

From Field Notes Towards a Knowledge Base

Piroska Lendvai, Steve Hunt

Department of Communication and Information Science
Tilburg University, The Netherlands
{p.lendvai,s.j.hunt}@uvt.nl

Abstract

We describe the process of converting plain text cultural heritage data to elements of a domain-specific knowledge base, using general machine learning techniques. First, digitised expedition field notes are segmented and labelled automatically. In order to obtain perfect records, we create an annotation tool that features selective sampling, allowing domain experts to validate automatically labelled text, which is then stored in a database. Next, the records are enriched with semi-automatically derived secondary metadata. Metadata enable fine-grained querying, the results of which are additionally visualised using maps and photos.

1. Introduction

Besides housing collections of artifacts, cultural heritage institutions such as museums store a vast amount of textual data, for example descriptions of objects, scientific literature, and catalogues. Even if such texts come from a highly specialised domain, featuring specific vocabulary and language use, it is possible to successfully apply general NLP and IE techniques to enable enhanced access to these raw resources. In this paper we focus on processing domain-specific textual data using general machine learning techniques, to create several components of a knowledge base. The main softwares used to this end are the Memory-Based Tagger (MBT) and the Alignment-Based Learner (ABL) algorithms, both freely available for research purposes.

First, we will describe the procedure of joint segmentation and labelling of so-called *field notes* from the nature history domain; this can be likened to supervised recognition of named entities. The main difference with previous work on NE recognition is that field book entries tend to merely list the various entities, thus we need to deal with syntactically elliptic texts, while precise identification of unit boundaries is even more crucial than in traditional NER. Below is the original text from a random field note entry:

```
Plethodon cinereus 1 [female] [plus]
12 juvs. Peaks of Otter, Blue Ridge
Parkway, Bedford County, Virginia,
U.S.A., 750 m, 7-VIII-1982,
12.45-14.00 h, Sharp Top Trail,
in mixed forest (predominantly
deciduous) on mountain slope, inside
hollow in rotten log, leg. M.S.
Hoogmoed. Juveniles were still
in capsules at time of capture,
but started hatching immediately
after capture, till 14.00 p.m.
on 8-VIII-1982. [bookLEFTside]
[female] back brick red with small
black spots, throat white with brown
spots, belly grey and white mottled.
```

We designed a selective sampling procedure to improve the results of supervised segmentation and labelling of these texts; this is explained in Section 3. In Section 4 a semi-supervised method is introduced that enables harvesting new metadata from a textual database. The interface that

provides access to the data stored in a knowledge base is described in Section 5.

2. Segmentation and Labelling of Field Notes

Our material comes from 80 books containing field notes on collecting reptile and amphibian specimens for the Dutch National Museum of Natural History, *Naturalis*¹. A short field note entry is shown below.

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

In this example, the first unit up to the first comma comprises the name of the animal collected, the next two comma-separated units describe the geographical location where the specimen was found, the following units its physical environment (the so-called biotope), the date and time of collecting, and finally the identification number assigned to this specimen. Due to the commas (which in other entries do not always tend to be consistently placed), these field notes are semi-structured; still, there are at least three aspects that make processing the field notes difficult: (i) the order and length of entering information about the specimens collected or observed on site is not standardised, (ii) a large number of optional information units (e.g., PROVINCE or SUBSPECIES) occur irregularly, and (iii) the texts are written in a mix of Dutch and English, using domain terminology in Latin.

At a later stage, in several phases, some of the entries in the books were typed over to a digital database, as employees of *Naturalis* gradually utilised one or another group of specimens in their research. This *Reptile and Amphibian database* has 37 columns. The above example would fill seven of the possible columns, namely:

```
GENUS Gonatodes
SPECIES humeralis
LOCATION post Tigri, New River
BIOTOPE On tree
COLLECTDATE 2-VII-1968
COLLECTTIME 16.30 h
REGISTRNR RMNH 16314
```

¹<http://www.naturalis.nl>

2.1. Experimental setup

Our first goal is to turn the plain text field note entries into database records automatically. We recast this as a token-based, supervised, sequence labelling task, where each token in an entry needs to be marked as belonging to one of the 37 columns of the existing Reptile and Amphibian database at Naturalis, currently containing 16,870 records. By populating it with data from (non-overlapping) field notes, the database grows about three times its current size, allowing automated access to much more information about the museum’s collection.

