Digital Library & Virtual Museum

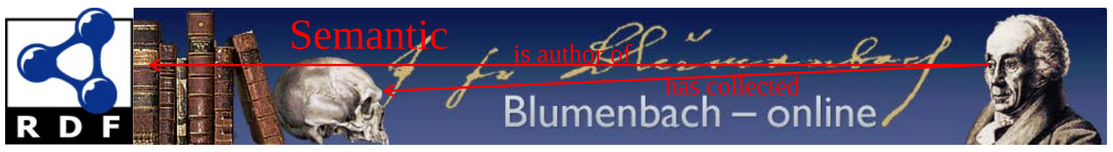
Digital Humanities Research Collaboration – Lower Saxony

Named Entity Recognition

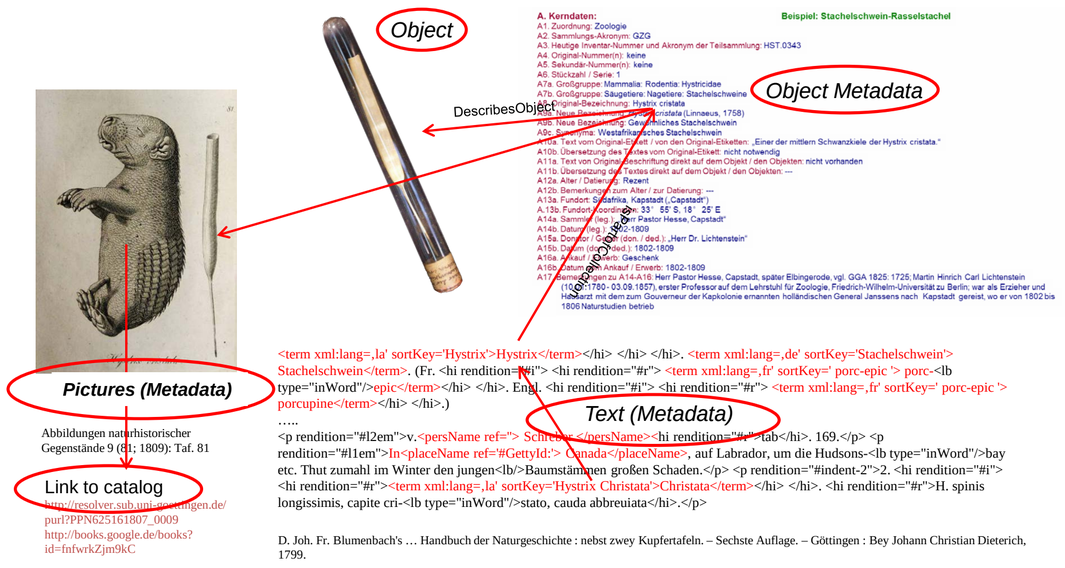
in historical texts from the natural history domain

GCDH
GÖTTINGEN CENTRE FOR DIGITAL HUMANITIES

Academy of Science and Humanities Goettingen



www.blumenbach-online.de



Abbildungen naturhistorischer Gegenstände 9 (81; 1809): Taf. 81

Abstract

In this poster we describe and evaluate a prototype system for recognizing and identifying named entities (persons, places, technical terms, objects and dates) in 18th century German texts written by Johann Fr. Blumenbach (1752 – 1840), professor at Gottingen and one of the fathers of physical anthropology. TEI encoded files are used as input for our system. A particular feature of the texts written by Blumenbach is the highly specialized scientific vocabulary originating from the field of natural history and including vernacular idioms that are out of use nowadays. Applying NER for this kind of documents is not an easy task and calls for a particular mix of list- and rule-based approaches to produce results of high precision and recall.

Problems and solutions

- Standard NER programs (e.g. GATE) have problems with historical and multilingual texts (difficulties for rule based approaches)
- List-based approaches need a lot of time to build the appropriate lists, especially for the recognition of persons
- Search and replace in TEI as opposed to plain text calls for special adaptations of the parser…
  Example: Ha-<lb type="inWord"/>sen
- The hierarchical structure of TEI can cause problems when tagging semantically related strings that do not occur together…
  Example: Hystrix Christata (Crested Porcupine)

