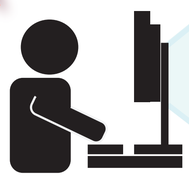
# Making Sense of Illustrated Handwritten Archives

Andreas Weber*[1], Mahya Ameryan*[2], Lise Stork*[3], Katherine Wolstencroft [3], Eulàlia Gassó Miracle [5], Siegfried Nijssen [3], Marco Wiering [2], Maarten Heerlien [5], Michiel Thijssen, Marti Huetink [4], Fons Verbeek [3], Aske Plaat [3], Joost Kok [3], Lissa Roberts [1], Jaap van den Herik [6], Lambert Schomaker [2].

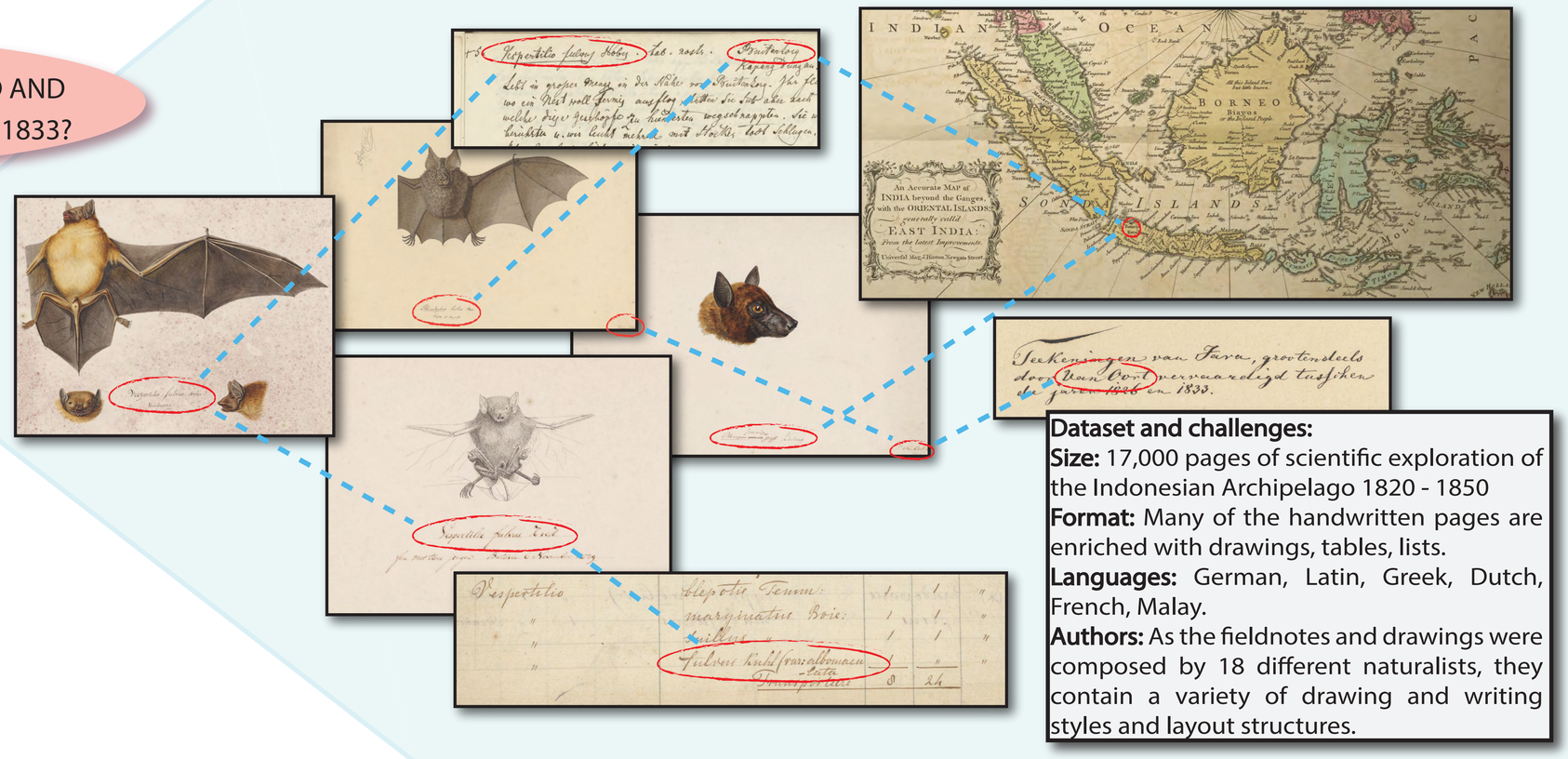
MAKING SENSE realizes a technologically advanced and user-friendly digital infrastructure to open up, enrich and connect illustrated handwritten archives. It combines both image and textual recognition, and allows for an integrated study of underexplored digitized scientific collections. This approach is applicable across the cultural heritage domain and is demonstrated using a 17,000 page account of the exploration of the Indonesian Archipelago between 1820 and 1850 ("Natuurkundige Commissie voor Nederlands-Indië"). This poster provides a project overview, presents the infrastructure's basic layout and sketches its realization in the period 2016-2020. Funding for this project is provided by the Netherlands Organization for Scientific Research (NWO) and BRILL publishers.