After digitising 15,000 hand-written pages, we have 40,749 field note entries at our disposal, containing 40 tokens on average: words, numbers, punctuation marks, as well as symbols indicating illegible phrases, drawings, and a layout marker in the original books.

We experimented with two supervised machine learning algorithms for the joint segmentation and labelling task: conditional random fields, and memory-based tagging. A CRF algorithm defines a conditional probability distribution over label sequences, given a particular observation sequence (i.e., a sequence of tokens), rather than a joint distribution over both label and observation sequences (see (Lafferty et al., 2001) for details). We used the default CRF++ package built by Taku Kudo². MBT is a memory-based tagger generator that classifies sequences based on stored examples and a frequency-thresholded vocabulary (Daelemans et al., 2003). MBT may be fine-tuned using algorithmic parameters of TiMBL 5.2, a memory-based software package³.

The classifiers were trained on 300 entries and tested on 200 held-out entries, both manually annotated as described in (Canisius and Sporleder, 2007). Only the two left and two right context tokens were used as features to classify the focus token. We employed IOB-encoding of spans of labels in both classifiers’ experiments; due to technicalities, MBT was run as a multilabel classifier, while CRF as a binary one.

2.2. Results

Our observation is that MBT outperforms CRF with the results of 0.88 accuracy, 0.82 precision, and 0.86 recall, yielding 0.84 F score (cf. Table 1). The performance of these supervised sequence classifiers shows improvement over both the supervised and unsupervised methods reported by Canisius and Sporleder on the same test set.

Table 1 reports general results as computed over all label types. If we analyse the scores obtained on the individual entities, the classifiers’ tendencies are similar. Using MBT, the highest F score (0.97) is achieved on the classes REGISTRNR and LAND, while the lowest (0.53) on SPECIALREMARKS. CRF obtains the best F score on the LAND label as well (0.91 F), but a much lower result than MBT on REGISTRNR (0.76). CRF’s lowest score (0.30) is also on identifying text belonging to SPECIALREMARKS, but it does this more poorly than MBT.

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

Table 1: Joint segmentation and labelling scores on field note entries using conditional random fields (CRF) and the memory-based tagger (MBT).

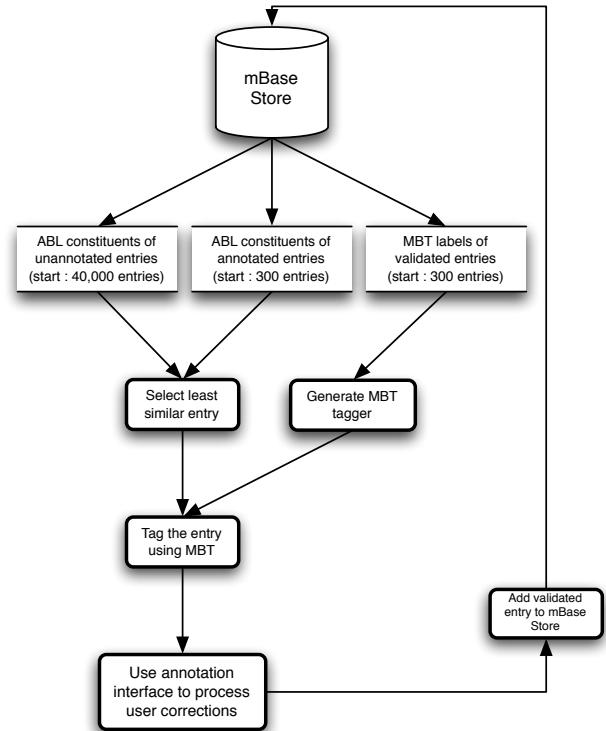


Figure 1: Flowchart of the selective sampling procedure.

3. Postprocessing with Selective Sampling

Our results on segmenting plain text into a database are good, but not perfect, whereas errors should not be allowed in the museum collection database. To correct labelling errors, we designed an interactive annotation tool that lets a human expert correct the automatically labelled output. The tool integrates a selective sampling loop.

Selective sampling is an active learning method (Dagan and Engelson, 1995), where the sampling procedure is usually based on the diversity of classification results from several classifiers, or on confidence measures. In our implementation, in each loop an unannotated sentence is picked for validation that bears the least similarity to the examples already in memory. This sentence is then automatically labelled by MBT, and is shown via the tool’s interface to a domain expert. After corrections, the validated sentence is added to the pool of training instances, from which MBT generates a new tagger. Next, a new sentence is picked for validation, we label it using the newly generated MBT tagger, and show it to the human annotator. The process flow is illustrated in Figure 1.

