# A Workflow for the Semantic Annotation of Field Books and Specimen Labels

Lise Stork[‡], Andreas Weber[§], Eulàlia Gassó Miracle[|], Katherine Wolstencroft[‡]

‡ Leiden Institute of Advanced Computer Science, Leiden, Netherlands
§ University of Twente, Twente, Netherlands
| Naturalis Biodiversity Center, Leiden, Netherlands

Corresponding author: Lise Stork (l.stork@liacs.leidenuniv.nl)

## Abstract

Geographical and taxonomical referencing of specimens and documented species observations from within and across natural history collections is vital for ongoing species research. However, much of the historical data such as field books, diaries and specimens, are challenging to work with. They are computationally inaccessable, refer to historical place names and taxonomies, and are written in a variety of languages.

In order to address these challenges and elucidate historical species observation data, we developed a workflow to

(i) crowd-source semantic annotations from handwritten species observations,

(ii) transform them into RDF (Resource Description Framework) and

(iii) store and link them in a knowledge base.

Instead of full-transcription we directly annotate digital field books scans with key concepts that are based on Darwin Core standards. Our workflow stresses the importance of verbatim annotation. The interpretation of the historical content, such a resolving a historical taxon to a current one, can be done by individual researchers after the content is published as linked open data. Through the storage of annotion provenance, who created

the annotation and when, we allow multiple interpretations of the content to exist in parallel, stimulating scientific discourse.

The semantic annotation process is supported by a web application, the Semantic Field Book (SFB)-Annotator, driven by an application ontology. The ontology formally describes the content and meta-data required to semantically annotate species observations. It is based on the Darwin Core standard (DwC), Uberon and the Geonames ontology. The provenance of annotations is stored using the Web Annotation Data Model. Adhering to the principles of FAIR (Findable, Accessible, Interoperable & Reusable) and Linked Open Data, the content of the specimen collections can be interpreted homogeneously and aggregated across datasets. This work is part of the Making Sense project: makingsenseproject.org. The project aims to disclose the content of a natural history collection: a 17,000 page account of the exploration of the Indonesian Archipelago between 1820 and 1850 (Natuurkundige Commissie voor Nederlands-Indie)

With a knowledge base, researchers are given easy access to the primary sources of natural history collections. For their research, they can aggregate species observations, construct rich queries to browse through the data and add their own interpretations regarding the meaning of the historical content.

## Keywords

Linked Data, Biodiversity, Natural History Collections, Ontologies, crowd-sourcing, Semantic Annotation, History of Science

## Presenting author

Lise Stork