Studying the History of Pre-Modern Zoology with Linked Data and Vocabularies

Molka Tounsi[1], Catherine Faron Zucker[1], Arnaud Zucker[1], Serena Villata[2], and Elena Cabrio[2]

[1] Univ. Nice Sophia Antipolis, France
tounsi.molka@etu.unice.fr, faron@unice.fr, zucker@unice.fr
[2] Inria Sophia Antipolis Méditerranée, France
serena.villata@inria.fr, elena.cabrio@inria.fr

Abstract. In this paper we first present the international multidisciplinary research network Zoomathia, which aims the study of the transmission of zoological knowledge from Antiquity to Middle Ages through varied resources, and considers especially textual information, including compilation literature such as encyclopaedias. We then present a preliminary work in the context of Zoomathia consisting in (i) extracting pertinent knowledge from mediaeval texts using Natural Language Processing (NLP) methods, (ii) semantically enriching semi-structured zoological data and publishing it as an RDF dataset and its vocabulary, linked to other relevant Linked Data sources, and (iii) reasoning on this linked RDF data to help epistemologists, historians and philologists in their analysis of these ancient texts.

Keywords: History of Zoology, Semantic Analysis of Mediaeval compilations, Linked Data and Vocabularies

1 Introduction

Scholars concerned with cultural issues in Antiquity or Middle Ages have to deal with a huge documentation. The literary material is a significant part of this material, but the commonly used technology supporting these researches is to date far from satisfactory. In spite of pioneering undertakings in digitization since the 70's, historians and philologists still have access to few tools to operate on texts, mostly limited to lexical searches. Therefore they stand in need for more intelligent tools, in order to overcome this word-dependency, to access the semantics of texts and to achieve more elaborated investigations.

The Semantic Web has an increasing role to play in this process of providing new methodological implements in cultural studies. During the last decade, several works addressed the semantic annotation and search in Cultural Heritage collections and Digital Library systems. They focus on producing Cultural Heritage RDF datasets [1, 4], aligning these data and their vocabularies on the Linked Data cloud [2, 7], and exploring and searching among heterogenous semantic data stores [5, 8, 3, 6].

The international research network Zoomathia[3] has been set up to address this challenge in the area of History of Science. It aims to develop interconnected researches on History of Zoology in pre-modern times and to raise collaborative work involving philologists, historians, naturalists and researchers in Knowledge Engineering and Semantic Web. In this context, we conducted a preliminary work, presented in this paper, on the fourth book of the late mediaeval encyclopaedia *Hortus Sanitatis* (15th century), which compiles ancient texts on fishes. Each chapter of this book is dedicated to one fish, with possible references to other fishes. In this work we aim at *(i)* automating information extraction from these texts, such as zoonyms, zoological sub-discipline (ethology, anatomy, medicinal properties, etc.); *(ii)* building an RDF dataset and its vocabulary representing the extracted knowledge, and link them to the Linked Data; and finally, at *(iii)* reasoning on this linked data to produce new expert knowledge. We build upon the results of two previous French research projects on structuring mediaeval encyclopaedias in XML according to the TEI model and manualy annotating author sources (SourceEncyMe project[4]) and zoonyms (Ichtya project[5]).

The paper is organized as follows: Section 2 presents the general aim of Zoomathia. Section 3 presents our work on knowledge extraction from the mediaeval encyclopaedia *Hortus Sanitatis*, while Section 4 describes the publication of a linked RDF dataset and its vocabularies. Section 4.3 presents preliminary work on the exploitation of these data to support the study of the history of pre-modern zoology, and Section 5 concludes the paper.

2 The Zoomathia Research Network

Zoomathia primarily focuses on the transmission of zoological knowledge from Antiquity to Middle Ages. Manual search and computing on ancient and mediaeval texts enable to address the quantitative dimension of data but fail to answer the epistemological demands, which concern the scientific relevancy and the diachronic features of the documentation. A large range of investigations on specific topics is inaccessible through simple lexical queries and requires a rich, scientific and semantic annotation. When investigating, for example, on ethological issues (such as animal breeding, intraspecific communication or technical skills) or on pharmaceutical properties of animal products, we have to face a scattered documentation and a changing terminology hampering a direct access to and a synthetic grasp of the topics studied. An automatized and semantic-based process will help to link and cluster together the related data, compare evidences in a diachronic approach and to figure out the major trends of the cultural representations of animal life and behaviour.

[3] http://www.cepam.cnrs.fr/zoomathia/

[4] http://atelier-vincent-de-beauvais.irht.cnrs.fr/encyclopedisme-medieval/programme-sourcencyme-corpus-et-sources-des-encyclopedies-medievales

[5] http://www.unicaen.fr/recherche/mrsh/document_numerique/projets/ichtya

In this network, we aim at both *(i)* identifying a corpus of zoology-related historical data, in order to progressively encompass the whole known documentation, and *(ii)* producing a common thesaurus operating on heterogeneous resources (iconographic, archaeological and literary). This thesaurus should enable to represent different kinds of knowledge: zoonyms; historical period; geographical area; literary genre; economical context; zoological sub-discipline (ethology, anatomy, physiology, psychology, animal breeding, etc.). The aim is to synthesize the available cultural data on zoological matters and to crosscheck them with a synchronic perspective. This would enable to reach the crucial concern, i.e. to precisely assess the transmission of zoological knowledge along the period and the evolution of the human-animal relations. Finally, this thesaurus should be published on the Linked Data and linked to modern reference sources (biological and ecological) to appraise the relevance of the historical documentation.