## QUERY

WHICH BAT SPECIES WERE COLLECTED AND DRAWN IN JAVA IN THE PERIOD 1820 - 1833?



- Date
- Person
- Visual features
- Species names
- Place
- ....

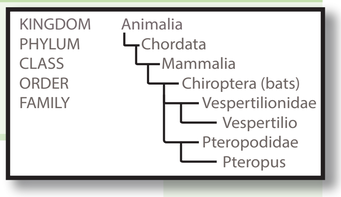



**Dataset and challenges:**
**Size:** 17,000 pages of scientific exploration of the Indonesian Archipelago 1820 - 1850
**Format:** Many of the handwritten pages are enriched with drawings, tables, lists.
**Languages:** German, Latin, Greek, Dutch, French, Malay.
**Authors:** As the fieldnotes and drawings were composed by 18 different naturalists, they contain a variety of drawing and writing styles and layout structures.

## PROCESS

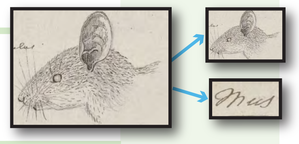
### Lexicon & Ontologies

Identify and construct vocabularies and ontologies that can be used as background knowledge and the formal representation of these resources.




### Preprocessing

Apply noise removal, binarization and normalization on page images.

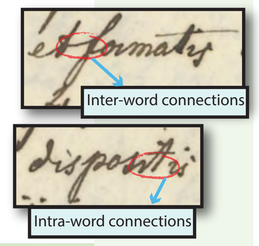


### Layout Analysis

Extract regions of interests (ROI) from document images through a geometrical and logical analysis.

### Text and Picture Recognition

Recognize page segments and form hypotheses about their content. The historical collection contains text, drawings of animals and plants and tables with numerical data. The challenge is to extract as much information from a scanned image. We will use layout analysis and segmentation to arrive at text and object classification using (deep) machine learning. Already the low-level problem of segmentation requires knowledge.




