

Text Mining for Historical Documents

Motivation and Case Studies

Caroline Sporleder

Computational Linguistics/MMCI
Universität des Saarlandes

Wintersemester 2011/12

22.02.2012

Museums, archives and libraries possess large collections of data

- artefacts
- books, manuscripts
- meta-data: catalogues, field books, reports etc.

More and more digitisation projects

- governments have come to see CH as a valuable asset
- digitised data can be accessed more easily
- digitisation as a safeguard against data loss

Digitisation offers opportunities

- easier data access (searching, browsing)
- presentation of data (visualisation)
- knowledge discovery
- support for curation (partial automisation, consistency checking)

But to make the most of digitised data, we need sophisticated tools

- information retrieval (data indexing and searching)
- information extraction (linguistic data analysis)
- automatic data linking
- discovery of trends and interdependencies
- data presentation (for experts and non-experts)
- meta-data enrichment (linguistic disambiguation, semantic tagging, automatic transcription of audio data etc.)
- semi-automatic curation (data completion, error detection, consistency enforcement)

⇒ **text mining and natural language processing (NLP) play a big role because much of primary and most meta-data are textual**

Case Study: Naturalis

The Dutch National Museum of Natural History



Naturalis: The Collection (1)



- more than 10 million specimens:
 - 5,250,000 insects
 - 2,290,000 invertebrates
 - 1,000,000 vertebrates
 - 1,160,000 fossils
 - 440,000 stones and minerals
- 150,000 species
- 10% of the Earth's biodiversity

Naturalis: The Collection (2)



For each of the 10M specimens

- a label attached to the specimen, providing basic details (biological name, where and when found, inventory number)
- an entry in a register book
- usually an entry in a field book

Additionally, for many specimens

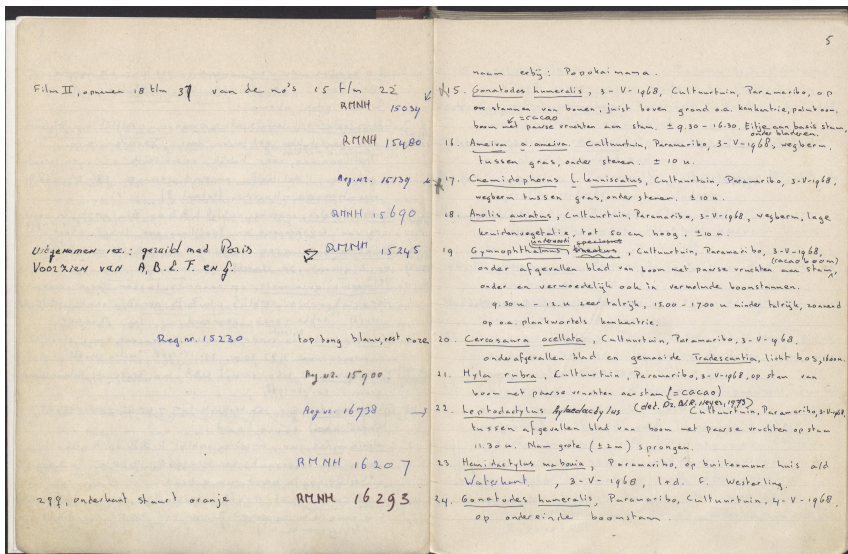
- an entry in a specimen data base
- a photo
- meta-data in the form of research papers etc. written about them

Also:

domain ontologies, taxonomic descriptions, maps etc.

Convert field and register books into data bases

- take high quality digital photos of pages
- transcribe them manually



Digitisation of Fieldbooks

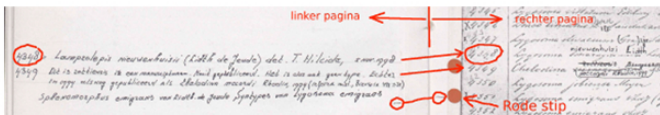


Example: Typist Guidelines

	code	bladzijdenummer
A	3023	121
	<i>Leptodactylus rugosus myersi</i> 2 ex. km 117, 4 km z Barubakreek, 23-ix-1980, 20.30-21.30u. open granierplaat met schaarse vegetatie, omgeving savannebos, zittend op graniet aan rand vegetatie, 100m	
B	3024	
	<i>Hyla minuta</i> 1 ex. km 120, 23-IX-1980, 21.45 u. kwakend aan rand van weg, op striuk, 180 cm boven oppervlak van poeltje, secundair bos, 100m	
C	3025	1 [vrouw]
	Waarnemingen <i>Geochelone denticulata</i> km 119 van Zanderij, 16 km W van Witagron, weg naar Avanavero, dist. Saramacca, Suriname, 26-IX-1980, 9.00u, op weg in gebied met ondergelopen bos, waarnemer M.S. Hoogmoed.	