genus species

<term xml:lang= 'la' sortKey='Hystrix'><rs type=„Zoologie" ref="3">Hystrix</rs></term></hi> </hi> </hi>. <term xml:lang= 'de' sortKey='Stachelschwein'>Stachelschwein</term>. (Fr. <hi rendition="#i"> <hi rendition="#r"> <term xml:lang=,fr' sortKey=' porc-epic '> porc-<lb type="inWord"/>epic</term></hi> </hi>. Engl. <hi rendition="#i"> <hi rendition="#r"> <term xml:lang=,fr' sortKey='porc-epic '> porcupine</term></hi> </hi>.)
…..
v.<persName xml:lang="de" ref="http://thesaurus.cerl.org/record/cnp01362609"><surname>Schreber</surname></persName><hi rendition="#r">tab</hi>. 169. In<placeName ref='#GettyId: 7005685'> Canada</placeName>, auf Labrador, um die Hudsons-<lb type="inWord"/>bay etc. Thut zumahl im Winter den jungen<lb/>Baumstämmen großen Schaden. 2. <hi rendition="#i"> <hi rendition="#r"><term xml:lang= 'la' sortKey='Hystrix Christata'>Christata</term></hi> </hi>. <hi rendition="#r">H. spinis longissimis, capite cri-<lb type="inWord"/>stato, cauda abbreuiata</hi>.

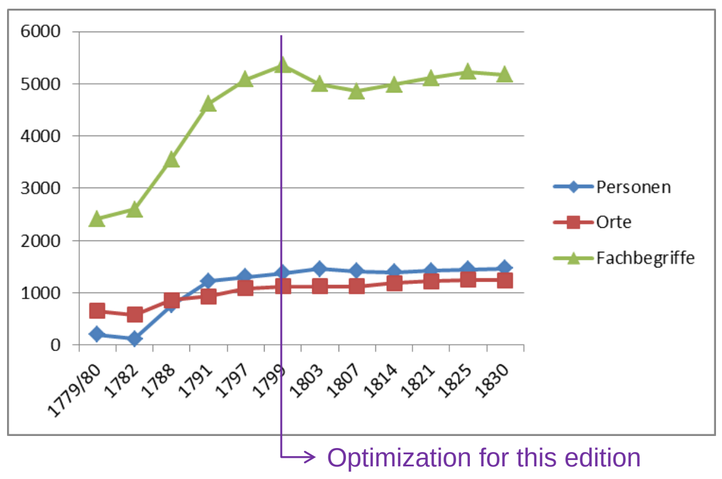
Results for 12 editions of the same book

Absolute Numbers of tagged strings in the different editions of Blumenbachs‘ natural history manual from 1779 - 1830

e.g.<term xml:lang= 'la' sortKey='Hystrix Christata'>

e.g.<placeName ref='#GettyId: 7005685'> Canada

e.g.<persName xml:lang="de" ref="http://thesaurus.cerl.org/record/cnp01362609">



Lessons learned so far…

- Usage of existing indexes (e.g. Lémery) and thematically similar word lists can facilitate the recognition and increase the recall
- List enrichment via authority files (CERL, Getty, GND) allows for person identification and works faster than an integration *on the fly*
- Specially adapted tools for correction and maintenance of the lists facilitate and speed up the work
- Combination of list- and rule-based NER seems most promising for this particular corpus of historical texts on natural history

Contact:
Dr. Jörg Wettlaufer
Akademie der Wissenschaften zu Göttingen (ADWG)
Göttingen Centre for Digital Humanities (GCDH)
Tel.: +49 (0)551 39 20477
jwettla@gwdg.de
www.gcdh.de | www.digihum.de

Sree Ganesh Thotempudi
Digital Humanities Research Collaboration – Lower Saxony
Göttingen Centre for Digital Humanities (GCDH)
Tel.: +49 (0)551 39 20479
sree-ganesh.thotempudi@gcdh.de
www.gcdh.de

Literature

- **Sören Auer, Jens Lehmann, and Axel-Cyrille Ngonga Ngomo: Introduction to Linked Data and Its Lifecycle on the Web, in: A. Polleres et al. (Eds.): Reasoning Web, Springer 2011, p. 1-75.**
- **Nadeau, D.: Semi-Supervised Named Entity Recognition: Learning to Recognize 100 Entity Types with Little Supervision, Diss. Ottawa 2007.**

Gefördert von
Niedersächsisches Ministerium für Wissenschaft und Kultur
VolkswagenStiftung
Digital Humanities Forschungsverbund - DHFV