### Integration

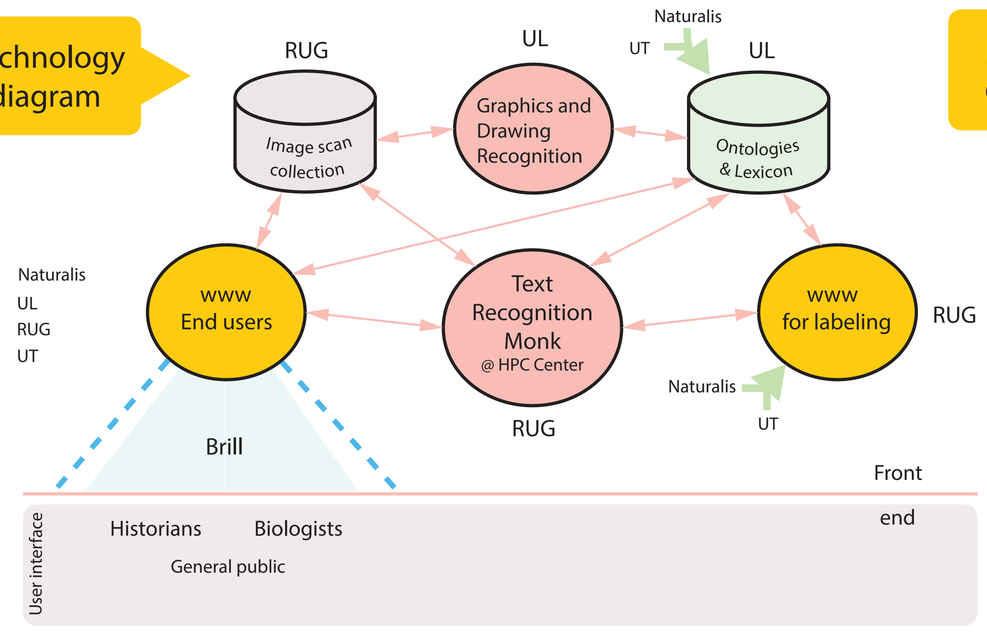
Select background knowledge that can be used to improve the accuracy of the recognition process. Develop algorithms based on probabilistic logic programming to integrate background knowledge, candidate words and candidate images.

### Outreach

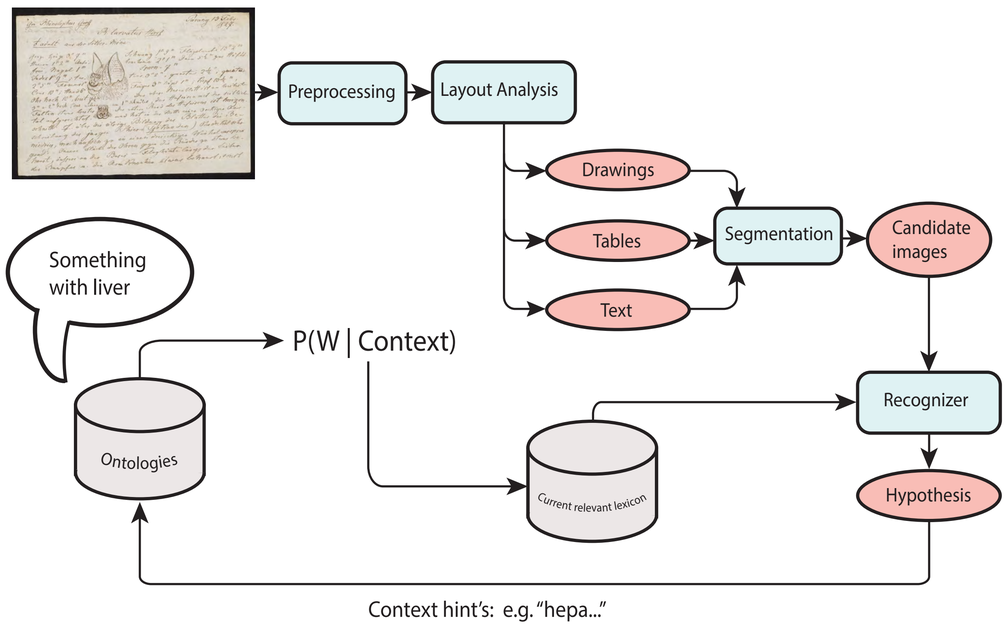
Publish extracted knowledge as Linked Data. Cross-match enriched results with Naturalis specimen collection databases as well as other cultural heritage resources.

## INFRASTRUCTURE

Technology diagram




Process diagram




* These authors have made equal contributions to this poster and the accompanying screencast.

For more information, see also: www.brill.com/makingsense

[1] STePS, University of Twente (UT)
[2] ALICE, University of Groningen (RUG)
[3] LIACS, Leiden University (UL)
[4] Brill publishers (Leiden, the Netherlands)
[5] Naturalis Biodiversity Center (Naturalis)
[6] LCDS, Leiden University (UL)