# Towards Semantic Enrichment of Newspapers: A Historical Ecology Use Case

Marieke van Erp[1], Thomas van Goethem[2], Katrien Depuydt[3], and Jesse de Does[3]

[1] Vrije Universiteit Amsterdam
marieke.van.erp@vu.nl
[2] Radboud University Nijmegen
tgoethem@science.ru.nl
[3] Instituut voor de Nederlandse Taal
{Katrien.Depuydt,Jesse.Dedoes}@ivdnt.org

**Abstract.** Historical ecology research relies on historical accounts of human-animal interactions to study this interaction through space and time. Newspaper archives are a rich source of information, but require careful querying and filtering to collect the relevant information. Traditionally, this is a laborious manual task. In this position paper, we describe our ongoing work on semantically enriching a newspaper collection to create a knowledge base to support historical ecological work.

**Keywords:** text enrichment, historical ecology, lexicology

## 1 Introduction

Historical research often relies on manual inspection of documents. Historical ecology investigates, a.o., the occurrence of particular animals in a distinctive region over time. When using a large newspaper corpus, this would mean having to sift through thousands of documents to identify whether each document is relevant or not, before the researcher can even begin to analyse the content at a more detailed level. In this position paper, we present ongoing work on the creation of a knowledge base that provides a semantic enrichment layer over a large newspaper corpus. We highlight the challenges we discovered in our first data analysis as well as the solutions we intend to implement to resolve these.

## 2 Background and Related Work

Historically, humans have had an ambivalent relationship with animals, perceiving animals not only as sources of food, tools or totems, but also as threats and nuisances. Many birds, small mammals and insects were believed to carry diseases or to be harmful to crops or livestock. Furthermore, large predatory species (e.g. wolf) or venomous species (e.g. viper) were feared for injuring or killing humans [1]. These perceived threats have led to a 'cultural fear' of pest

species, which has been reinforced through storytelling and mythology [2]. Recently, our relationship to many of these so-called "vermin" species has changed. Some species are now valued as key species in nature rehabilitation, while others are reintroduced to our country. It is therefore becoming increasingly relevant to understand how these historical relationships between man and nature relate to the present time. A comprehensive historical study on pest and nuisance species is lacking for the Netherlands. Newspapers reporting on interactions with pest and nuisance species may be an important source of information for such a study.

Currently, the majority of such newspaper analyses are done manually. They involve sending a query to the newspaper interface and clicking every article link, reading the article and recording whether it is relevant to the research question or not. We propose to automate the classification of newspaper articles and storing the results in a knowledge base that contains structured, semantic information about and extracted from the articles along with a link to the original articles[4] to help researchers focus their time on a deeper analysis of the relevant articles.

## 3 Resources

A mix of unstructured and structured resources is used. The newspaper corpus is the main information source, but structured resources to systematically query the newspapers and inform the language technology tools are used.

**National Library Newspaper Corpus** The Dutch National Library has made available the original texts from 1.3 million newspapers, 1.5 million magazine pages and 320,000 books from the 15th to the 21st century through the Delpher portal.[5] We focus on newspaper articles published between 1800 and 1940, for two reasons: 1) The OCR quality on these is most likely better than on the older material and 2) This period also saw the "biological reveil", a reawakening of interest in biological, in the Netherlands, which also may be reflected in mentions of animals in newspapers [3].

**Taxonomic Resources and Lexicons** A list of pest and nuisance species compiled in the ATHENA project,[6] is used, which provides the latin name and its common vernacular name. However, due to the local and temporal variance in animal names we also employ diachronic lexicons that each contain Dutch language variations across time and dialects[7][8] [4].

[4] Due to copyright restrictions it is not possible to include all article texts in the knowledge base, but the articles are freely accessible through the Dutch National Library newspaper portal.

[5] `http://www.delpher.nl`

[6] `http://www.athena-research.org/`

[7] `http://ivdnt.org/onderzoek-a-onderwijs/projecten/gigant`

[8] `http://ivdnt.org/onderzoek-a-onderwijs/projecten/diamant`

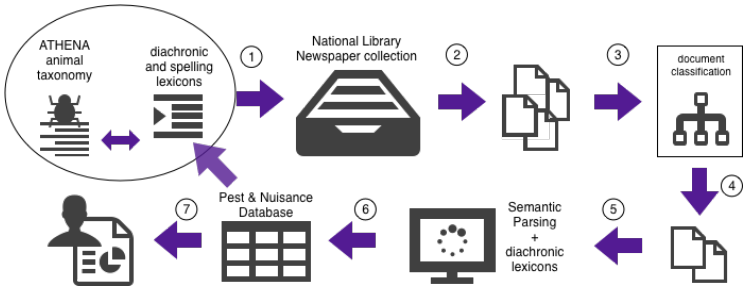


**Fig. 1.** SERPENS Workflow

## 4 Design of the Knowledge Base

We aim to create a knowledge base containing information about animal reports. We start by broadly querying the National Library Newspaper Corpus through the taxonomic and diachronic resources (Step (1) in Figure 1). This results in many articles returned from the newspaper collection that are not necessarily about animals, such as persons with last name "De Wolf") (2). We therefore employ document classification to filter out irrelevant documents (3). Our initial analysis on this is presented in Section 5.

