Knowledge extraction from archives of natural history collections

Stork, L.

CHAPTER 4

Semantic Annotation

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

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

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

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

Specifically, this chapter aims to answer two research questions **Q.2** (*What types of research questions do domain experts formulate regarding the archival content of NHCs, and how can we make the content machine-readable to facilitate such queries?*) and **Q.3** (*How can we accommodate a transparent and FAIR approach to enriching the archival content of NHCs, facilitating and encouraging scientific discourse over the content?*).

4.1 Introduction

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

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

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

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

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

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

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

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

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

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

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

4.2 Development of a Semantic Model

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

4.2.1 Requirements

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

Elucidating Content

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

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

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

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

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

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

Serving Structured Annotations to the Semantic Web

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

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

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

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

4.2.2 Semantics for Biodiversity

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

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

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

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

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

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

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

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

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

4.2.3 Data Elucidation by Domain Experts

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

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

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

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

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

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

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

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

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

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

4.2.4 The NHC-Ontology

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

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

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

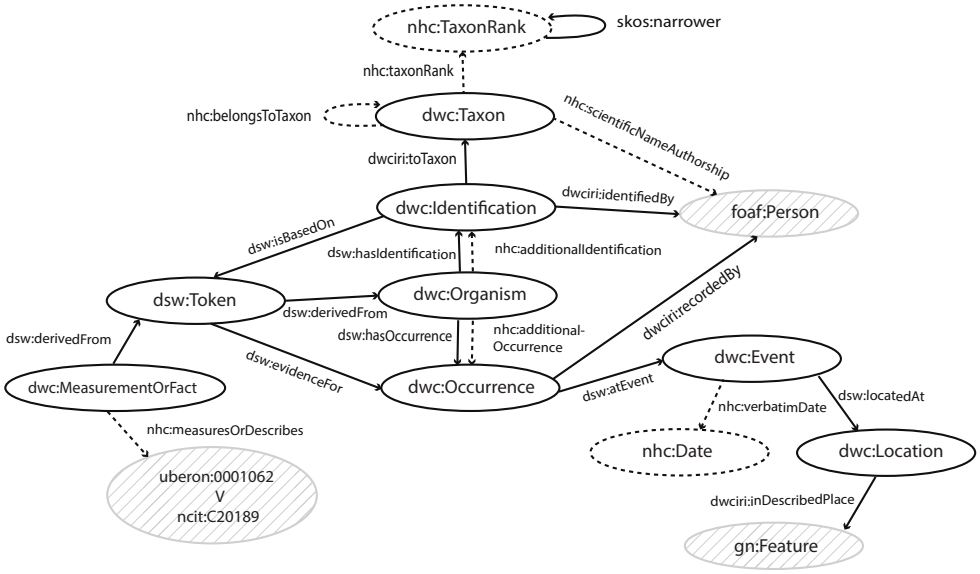

Figure 4.1: The NHC-Ontology, an extension of the DSW ontology for annotating NHCs. Gray striped classes indicate classes from external ontologies, whereas classes and properties with a dotted line pattern indicate additions to the DSW ontology.