...
[lemma] 3023 Leptodactylus myersi 2 ex. km 117, 4 km z Barubakreek, 23-ix-1980, 20.30-21.30u. Open granierplaat met schaarse vegetatie, omgeving savannebos, zittend op granier aan rand vegetatie, 100m.
[lemma] 3024 Hyla Minuta 1 ex. km. 120, 23-IX-1980, 21.45 u. kwakend aan rand van weg, op striuk, 180 cm boven oppervlak van poeltje, secundair bos, 100m ex in slechte conditie, werd 's nachts opgegeten door <i>Leptodactylus rugosus</i> (3023) en volgende dag bij doden weer uitgebraakt. gedeeltelijk verteeld.
[lemma] dia's 1980-XII, 2/4 Waarnemingen <i>Geochelone denticulata</i> 1 [vrouw] km 119 van Zanderij, 16 km W van Witagron, weg naar Avanavero, distr. Saramacca, Suriname, 26-IX-1980, 9.00u, op weg in gebied met ondergelopen bos, waarnemer M.S. Hoogmoed.
...

Example: Typist Guidelines



[lemma]

4348

Lygosoma nieuwenhuisii Lidth smarafarimen Lus. Sandakanbaai J. Chr. Prakka

[links]

4348 = Lampeolepis nieuwenhuisii (Lidth de Jeude) det. T. Hilcida, 5 nov. 1998

[lemma]

4349

Chelodina mccodi Rhodin, 1994 (paratype) Rotti: D. H. ten Kate

[rode stip]

[links]

4349 Dit is zottiensis is een manuscriptnaam. Nooit gepubliceerd. Het is dus ook geen type. Echter in 1994 alsnog gepubliceerd als Ehelodina mccodrdi Rhodin, 1994 (referred mat., Breviora 498:1-31)

[lemma]

4350

Lygosoma jobiense Majer Salawatti Bernstein

[lemma]

X 4351

Lygosoma emigrans vhdj (types) Somba D'H. ten Kate

[rode stip]

[links]

Sphenomorphus emigrans van Lidth de Jeude Syntypes van Lygosoma emigrans

...

Example: Typist Guidelines



5 | Parasitus fimetarius (Berl. 19) Nph. II dors. vent. Faure 1445 | RMNH Acari P3036 | in mest Weimar III.19

Transcription of Fieldbooks

- all fieldbooks relating to Reptiles and Amphibians Collection
- 15,000 handwritten pages
- manually transcribed by typists
- simple guidelines on how to deal with
 - non-ASCII characters
 - text written in the margins
 - illegible passages
 - etc.
- transcriptions completed in around 8 months
- <5% error rate

1 ex. *Phyllobates femoralis* At base of tree on small island, primary forest, 20.45-22.00 u. RMNH 23865

Lithodytes lineatus, Brownsberg, aan voet, onder stuk rot hout, 13.07.1968, 8.45 u., RMNH 26076

Dorsolateraal strepen heldergeel, tekening op dijen vuurrood, veel feller als bij *Phyllobates femoralis*.

Gonyocephalus auritus Meyer, 3 ex. (1 juv.), Misool. Hoedt 1867.

RMNH 17656 *Eleutherodactylus zeuctotylus* 1 [vrouw] Lelygebergte, 4 km N.O. van airstrip, distr. Marowijne, Suriname, 19-VIII-1975, onder stuk hout, 610m, 1 [plus] d M. S. Hoogmoed.

What can you do with it? (1)

 Database Fieldbooks

Registration Number	14655		
Class	Amphibia		
Order	Serpentes	Deviates from expected value 'Reptilia' (accuracy of ~99%)	
Family	Colubridae		
Genus	Psammophis		
Species	sibillans		
Sub Species			
No. of Specimens	1		
Sex			
Storage Method	alcohol		
Special Remarks	Geen verdere gegevens bekend.		
Attribute			
Collector	Buttkofer, J.	Collection Date	- -1881
Label Data		Collection Number	944
Country	Liberia	Country ID	132
Province/State		Altitude	
Place	Schieffelin's ville	Coordinates	
Biotope		Determinator	
Location		Determination Date	
Author	(Linnaeus, 1758)	Recorder	Grouw, H.J. van
Publication		Record Date & Time	2000-06-26
Printed	j	Inventory Number	0
Globally Unique ID	{BB98DD11-4B3E-11D4-A2CB-00104BCC2C29}	Expedition	buttkoferliberia1881



What can you do with data? (2)

Psammophis sibilans

Database
Fieldbooks

Search

Results 1-8 of 8 for 'Psammophis sibilans'



14655 - *Psammophis sibilans*

Taxonomy : Amphibia - Serpentes - Colubridae - *Psammophis sibilans*

Location : Schieffelin's villa, Liberia

Collection Date : -1881

[Google Images](#) [Specimen Photo](#) [Google Image](#) [Wikipedia](#)

Corrections available for klasse



14652 - *Psammophis sibilans*

Taxonomy : Reptilia - Serpentes - Colubridae - *Psammophis sibilans*

Location : Robertsport, Grand Cape Mount, Liberia

Collection Date : 31-07-1881

[Google Images](#) [Google Image](#) [Wikipedia](#)



14649 - *Psammophis sibilans*

Taxonomy : Reptilia - Serpentes - Colubridae - *Psammophis sibilans*

Location : Grand Cape Mount, Liberia

Collection Date : 11-10-1881

[Google Images](#) [Google Image](#) [Wikipedia](#)



14653 - *Psammophis sibilans*

Taxonomy : Reptilia - Serpentes - Colubridae - *Psammophis sibilans*

Location : Robertsport, Grand Cape Mount, Liberia

Collection Date : 09-08-1881

[Google Images](#) [Google Image](#) [Wikipedia](#)



14654 - *Psammophis sibilans*

