

FROM FIELD NOTES TOWARDS A KNOWLEDGE BASE

Ervina Cergani
Max Planck Institute für Informatik
Saarland University

From fields notes towards a knowledge base

- Convert text cultural heritage data to a domain-specific knowledge base, using machine learning techniques:
 - Digitized expedition field notes => segmented and labelled automatically (supervised recognition of named entity)
 - Postprocessing with selective sampling
 - Enrichment of the records with secondary metadata
 - Visualizing the results using maps and photos

From fields notes towards a knowledge base

- Cultural heritage institutions (museums):
 - Collection of artifacts
 - Textual data
 - Description of objects
 - Scientific literature
 - Catalogues
- Highly specialized domain => NLP and IE techniques enables enhanced access to this resources.

Segmentation and Labelling of Field Notes

- Material: 80 books containing field notes on collecting reptile and amphibian specimens for the Dutch National Museum of Natural History, *Naturalis*

Gonatodes humeralis, post Tigri, New River, On tree, 2-VII-1968,
16.30 h. RMNH 16314

Segmentation and Labelling of Field Notes

- Three aspects that processing the field notes difficults:
 - The order and length of entering information about the specimens collected or observed is not standardized.
 - Large number of optional information units occur irregularly.
 - The texts are written in a mix of Dutch and English, using domain terminology in Latin.

Segmentation and Labelling of Field Notes

- The Reptile and Amhibian Database => 37 columns
- Our example => fill 7 columns

GENUS	Gonatodes
SPECIES	humeralis
LOCATION	post Tigri , New River
BIOTOPE	On tree
COLLECTDATE	2-VII-1968
COLLECTTIME	16 . 30 h
REGISTRNR	RMNH 16314

Segmentation and Labelling of Field Notes

- Goal: Turn the text field note entries into database records automatically.
- Token-based, supervised, sequence labelling task
=> each token in an entry needs to be marked as belonging to one of the 37 columns.
- Currently: 40'749 field note entries
 - 40 tokens on average.

Segmentation and Labelling of Field Notes

- Experimental Setup:
 - Two supervised machine learning algorithms:
 - Conditional Random Fields (CRF), defines a conditional probability distribution over label sequences.
 - Memory-Based Tagging (MBT), classifies sequences based on stored examples and a frequency-threshold vocabulary.
 - The classifiers were trained on 300 entries and tested on 200 held-out entries, both manually annotated.