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

Listing 4.1: An example annotation

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

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

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

4.3 Semantic Annotation

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

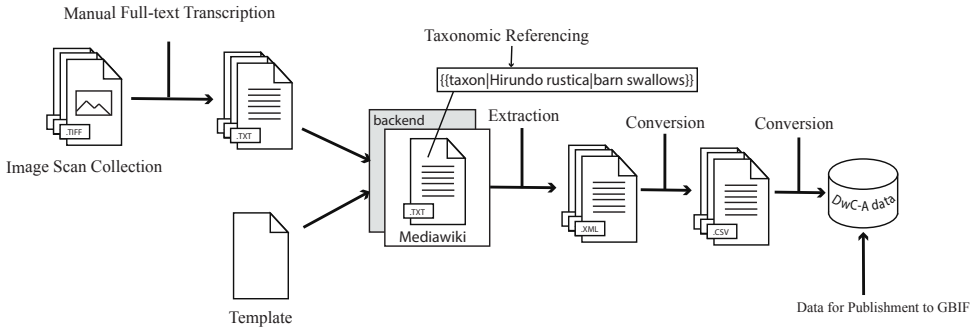

Figure 4.2: *From Documents to Datasets* (35) system design

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

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

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

4.3.1 System Design

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

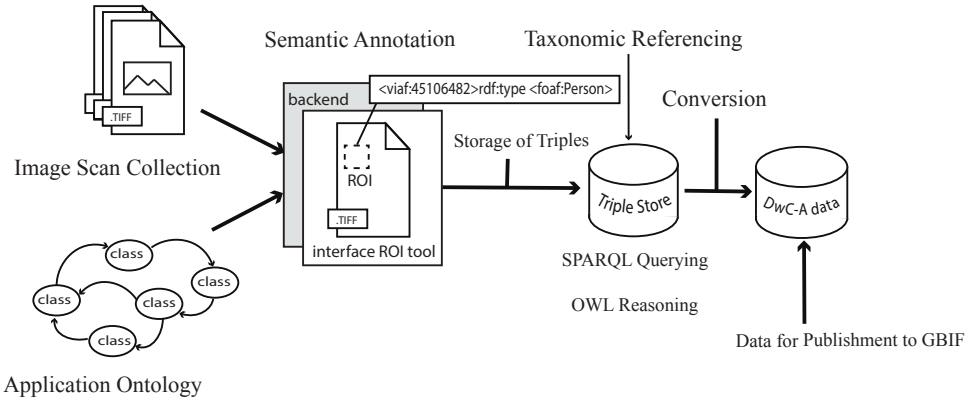

Figure 4.3: The proposed system for semantically annotating manuscripts from NHCs.