Simply obtaining a set of relevant documents (4) is already useful to ecological historians, but we would like to dive further into the documents and classify what type of animal report it is (5). We will investigate what level of specificity the tools can handle. This results in a knowledge base that contains document classifications, links to the Delpher sources, animal mentions, its spelling variations, document metadata such as publication date and article length, and factual information extracted from the documents (6). New mentions will be fed back to enrich the lexicons.The knowledge base enables humanities and biology researchers to study pest and nuisance species across species, space and time (7).

## 5 European polecats and lynxes

As a first use case, we chose to investigate documents mentioning 'bunzing' (European polecat) and 'lynx' (Lynx). These two species are chosen as a first query on the database, which returns relatively modest result sets (2,515 and 5,530 documents respectively) showing a wide variety of topics in the documents.

**Language variation**
As our research covers a relatively long period of time, as well as a corpus that contains quite some local newspapers, we expect to find a fair amount of language variation. Indeed, through the diachronic lexicons as well as the returned hits, we find a reasonable set of terms to expand our query with. Table 1 lists the

query terms employed for "bunzing" (European polecat), what type of language variation they express and in the last column the number of hits in the corpus.

**Table 1.** Bunzing variations found through the lexicon

| Term | Type | Hits |
|---|---|---|
| Bunzing | base name | 2,515 |
| Bunsing | spelling variation | 1,319 |
| Bonzing | form variation | 47 |
| Bunzel | morphological variation | 617 |
| Bunsel | morphological variation | 67 |
| Ulk | synonym | 21,993 [9] |
| Ulling | synonym | 1,153 |
| Fret[10] | related | 25,830 [11] |
| Eierdief[12] | hyperonym | 98 |

**Document categories**

The hypothesis that will be tested is that the negative perception of native species has subsided with the passing of time, while it has grown for invasive (non-native) species. Categorization helps to dissect the newspaper corpus, making it more easy to measure perception and understand its determinants. The categories have been chosen to be broad enough to be applicable to a broad range of species over an extended period of time, but specific enough for a meaningful analysis.

Initially, we set out to classify whether a document is about the animal or not. However, upon inspecting the document sets, we discovered that documents that may not directly be about the animal, may also be informative and useful to include in the research on species perceptions. For example, an uptick in ads for 'bunzinghonden' (dogs used in the hunt for polecats) or 'bunzingklemmen' ('polecat traps') can be indicative for the species to be a nuisance and thus for people to want to get rid of them. Furthermore, figurative language use that has a positive or negative connotation ('eyes like lynx' or 'stinks like a polecat') also says something about the public perception of a species. Upon annotating about 100 examples with three annotators (consisting of a historical ecologist, a lexicologist and a computational linguist), we came to the following categories, which largely correspond to categories described in [1, 5].

**Natural history** General articles about the animal, e.g. it subsists on birds or x number were stuffed and became part of a museum collection

**Nuisance, material damages** The article mentions the animal as causing material damages, e.g. beetles damaging crops or lynxes killing chickens

[9] 'Ulk' was a German satirical magazine whose cartoons were sometimes republished in Dutch newspapers.

[10] Fret (ferret) is a domesticated polecat

[11] We found OCR errors that map 'het' ('the') to 'fret', further stressing the need for OCR correction and automatic document classification to yield relevant documents.

