

# Named Entity Recognition in historical texts from the natural history domain



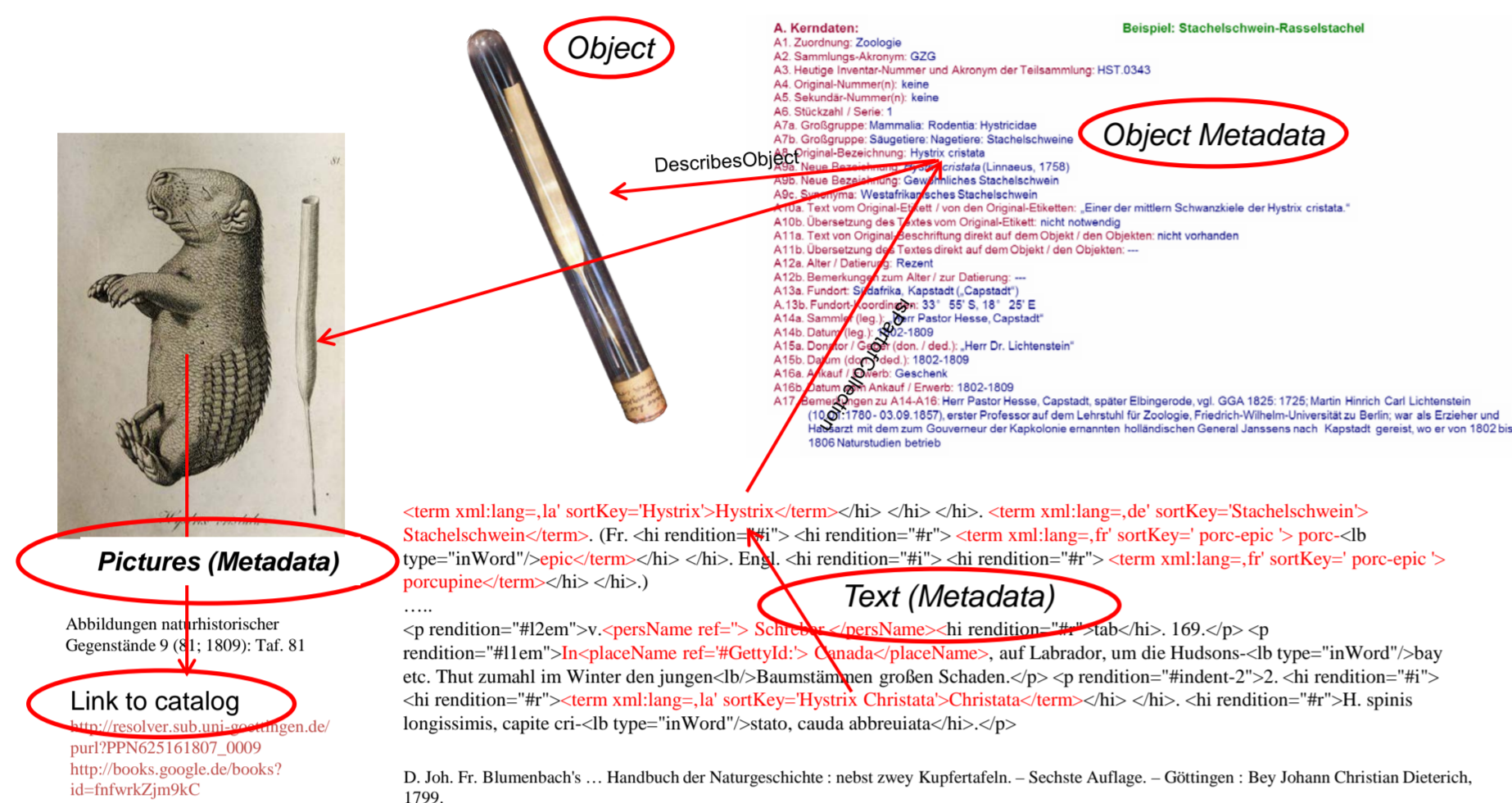
Academy of Science and Humanities  
Goettingen



www.blumenbach-online.de

## 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 Göttingen 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.