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

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

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

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

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

4.3.2 The Semantic Field Book Annotator

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

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

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

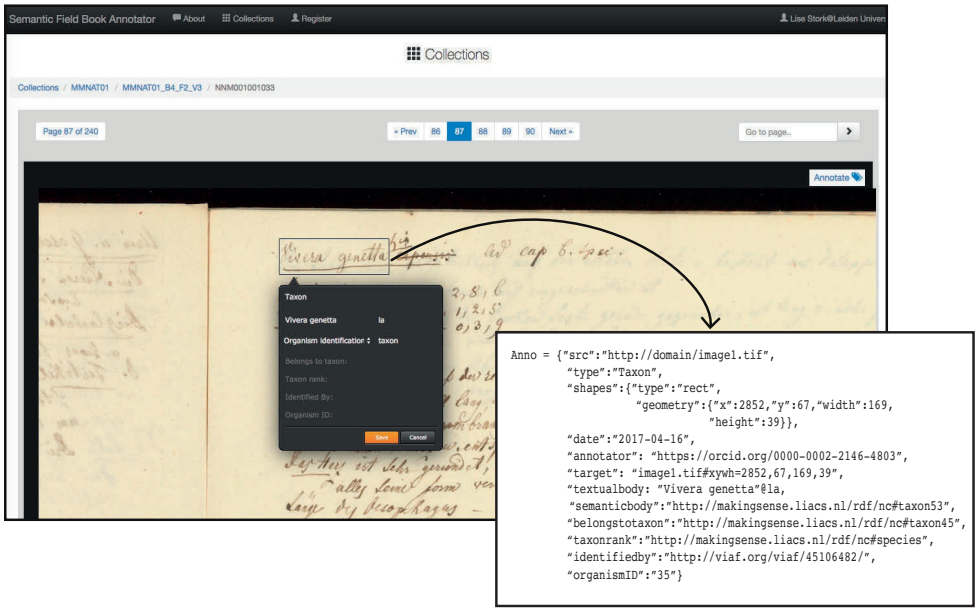

Figure 4.4: The annotation process using the *Semantic Field Book Annotator*

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

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

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

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

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

4.3.3 Towards Semi-Automated Annotation

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

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

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

4.4 Qualitative Evaluation

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

4.4.1 The Annotation Process

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

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

4.4.2 The Data

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

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

Table 4.2: Annotation specifications

Total Annotations

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

Annotations per class

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

Predicate specifics

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

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

4.4.3 Semantic Queries

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

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

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

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

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

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

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

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

Table 4.3: Types of expert queries

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

4.5 Conclusions

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

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

4.6 Ongoing and Future Work

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

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

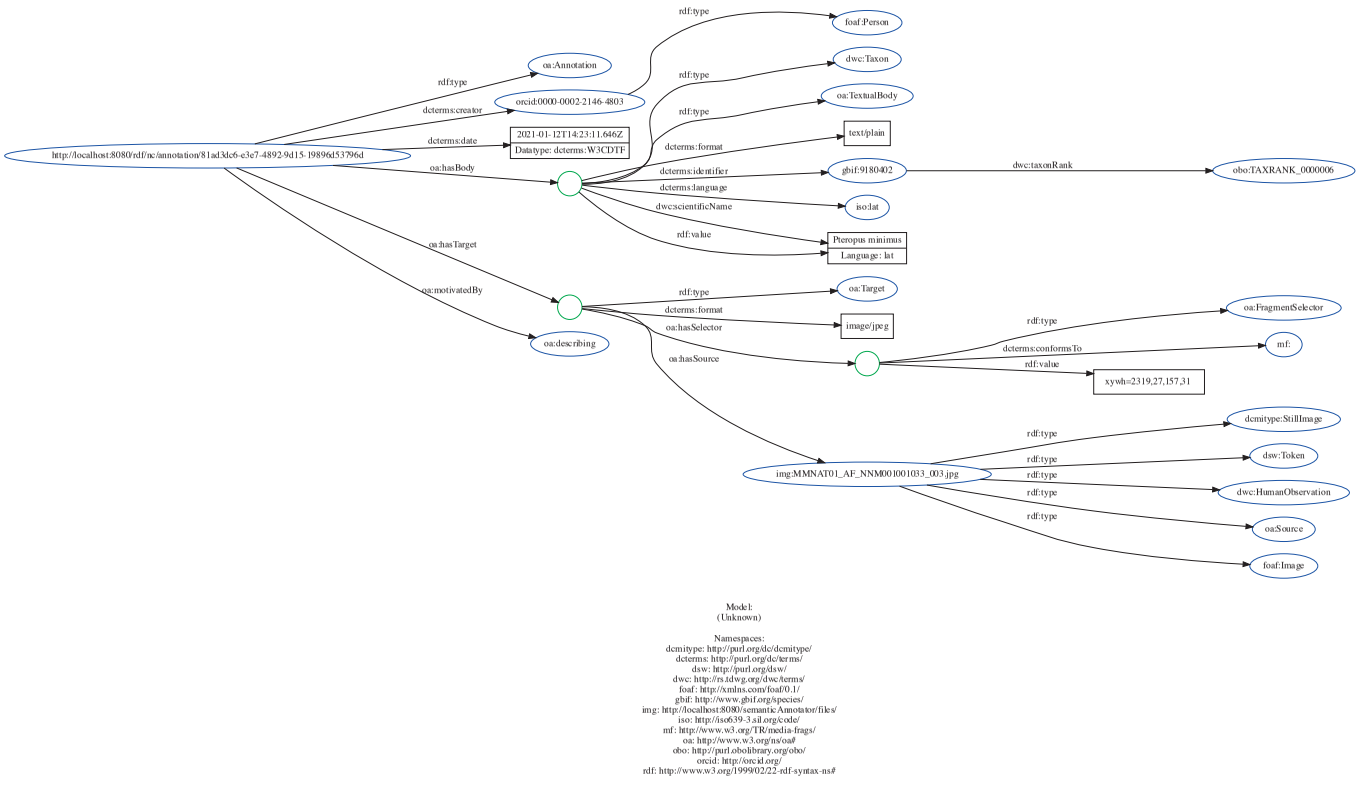

Figure 4.5: Example semantic annotation of an annotation, using the refined data model. Courtesy: A. Kuzniar (2021). Other example annotations can be found here: `https://github.com/LINNAE-project/SFB-Annotator/tree/master/doc/models`

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

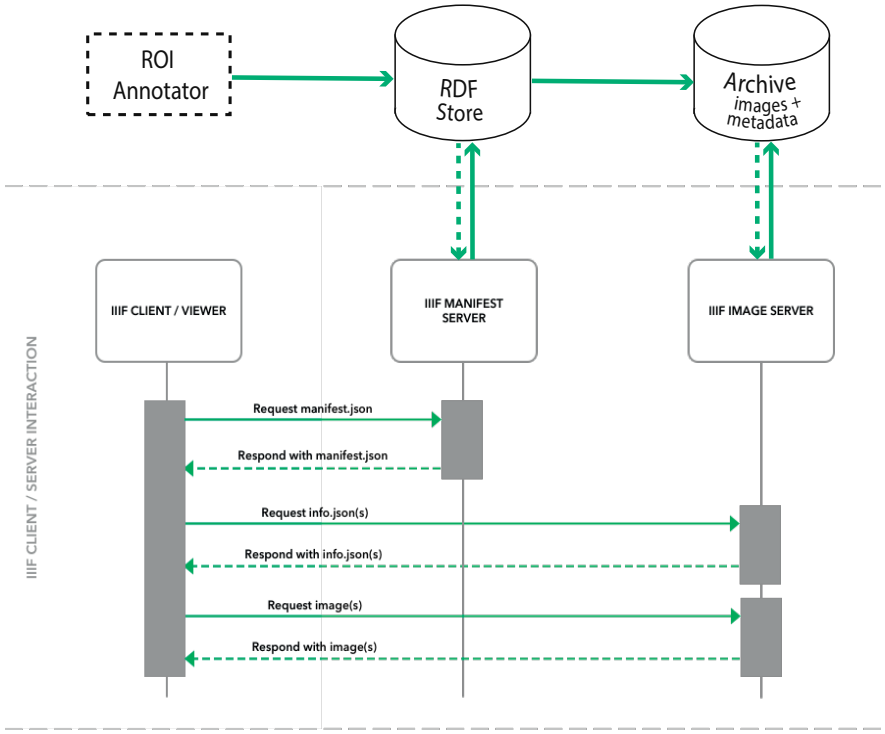

Figure 4.6: Proposed architecture of the SFB-Annotator.[4] Courtesy: A. Kuzniar (2020)

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

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

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

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