Figure 2 shows a screenshot of the annotation tool interface.

²available at <http://crfpp.sourceforge.net/>

³MBT and TiMBL are available at <http://ilk.uvt.nl/>

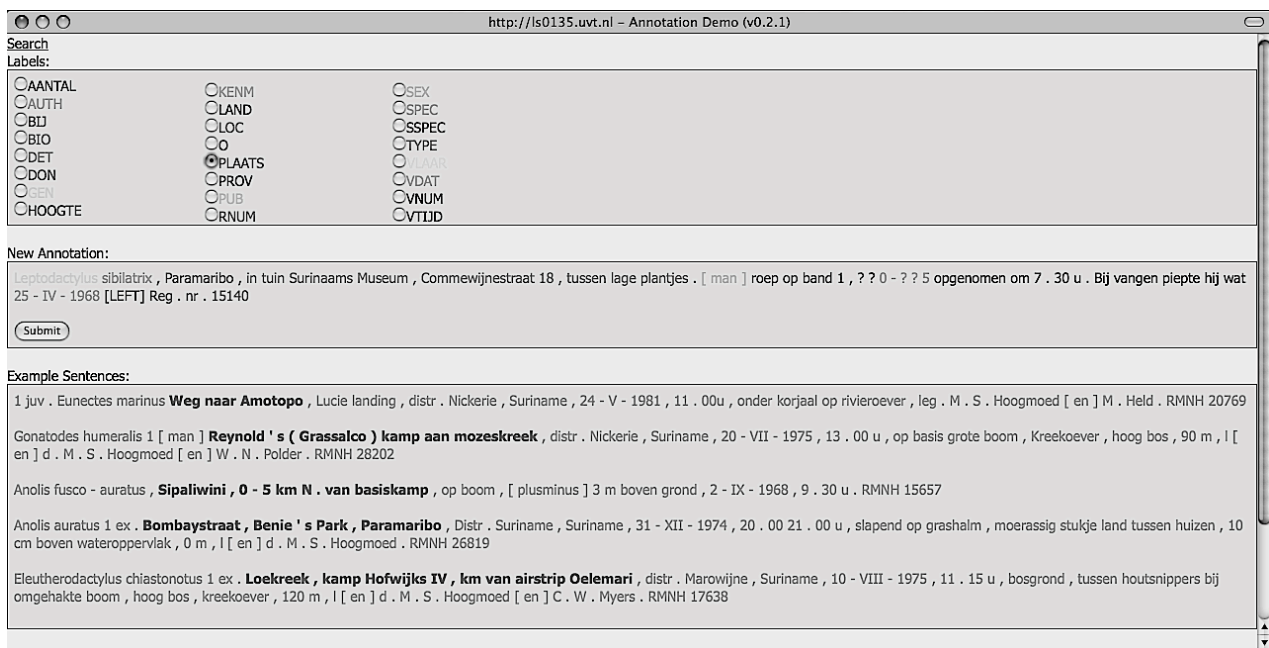


Figure 2: Interface for semi-automatic post-processing of segmentation and labelling errors on field notes. The PLACE label is selected.

The user is shown a field book entry as tagged by MBT. The span of words assigned to various labels (i.e., database columns) are displayed in distinct colours. Hovering above a word with the mouse displays the name of the assigned label.

If correction of the sentence in focus is necessary, the new column's button is to be selected in the upper section of the interface. Then, clicking on incorrectly labelled words assigns these the selected new label. In the bottom section of the screen the user can observe already manually validated sentences (the selected label in boldface) that provide examples to facilitate labelling decisions. When the displayed focus sentence is judged correct, the Submit button is clicked, and the validated sentence is added to the training data.

3.1. The Alignment-Based Learning algorithm

The alignment-based learning algorithm (ABL) is an unsupervised, symbolic, structure bootstrapping system (van Zaanen, 2001). ABL finds the grammar that underlies a corpus of plain text sentences, without using any external sources of information. It aligns the sentences in an input corpus and creates hypotheses where unequal parts of sentences are syntactic constituents, clustered and judged interchangeable in their given context. Constituents are found based on parametric string edit distance metrics, in a procedure that can be likened to bracketing spans of words in each sentence, after comparing it to every other sentence in the corpus. In the current study ABL is used with default settings⁴. We run ABL in a standalone procedure before the selective sampling loop, and store the induced constituents for each sentence.