[12] 'egg thief'

**Nuisance, immaterial damages** The article mentions the animal as a nuisance without material damages e.g. polecats found to walk over someone's face whilst they were in bed, or (possibly irrational) fear for a certain animal
**Pest control** Organised hunt to bring down the number of pest species, e.g. ad for hunting dogs
**Hunt for economic reasons** Hunting to use the fur, meat or other parts of the animal e.g. an article mentioning that the hunting season has started again
**Prevention** Non-lethal actions against pest species, e.g. advice in the newspaper on which plants keep away pest species
**Accidents** Mention of an unintentional encounter with the animal, e.g. roadkill
**Figurative** Figurative language featuring the animal e.g. eyes like a lynx
**Other** Articles not pertaining to the animal, e.g. a ship named 'Lynx' or a person whose last name is 'Bunzing'

**Fact and Fiction**
Another interesting dimension of the dataset is that it does not only cover 'news' but also other types of texts. Currently, the National Library corpus distinguishes 4 types of documents in its newspaper corpus: article, ad, announcement and illustration. In our result set, we also encounter crossword puzzles, feuilletons, poems and cartoons, which are all classified as 'article' in the metadata. This is understandable as the newspaper corpus has been processed largely automatically, but for our purposes it makes sense to distinguish at least between text with an 'imaginative' primary aim (a.o. fiction) and text with an informative primary aim (non-fiction), where we put crossword puzzles in the imaginative category. We are annotating the articles with these classes, and intend to train a classifier from this to automatically detect these categories. We expect that the crosswords are the easiest here (when looking at features such as the occurrence of horizontal, vertical and numbers), but jokes are more difficult to identify automatically e.g.:

*Guest: "Could you perhaps bring me a ferret?"*
*Waitress: "Why would you want one?"*
*Guest: "Perhaps it could find the hare that is hidden in this jugged hare"*[13]

**Document quality**
It is well-known that Optical Character Recognition is not perfect, especially not on older documents [6]. As a first attempt to identify which documents are of highest and lowest quality, we compare each OCRed text to a historical lexicon of known words and return the percentage of words recognised (cf. also [7, 8] for a lexical and a geometrical approach to quality assessment).

## 6 Discussion and Future Work

Semantic Web research revolves around structured data, but for humanities researchers, text is often the core source of investigation. We argue that some of

[13] Arnhemsche Courant, 09-01-1926, `http://resolver.kb.nl/resolve?urn=MMKB08:000106191:mpeg21:a0117`

the manual data collection work for humanities researchers can be alleviated through semantic enrichment of the texts. We propose to use a combination of language technology and structured resources to create a knowledge base as a more sophisticated entry point to collections.

However, working with historical textual collections is not without challenges; in this contribution. we have identified historical language variation, document classification, and document quality as major problems to overcome. By bringing together the knowledge of historical ecology, (historical) lexicography and computational linguistics, we believe we are in the best position to address these issues.

The knowledge base we are creating will be published through Timbuctoo.[14] Besides data access via SPARQL through an API, it provides a programming-free manner to access the data for humanities researchers. Currently, ports to visualisation tools such as Gephi[15] are being built. The annotated data and experiments thusfar can be found at: `http://www.github.com/clariah/serpens`.

## Acknowledgements

The research for this paper was made possible by the CLARIAH-CORE project financed by NWO: `http://www.clariah.nl`

## References

1. Lenders, H.J.R.: Ten a penny? deadly viper bites in the netherlands in a socio-economic perspective. Litteratura Serpentium **34** (2014) 290–316
2. Lenders, H.J.R., I. A. W. Janssen, I.: The grass snake and the basilisk: From pre-christian protective house god to the antichrist. Environment and History **20** (2014) 319 – 346
3. van Berkel, K.: Vóór Heimans en Thijsse: Frederik van Eeden sr. en de natuurbeleving in negentiende-eeuws Nederland. Volume 63. Koninklijke Nederlandse Akademie van Wetenschappen (2006)
4. Maks, I., van Erp, M., Vossen, P., Hoekstra, R., van der Sijs, N.: Integrating diachronous conceptual lexicons through linked open data. Presented at DHBenelux 2016 (9-10 June 2016)
5. Dirke, K.: Where is the big bad wolf? notes and narratives on wolves in swedish newspapers during the eighteenth and nineteenth centuries. In Masius, P., Sprenger, J., eds.: A fairy tale in question. Historical interactions between humans and wolves. The White Horse Press, Cambridge (2015) 101–118
6. Reynaert, M.: Non-interactive ocr post-correction for giga-scale digitization projects. In: CICLing, Springer (2008) 617–630
7. Springmann, U., Fink, F., Schulz, K.U.: Automatic quality evaluation and (semi-) automatic improvement of mixed models for OCR on historical documents. CoRR **abs/1606.05157** (2016)
8. Gupta, A., Gutierrez-Osuna, R., Christy, M., Capitanu, B., Auvil, L., Grumbach, L., Furuta, R., Mandell, L.: Automatic assessment of ocr quality in historical documents. In: Proc. AAAI. Volume in press. (2015)

[14] `https://github.com/HuygensING/timbuctoo`
[15] `https://gephi.org/`