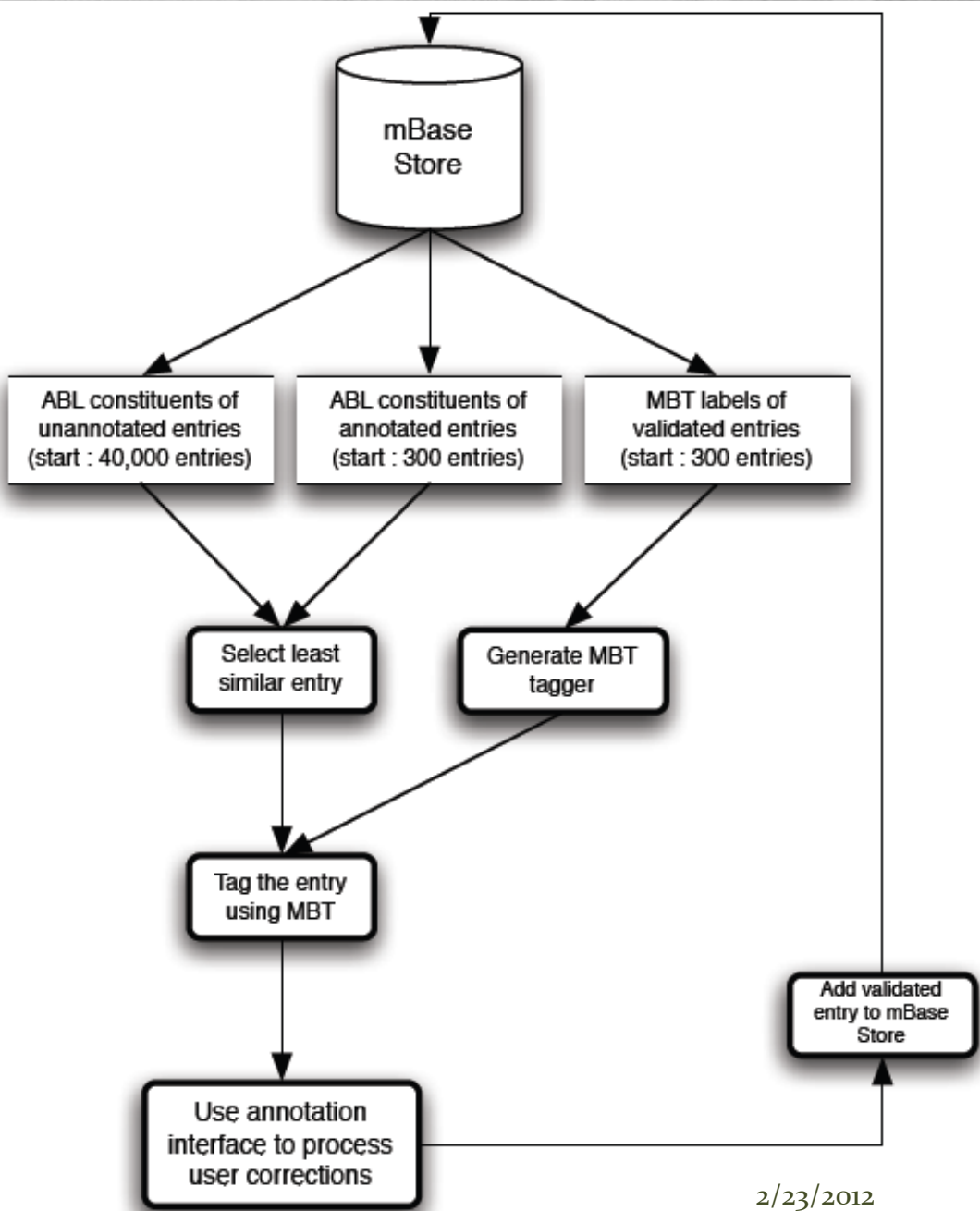
Segmentation and Labelling of Field Notes

- Experimental Results:

	CRF	MBT
accuracy	0.98	0.88
precision	0.71	0.82
recall	0.67	0.86
F score	0.69	0.84

- Analyses on individual entities:

	<u>MBT (F score)</u>		<u>CRF (F score)</u>	
REGISTERNR, LAND	0.97		0.91	LAND
			0.76	REGISTRNR
SPECIAL REMARKS	0.53		0.30	



Interface for semi-automatic post-processing of segmentation and labelling errors on field notes

http://s0135.uvt.nl - Annotation Demo (v0.2.1)

Search

Labels:

<input type="radio"/> AANTAL	<input type="radio"/> KENM	<input type="radio"/> OSEX
<input type="radio"/> AUTH	<input type="radio"/> LAND	<input type="radio"/> OSPEC
<input type="radio"/> BIJ	<input type="radio"/> LOC	<input type="radio"/> OSSPEC
<input type="radio"/> BIO	<input type="radio"/> O	<input type="radio"/> OTYPE
<input type="radio"/> DET	<input checked="" type="radio"/> PLAATS	<input type="radio"/> O/LAAR
<input type="radio"/> DON	<input type="radio"/> PROV	<input type="radio"/> OVDAT
<input type="radio"/> GEN	<input type="radio"/> PUB	<input type="radio"/> OVNUM
<input type="radio"/> HOOGTE	<input type="radio"/> ORNUM	<input type="radio"/> OVTIJD

New Annotation:

Leptodactylus sibilatrix, Paramaribo, in tuin Surinaams Museum, Commewijnestraat 18, tussen lage plantjes. [man] roep op band 1, ?? 0 - ?? 5 opgenomen om 7.30 u. Bij vangen piepte hij wat 25 - IV - 1968 [LEFT] Reg. nr. 15140

Example Sentences:

1 juv. *Eunectes marinus* **Weg naar Anotopo**, Lucie landing, distr. Nickerie, Suriname, 24 - V - 1981, 11.00u, onder korjaal op rivieroever, leg. M. S. Hoogmoed [en] M. Held. RMNH 20769

Gonatodes humeralis 1 [man] **Reynold ' s (Grassalco) kamp aan moezekreek**, distr. Nickerie, Suriname, 20 - VII - 1975, 13.00 u, op basis grote boom, Kreekoever, hoog bos, 90 m, I [en] d. M. S. Hoogmoed [en] W. N. Polder. RMNH 28202