⁴ABL is available from <http://www.ics.mq.edu.au/~menno/research/software/abl/>

3.2. Sampling

During the segmentation and labelling of field note entries, MBT performs a multi-label classification task, drawing on as little as 300 annotated sentences in its memory. Due to data sparseness, the tagger will experience difficulty in classifying unseen words and constructions. By adding automatically tagged and manually validated sentences to the set of training data, as they gradually become available, the classification model can improve.

The implemented sampling procedure selects most dissimilar sentences from the unannotated sentences, based on longest common substrings. Since the pool of unannotated sentences is in the magnitude of tens of thousand sentences, at the beginning of the sampling procedure many of these score equally dissimilar. To differentiate between these, we search for sentences that do share a minimal amount of syntactic components with the memory examples, in terms of ABL's hypothesis space. Minimally overlapping constituents between the examples in memory and the most dissimilar candidate sentences from the unannotated pool are dynamically identified, using frequency-based ranking.

4. Inducing Metadata

The next goal is to introduce more structure into the knowledge base. Our approach is to create new metadata, based on already existing metadata (i.e., the column names in the database). We extract a secondary layer of metadata from free text columns in a process we call *field expansion*, explained in (Lendvai, 2008).

4.1. Database Field Expansion

Free text fields of a database – such as SPECIALREMARKS or APPEARANCE – contain (fragmented) sentences, as opposed to fields whose contents express a single value, often

in a single word – such as SPECIES. Consider the following example from the 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.

If a researcher is searching for tape recordings of a certain year, he also needs to browse through slide identification numbers and specimen ID numbers, because retrieving numbers can only be done by accessing the entire SPECIALREMARKS field. A query would be more efficient if the various ID numbers of slides, tape recordings, registration numbers, etc. would be separately accessible, which we achieve by adding corresponding labels to the relevant spans of words.

During field expansion, the content (values) of complex fields are assigned new metadata labels. Our method performs this requiring no linguistic analysis of the cell contents, which is attractive for specialised domain texts, on which most NLP tools, trained on general corpora, might be suboptimal to use.

4.2. Candidate Selection

The extraction of field expansion candidates draws on the observation that most (syntactic) heads modified by a value can qualify as metadata. Our procedure therefore spots *empty constituents*, i.e., indications of a possible but unrealised constituent (i.e., the modifier), and stores the context word (i.e., the head) of the empty constituent as a metadata candidate. This approach can be seen as a heuristic model for retrieving head-modifier dependency relations from ABL’s hypothesis space. Here, too, the candidate labels are presented afterwards to a human expert for validation.

4.3. Experiments

The approach was tested on two datasets created from the original Reptile and Amphibian database: SPECRA from the SPECIALREMARKS field, and BIORA from the BIOTOPE field. The full content of each field belonging to these columns is regarded as a sentence. The words in the sentences are tokenised. All occurrences of numbers in SPECRA are masked by the symbol NUM. We then run ABL on each dataset and induce a grammar from these.

Table 2 describes the results of processing the two datasets: in terms of the induced grammar by ABL, the candidate extraction process, and examples of new, accepted metadata. Comparing the amount of tokens in each datasets and the number of proposed candidates, we characterise the magnitude of reduction in time needed for the human expert to browse through the processed columns to validate metadata. In SPECRA, for example, empty constituents are identified in the context of 181 words. From these 29 are accepted. These are integrated as new concepts in the ontology underlying the knowledge base.

4.4. Analysis

To illustrate the semantic range of the output, the bottom section of the table displays a few of the validated 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	bush
	died	creek
	formerly	forest
	length	ground
	loan	pool
	museum	river
	obtained	road
	photo	rock
	slide	swamp
	tank	vegetation
	university	water

Table 2: Field expansion results on two fields from the Reptile and Amphibian database.

candidates from each dataset. Some labels we translated from Dutch, some are the English original, such as ‘loan’ or ‘formerly’ in SPECRA. It is interesting to observe that in the BIORA dataset several candidates are extracted both in their English form (‘forest’, ‘ground’, ‘tree’, ‘road’) and in Dutch (‘bos’, ‘grond’, ‘boom’, ‘weg’). More detailed examples with actual values extracted for certain metadata fields are given in Table 3.

