadata for field books: Methods and practices of the field book project. *D-Lib Magazine*, 19(11/12):1, December 2013.

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

[31] 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*, 9(8):e102623, August 2014.

[32] L. Schomaker. Design considerations for a large-scale image-based text search engine in historical manuscript collections. *It - Information Technology*, 58(2):80–88, April 2016.

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

[34] M. C. Suárez-Figueroa, A. Gómez-Pérez, and M. Fernández-López. The neon methodology for ontology engineering. In *Ontology engineering in a networked world*, pages 9–34. Springer, Berlin, Heidelberg, 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*, 209:235–253, July 2012.

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

[37] A. van Camp and M. R. Kalfatovic. The field book project, 2010. last accessed: 30-03-2019.

[38] A. Weber, M. Ameryan, K. Wolstencorft, L. Stork, M. Heerlien, and L. Schomaker. 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.

[39] M. Wick and B. Vatant. The geonames geographical database, 2012. last accessed: 30-03-2019.

[40] 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*, 7(1):e29715, January 2012.