# Text Mining for Historical Documents
# Motivation and Case Studies

Caroline Sporleder

Computational Linguistics/MMCI
Universität des Saarlandes

Wintersemester 2010/11

21.02.2011

# IT and Cultural Heritage: Why bother? (1)

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

# IT and Cultural Heritage: Why bother? (2)

Digitisation offers opportunities

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

# IT and Cultural Heritage: Why bother? (3)

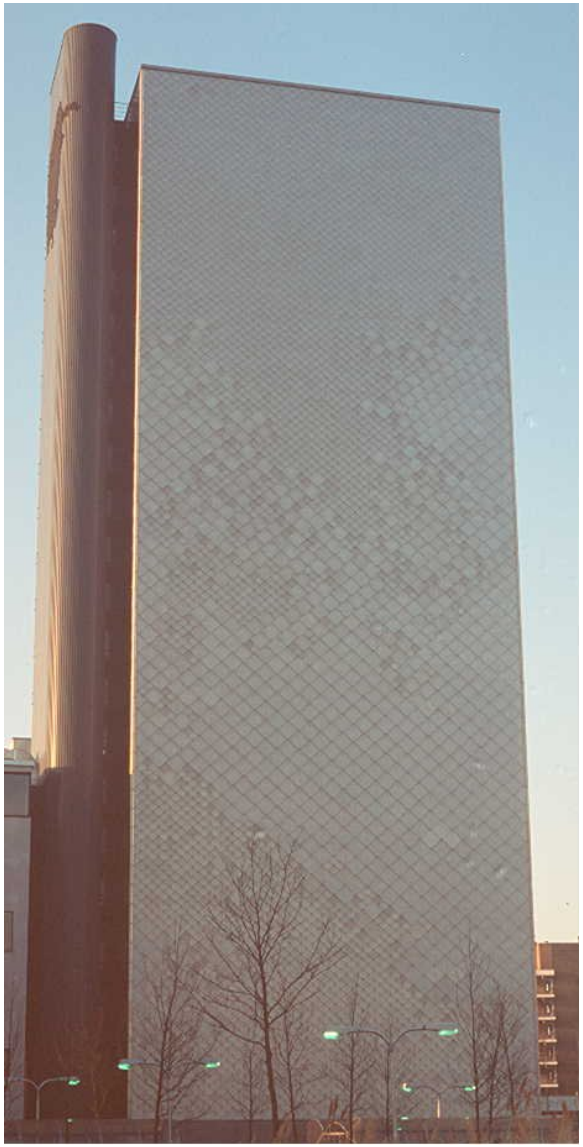
But to make the most of digitised data, we need sophisticated tools

- 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)

$\Rightarrow$ **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)

# Data and Meta-Data

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:

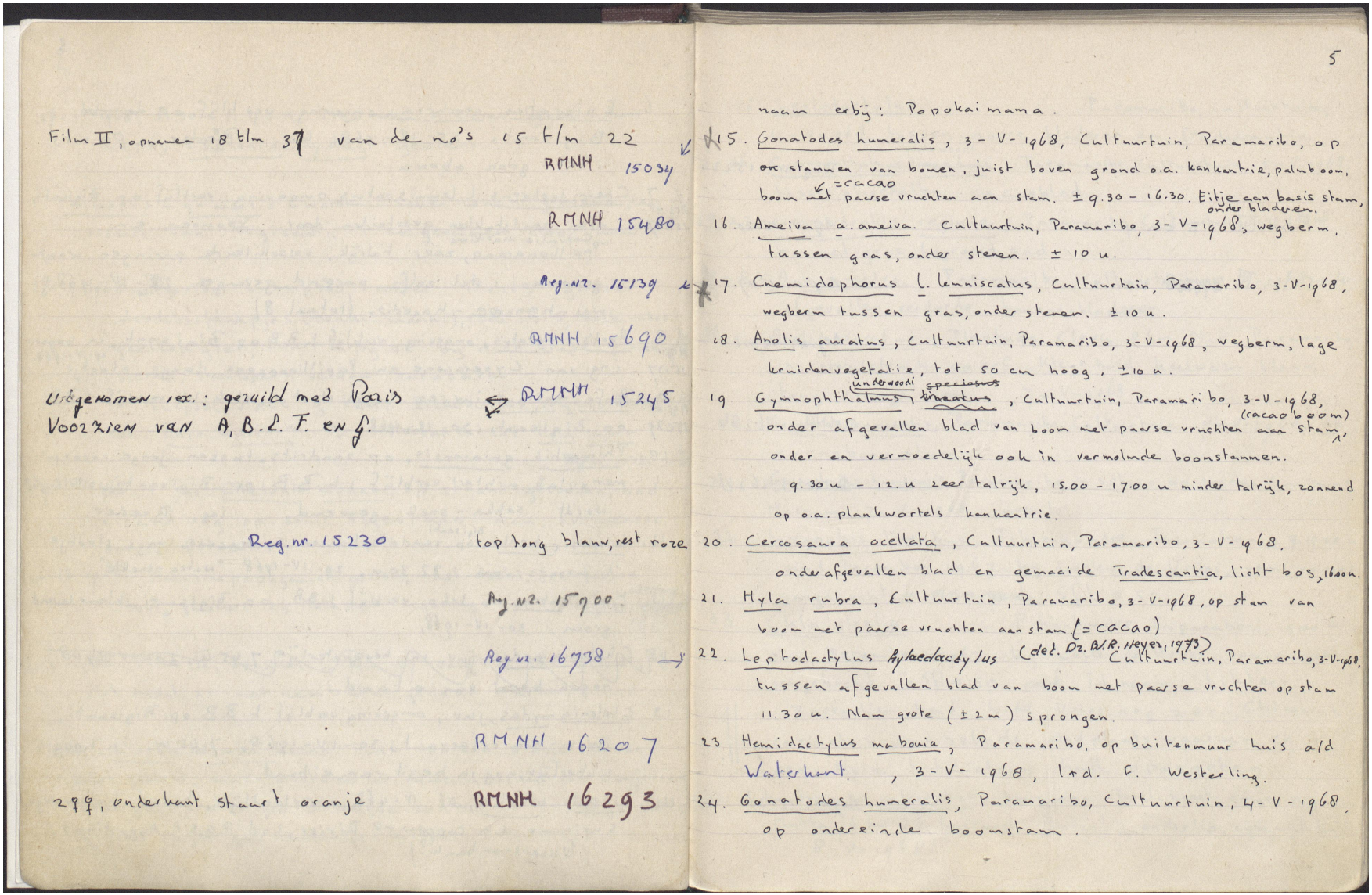
domain ontologies, taxonomic descriptions, maps etc.