Often, we extracted synonyms (e.g., ‘formerly’ and ‘originally’; ‘purchased’ and ‘obtained’), spelling variants, as well as semantically related word forms, such as both the nominalised and the inflected verb form (e.g., ‘loan’ and ‘loaned’). In the ontology these terms are collapsed into a single concept, but for the markup procedure it is important that several syntactic or language variations of one and the same term are detected. Field expansion is a method to induce an additional layer of metadata, and link the primary layer with the content of the fields through domain concepts of various granularity.

5. Disclosing Data

The goal of the integrated system, *mBase*, is to provide easier and more intuitive access to data from the museum. *mBase* runs on an open source XML management system, *eXist*⁵. It can be accessed through keyword search across the whole knowledge base or via more specific search of individual fields, which includes searching over specific date ranges and grouping by for instance species or expedition. The induced metadata labels can be used in combination with keyword search to filter retrieval results (e.g., only numbers pertaining to tape recordings, or only specimens loaned to certain musea should be presented, etc.). As the data is being continually updated it is also important to keep

⁵<http://exist.sourceforge.net/>

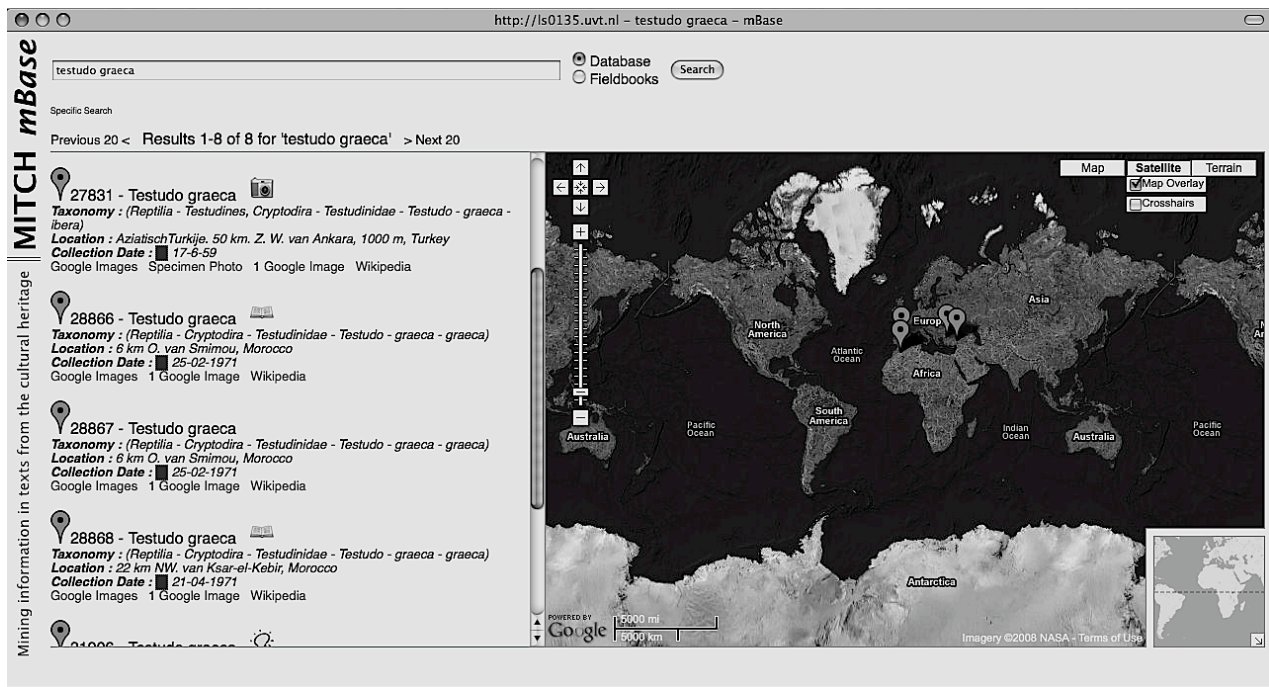


Figure 3: The search and visualisation display of the knowledge base.

Modifier	Head	Modifier
Hebrew	University	
Stanford	University	
Kansas	University	
South Australian	Museum	
British	Museum	
Natural History	Museum	
	forest	of <i>Quercus ilex</i>
	forest	in moss cushions
	forest	on loamy soil
	road	through cultivation
	road	under fallen <i>Cecropia</i> leaves
	road	after heavy rain
	rock	surface of hill under stones
	rock	boulder near road
	rock	in savanna

Table 3: Metadata candidates with assigned values (left or right modification) from the SPECIALREMARKS and BIOTOPE columns.

track of changes; it is thus possible to browse revisions and to search only particular versions of the knowledge base.