Anolis fusco - auratus, **Sipaliwini, 0 - 5 km N. van basiskamp**, op boom, [plusminus] 3 m boven grond, 2 - IX - 1968, 9.30 u. RMNH 15657

Anolis auratus 1 ex. **Bombaystraat, Benie ' s Park, Paramaribo**, Distr. Suriname, Suriname, 31 - XII - 1974, 20.00 21.00 u, slapend op grashalm, moerassig stukje land tussen huizen, 10 cm boven wateroppervlak, 0 m, I [en] d. M. S. Hoogmoed. RMNH 26819

Eleutherodactylus chiastonotus 1 ex. **Loekreek, kamp Hofwijks IV, km van airstrip Oelemari**, distr. Marowijne, Suriname, 10 - VIII - 1975, 11.15 u, bosgrond, tussen houtsnippers bij omgehakte boom, hoog bos, kreekoever, 120 m, I [en] d. M. S. Hoogmoed [en] C. W. Myers. RMNH 17638

Inducing Metadata

- Goal => Introduce more structure into the knowledge base.
- SPECIALREMARKS field of the database:
 - Slides MSH 1975-xviii-27/29, 1975-xix-20/25; tape recording 1975 II B 297-304. Acquired as gift from the British Museum (Nat. Hist.), BMNH 1975. 1348.

Inducing Metadata

	SPECRA	BIORA
# sents	2,641	694
# words / sent	11.8	6.5
# tokens	2,570	1,090
# production rules	5,305	1,402
# non-terminals	62,574	10,533
# terminals	1,703	886
# candidates	181	209
# accepted candidates	29	20
New metadata examples	born died formerly length loan museum obtained photo slide tank university	bush creek forest ground pool river road rock swamp vegetation water

2/23/2012

Disclosing Data

- Goal of the integrated system mBase => provide easier and more intuitive access to data from the museum.
- mBase runs on an open source XML management system, eXist.
- Access through keyword search across the whole knowledge base or via specific search on individual fields.
- Possible to browse revisions and to search only particular versions of the knowledge base.
- Linking to various other resources.

The search and visualization display of knowledge base

MITCH mBase

Mining information in texts from the cultural heritage

http://s0135.uvt.nl - testudo graeca - mBase

testudo graeca Database Search Fieldbooks

Specific Search

Previous 20 < Results 1-8 of 8 for 'testudo graeca' > Next 20

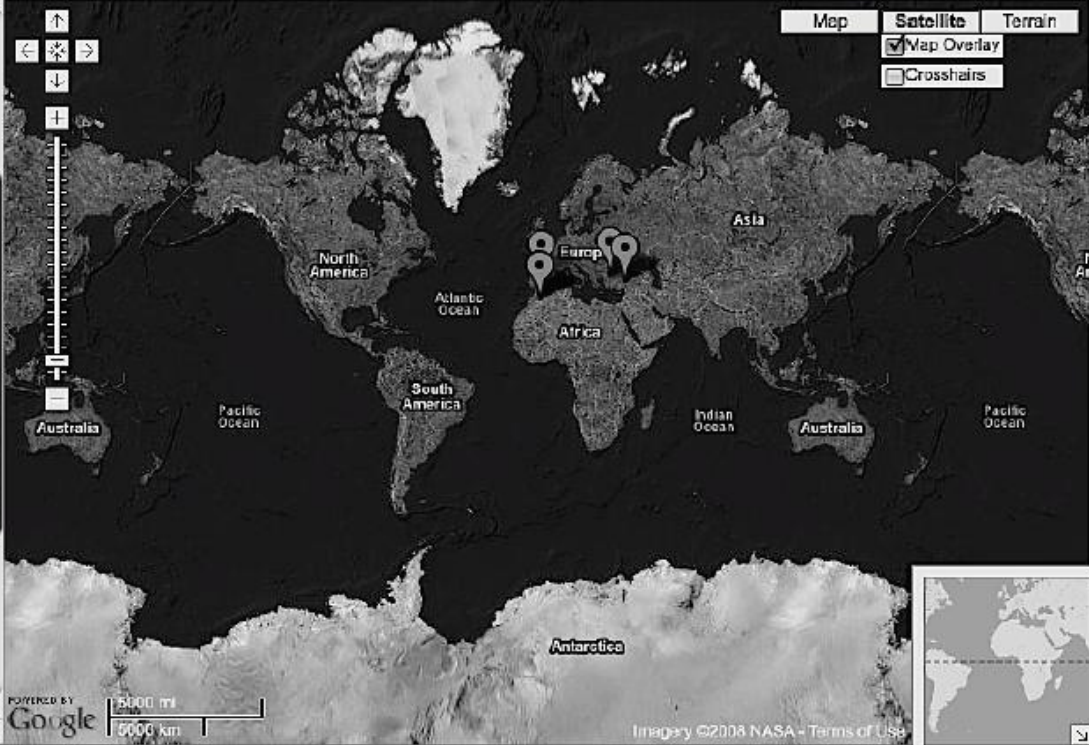
27831 - Testudo graeca
Taxonomy: (Reptilia - Testudines, Cryptodira - Testudinidae - Testudo - graeca - iberica)
Location: Aziatisch Turkije, 50 km. Z. W. van Ankara, 1000 m, Turkey
Collection Date: 17-6-59
Google Images Specimen Photo 1 Google Image Wikipedia