# Digitisation Efforts

Convert field and register books into data bases

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

# Digitisation of Fieldbooks



Film II, opnamen 18 t/m 37 van de no's 15 t/m 22

RMNH 15034

RMNH 15480

Reg.nr. 15139

RMNH 15690

Uitgenomen 1ex.: gezuid met Paris
Voorzien van A, B. L. F. en g.

RMNH 15245

Reg.nr. 15230

top long blauw, rest roze

Reg.nr. 15900

Reg.nr. 16738

RMNH 16207

2 ♀♀, onderkant staart oranje

RMNH 16293



naam erbij: Popokai mama.

15. Gonatodes humeralis, 3-V-1968, Cultuurtuin, Paramaribo, op en stammen van bomen, juist boven grond o.a. kankantrie, palmboom, boom met paarse vruchten aan stam (=cacao). ± 9.30 - 16.30. Eitje aan basis stam, onder bladeren.

16. Ameiva a. ameiva. Cultuurtuin, Paramaribo, 3-V-1968, wegberm, tussen gras, onder stenen. ± 10 u.

17. Chemidophorus l. lemniscatus, Cultuurtuin, Paramaribo, 3-V-1968, wegberm tussen gras, onder stenen. ± 10 u.

18. Anolis auratus, Cultuurtuin, Paramaribo, 3-V-1968, wegberm, lage kruidenvegetatie, tot 50 cm hoog, ± 10 u.

19. Gymnophthalmus (onderwoudi speciosus lineatus), Cultuurtuin, Paramaribo, 3-V-1968, onder afgevallen blad van boom met paarse vruchten aan stam (cacaoboom), onder en vermoedelijk ook in vermolmde boomstammen.
9.30 u - 12. u zeer talrijk, 15.00 - 17.00 u minder talrijk, zonnend op o.a. plankwortels kankantrie.

20. Cercosaura ocellata, Cultuurtuin, Paramaribo, 3-V-1968. onder afgevallen blad en gemaaide Tradescantia, licht bos, 16 u.