The interface supports linking to various other resources such as online geocoding SPECIALREMARKS and photo collections. Visualisation in this way allows for more intuitive browsing of specimens. This may also help highlighting erroneous entries in the dataset: e.g., spotting a specimen collected in a different continent than all other examples might suggest inconsistent data. Figure 3 shows a screenshot of the search and visualisation display. When presenting query results, the system uses icons to signal if a particular record is linked to a fieldbook entry, a photo, or has a potentially erroneous field value in the database.

Going to the record view of the mBase interface (see Figure

4) lets the user inspect all details of a record. It is also used to display the segmented field note associated to the given record (if any), alongside the structured database. The labels appearing in the segmented field notes are linked to their predicted database columns. This information is used to provide a convenient method for entering new database entries by a couple of mouse clicks.

When browsing for and displaying records, the mBase interface alerts the user to anomalous entries marked as such by Timpute, a perl wrapper around TiMBL, which uses a database's own features to spot errors and offer corrections in database cells (cf. (Sporleder et al., 2006)). The confidence of the error detections is displayed together with suggested corrections for the user to review.

6. Concluding remarks

We described the process of converting cultural heritage text into a searchable knowledge base, using machine learning. This includes segmentation of large amounts of texts into database fields, training a classifier on a small set of labelled examples. To postprocess the results, an annotation tool based on selective sampling (drawing on unsupervised grammar induction) is implemented. We explained a database field expansion method to create metadata, the results of which are integrated in the knowledge base. The interface to the knowledge base allows for interactively importing automatically tagged field notes to records, and displays query results using maps and specimen photos. In follow-up work within our project we plan to report on the design of the ontology underlying the knowledge base.

7. Acknowledgements

We thank Caroline Sporleder for sharing the tokenisation script and the annotations, and Svitlana Zinger for the longest common substring implementation.

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

Database Fieldbooks Search

Specific Search

MITCH mBase

Mining information in texts from the cultural heritage

Registration Number	20643		
Class	Amphibia		
Order	Sauria	Deviates from expected value 'Reptilia' (accuracy of ~99%)	
Family	Gymnophthalmidae		
Genus	Tretioscincus		
Species	agilis		
Sub Species	Specimen Image not available		
No. of Specimens	1		
Sex	m		
Storage Method	alcohol		
Special Remarks			
Attribute			
Collector	Hoogmoed, M.S. & Polder, W.N.		
Label Data	Collection Date	14-08-1975	
Country	Collection Number	1975-MSH1633	
Province/State	Country ID	220	
Place	Altitude	650	
Biotope	Coordinates		
Location	Determinator		
Author	Determination Date		
Publication	Recorder	Grouw, H.J. van	
Printed	Record Date & Time	2001-07-16 09:58:22	
Globally Unique ID	Inventory Number	0	
fieldbook Text	Expedition	mshsurfg1975	
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

Figure 4: The record view of the knowledge base.

8. References

- S. Canisius and C. Sporleder. 2007. Bootstrapping information extraction from field books. In *Proc. of EMNLP-CoNLL*.
- W. Daelemans, J. Zavrel, A. van den Bosch, and K. van der Sloot. 2003. MBT: Memory-Based Tagger, version 2.0, Reference guide. Technical report, ILK research group technical report series 03-13, Tilburg.
- I. Dagan and S. Engelson. 1995. Selective sampling in natural language learning. In *Proc. of IJCAI-95 Workshop*.
- J. Lafferty, A. McCallum, and F. Pereira. 2001. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proc. of ICML-01*.
- P. Lendvai. 2008. Alignment-based expansion of textual database fields. In A. Gelbukh, editor, *CICLing 2008. LNCS, vol. 4919*. Springer Berlin / Heidelberg.
- C. Sporleder, M. van Erp, T. Porcelijn, and A. van den Bosch. 2006. Spotting the 'odd-one-out': Data-driven error detection and correction in textual databases. In *Proc. of the EACL 2006 Workshop on Adaptive Text Extraction and Mining (ATEM-06)*.
- M. van Zaanen. 2001. *Bootstrapping Structure into Language: Alignment-Based Learning*. Ph.D. thesis, School of Computing, University of Leeds, UK.