28866 - Testudo graeca
Taxonomy: (Reptilia - Cryptodira - Testudinidae - Testudo - graeca - graeca)
Location: 8 km O. van Smimou, Morocco
Collection Date: 25-02-1971
Google Images 1 Google Image Wikipedia

28867 - Testudo graeca
Taxonomy: (Reptilia - Cryptodira - Testudinidae - Testudo - graeca - graeca)
Location: 6 km O. van Smimou, Morocco
Collection Date: 25-02-1971
Google Images 1 Google Image Wikipedia

28868 - Testudo graeca
Taxonomy: (Reptilia - Cryptodira - Testudinidae - Testudo - graeca - graeca)
Location: 22 km NW. van Ksar-el-Kebir, Morocco
Collection Date: 21-04-1971
Google Images 1 Google Image Wikipedia

24006 - Testudo graeca



Map Satellite Terrain
Map Overlay
Crosshairs

North America Atlantic Ocean South America Pacific Ocean Indian Ocean Australia Antarctica

POWERED BY Google 5000 mi 5000 km Imagery ©2008 NASA - Terms of Use

The record view of the knowledge base

MITCH mBase

http://ls0135.uvt.nl - 20643 / *Tretioscincus agilis* - mBase

Database
 Fieldbooks

Specific Search

Registration Number	20643		
Class	Amphibia		
Order	Sauria	Deviates from expected value 'Raptalia' (accuracy of ~99%)	
Family	Gymnophthalmidae		
Genus	Tretioscincus		
Species	agilis		
Sub Species		Specimen Image not available	
No. of Specimens	1		
Storage Method	m alcohol		
Special Remarks			
Attribute			
Collector	Hoogmoed, M.S. & Polder, W.N.	Collection Date	14-08-1975
Label Data		Collection Number	1975-MSH1633
Country	Suriname	Country ID	220
Province/State	Marowijne	Altitude	650
Place	Lely, S. of Airstrip	Coordinates	
Biotope		Determinator	
Location		Determination Date	
Author	(Ruthven, 1916)	Recorder	Grouw, H.J. van
Publication		Record Date & Time	2001-07-16 09:58:22
Printed	j	Inventory Number	0
Globally Unique ID	{4A715F04-79D0-11D5-A2CB-00104BCC2C29}	Expedition	mhsurg1975
fieldbook Text	Tretioscincus agilis1 [man] Lelygebergte , tussen kamp IV en airstrip , Z . van airstrip , distr . Marowijne , Suriname , 14 - VIII - 1975 , 11 . 15 u , op basis boom , tussen Lianen en kruiden , hoog bos , 650 m , [en] d . M . S . Hoogmoed en W . N . Polder [LEFT] Reg . nr . 20643		

Submit Changes

2/23/2012

16

Concluding remarks

- Converting cultural heritage text => searchable knowledge base.
 - Segmentation of large amounts of texts into database fields.
 - An annotation tool based on selective sampling is implemented to postprocess the results.
 - A database field expansion method to create metadata.
 - The interface of the knowledge base.
- Follow-up work: Report on the design of the ontology underlying the knowledge base.

Thank you for your attention!