21. Hyla rubra, Cultuurtuin, Paramaribo, 3-V-1968, op stam van boom met paarse vruchten aan stam (=cacao)

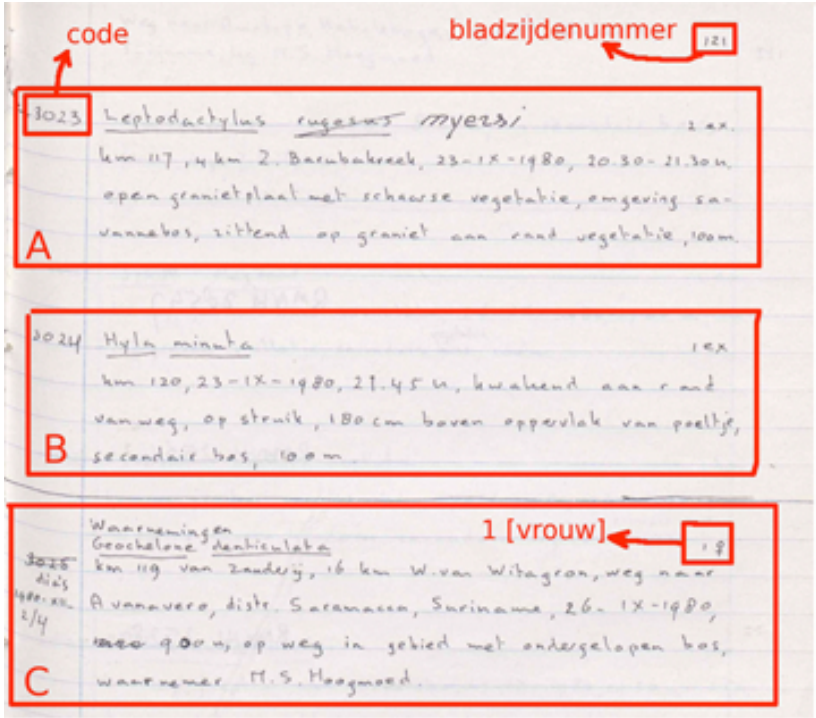
22. Leptodactylus hylaedactylus (det. Dr. W.R. Heyer, 1973) Cultuurtuin, Paramaribo, 3-V-1968, tussen afgevallen blad van boom met paarse vruchten op stam 11.30 u. Nam grote (± 2 m) sprongen.

23 Hemidactylus mabouia, Paramaribo, op buitenmuur huis a/d Waterkant, 3-V-1968, l.d. F. Westerling.

24. Gonatodes humeralis, Paramaribo, Cultuurtuin, 4-V-1968, op onderzijde boomstam.

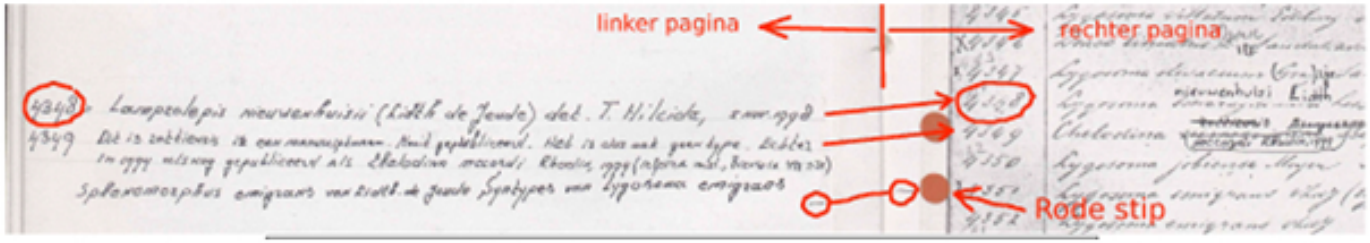
# Digitisation of Fieldbooks

# Example: Typist Guidelines




```
...

[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, secondair bos, 100m

ex in slechte conditie, werd 's nachts opgegeten door Leptodactylus rugosus
(3023) en volgende dag bij doden weer uitgebraakt. gedeeltelijk verteeld.


[lemma]
dia's 1980-XII_ 2/4
Waarnemingen
Geochelone denticulata
1 [vrouw]
km 119 van Zanderij, 16 km W.van Witagron, weg naar Avanavero, distr. Saramacca,
Suriname, 26-IX-1980, 9.00u, op wef in gebied met ondergelopen bos, waarnemer
M.S. Hoogmoed.

...
```

# Example: Typist Guidelines




```
[lemma]
4348
Lygosoma nieuwenhuisi 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 nanusciptnaam. 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

# Fieldbook Transcript

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, l [plus] d M. S. Hoogmoed.

# What can you do with it? (1)

MITCH mBase

ng information in texts from the cultural heritage

Database
Fieldbooks
Search

| | | | |
|---|---|---|---|
| **Registration Number** | 14655 | | |
| **Class** | Amphibia | | |
| **Order** | Serpentes | | |
| **Family** | Colubridae | | |
| **Genus** | Psammophis | | |
| **Species** | sibillans | | |
| **Sub Species** | | | |
| **No. of Specimens** | 1 | | |
| **Sex** | | | |
| **Storage Method** | alcohol | | |
| **Special Remarks** | Geen verdere gegevens bekend. | | |
| **Attribute** | | | |
| **Collector** | Buttikofer, J. | **Collection Date** | - -1881 |
| **Label Data** | | **Collection Number** | 944 |
| **Country** | Liberia | **Country ID** | 132 |
| **Province/State** | | **Altitude** | |
| **Place** | Schieffelins 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-00104BBC2C29} | **Expedition** | buttikoferliberia1881 |

Deviates from expected value 'Reptilia' (accuracy of ~ 99%)

# What can you do with it? (2)