**Object**

**Object Metadata**

**Pictures (Metadata)**

**Text (Metadata)**

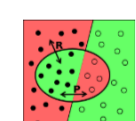
Abbildungen naturhistorischer Gegenstände 9 (81; 1809): Taf. 81

Link to catalog

D. Joh. Fr. Blumenbach's ... Handbuch der Naturgeschichte: nebst zwey Kupferplatten. – Sechste Auflage. – Göttingen: Bey Johann Christian Dieterich, 1799.

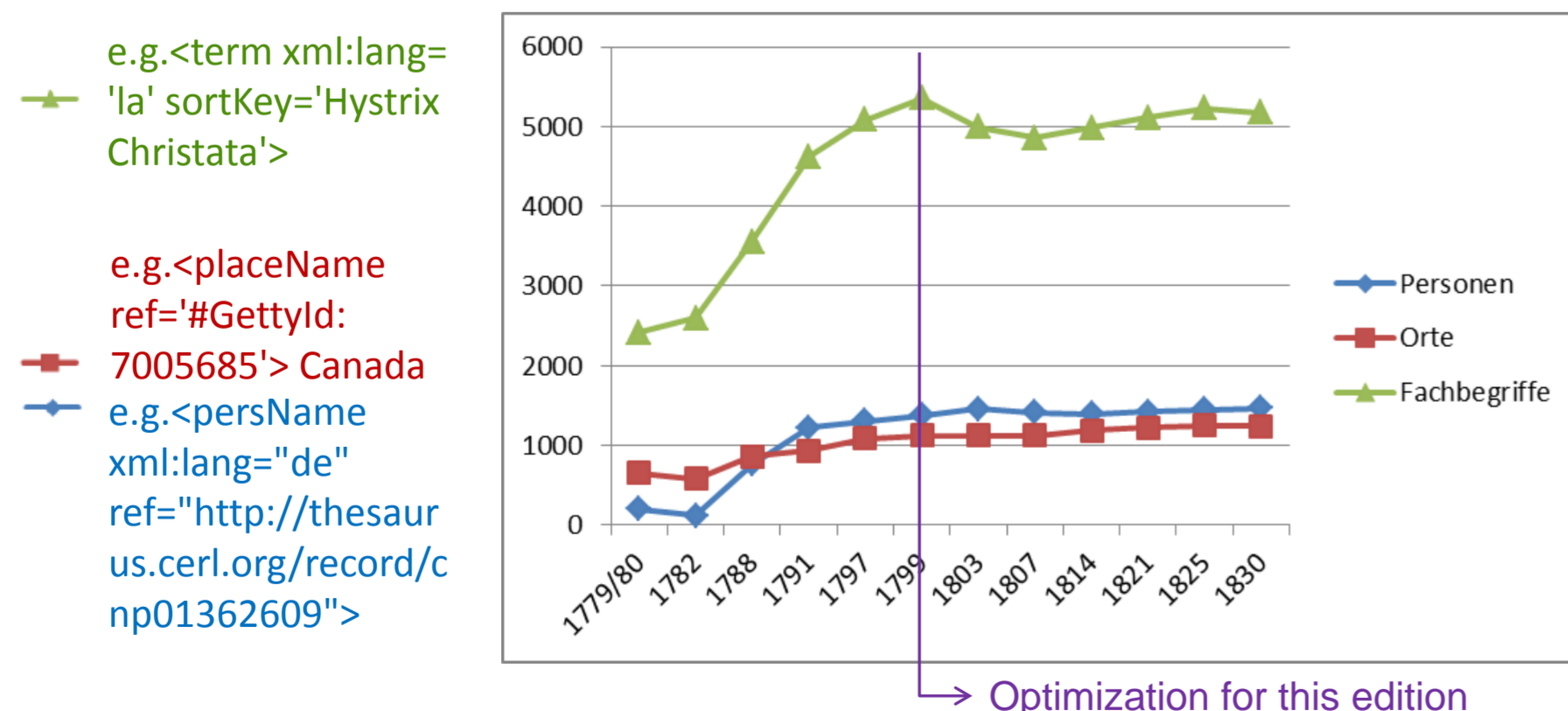
## Problems and solutions

- 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)



## Results for 12 editions of the same book

Absolute Numbers of tagged strings in the different editions of Blumenbach's natural history manual from 1779 - 1830



## 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.