3 Knowledge Extraction from Historians and Texts

3.1 Interviews of Historians

We conducted several interviews with three Historians participating in Zoomathia to explicit a list of major knowledge elements which would be useful in the study of the transmission of ancient zoological knowledge in mediaeval texts. Among them, let us cite the presence (or absence) of zoonyms in the corpus texts, variant names or name alternatives given to an animal (polyonymy), the relative volume of textual records devoted to a given zoonym, references to a zoonym and frequency of occurrences related to it out of their dedicated chapter, geographical location of the described animals, numerical data in the text (size, longevity, fertility, etc.) and other animal properties related to zoological sub-disciplines (ethology, anatomy, physiology, psychology, animal breeding, etc.).

3.2 Extraction of Zoonyms and Animal Properties from Texts

We processed two versions of book 4 of *Hortus Sanitatis*, the original Latin text and its translation in French. We used the XML structured version of these texts, identifying the 106 chapters of the book, divided in paragraphs, themselves including citations. We used TreeTagger to parse Latin and French texts and determine the lemmas and part of speech (PoS) of each word in the text. We searched for the resources available to support the knowledge extraction process. A lexicon of fish names in French and in Latin has been provided by the Ichtya project and we — Knowledge Engineers and Historians — collaboratively built a thesaurus of zoological sub-disciplines and concepts involved in the descriptions relative to these sub-disciplines. Then we defined two sets of syntactic rules for French and Latin to recognize zoonyms from the lexicon of fish names among the lemmas identified in the texts. For instance one of the rules to recognize that a Latin text deals with longevity is the occurence of the verb *vivere* followed by a numeric value followed by the noun *annis* (ablative plural of *annus*).

We conducted a similar processing of the same two texts to extract zoological sub-disciplines and animal properties. We defined two sets of syntactic rules to extract this information from the Latin and French text (39 rules for French and 10 rules for Latin). For instance the Latin verbs *curare* (heal) or *sanare* (cure) with an animal name as subject are used to identify the therapeutic topic; the verbs *comedere* or *pascere* or *deglutire* (eat) are used to identify the diet topic.

Evaluation The analysis of the results of the automatic annotation process was conducted by knowledge engineers and validated by philologists involved in the manual annotation. For the evaluation of the extraction of zoonyms we considered chapters 1 to 53 of book 4 of *Hortus Sanitatis*. We compared the results of the automatic annotation with those of the manual annotation of zoonyms conducted within the past Ichtya project. F-measure equals to 0.93 for both the annotation of the Latin text and the French text. Most missing annotations are due to the fact that the parsing tool is unable to deduce the exact lemma of some words, especially for Latin words. Among 65 missing annotations, 51 (rare) fish names were not annotated because TreeTagger does not recognize them (e.g., *loligo*). Other missing annotations concern composed names and are due to a mismatch between the complete fish name in the reference lexicon and the short name used in the text to be annotated (e.g. *locusta* instead of *locusta marina*). Conversely, most annotation errors are due to ambiguities between marine animal names and terrestrial animals. For instance, lemma *lupus* (wolf) is present in the provided lexicon of fish names (*wolffish*) and there are some comparisons in the text with the (terrestrial) wolf[6].

For the evaluation of the automatic extraction of animal properties, we manually annotated the 25 first chapters of *Hortus Sanitatis* to use it as a reference version. F-measure is above 0.7 for both the annotation of the Latin text and the French text. Most wrong annotations are related to anatomy. These annotations are due to a confusion between human and animal anatomical parts appearing in the text, when the text deals with the therapeutic power of some animal on a human organ. For instance, the detection of lemma *dentes* (tooth) in the text leads to the annotation of the text with the anatomy topic, whereas, in some cases, the text describes a therapeutic effect of the animal on (human) teeth[7].

4 From Unstructured Data to Semantic Data

The extracted knowledge has first been used to enrich the available XML annotation of *Hortus Sanitatis*. Then we translated the whole XML annotation (text structure, source authors, zoonyms and animal properties) into an RDF dataset and vocabularies and exploited it with SPARQL queries.

[6] "And although this is the case for all fishes, it is however more obvious in him (*wolffish*), as it is also for the wolf and the dog among the beasts"

[7] "[Human] teeth are cleaned using conch shell ash."

4.1 RDF Dataset

An RDF dataset describing *Hortus Sanitatis* has been automatically generated by writing an XSL stylesheet to be applied to its XML annotation. Listing 1.1 presents an extract of it describing quotation 4 of paragraph 3 of chapter 20. It is a citation of Aristotle, refering to the *crocodile* zoonym and addressing the *therapeutics* and *anatomy* topics.

```
<http://zoomathia.unice.fr/HortusSanitatis/FR.hs.4.25.3/cit4>
  a tei:Citation;
  tei:hasHead "FR.hs.4.25.3.cit4";
  tei:hasBibliography [ a tei:Bibliography;
    tei:hasAuthor <http://zoomathia.unice.fr/auteurs/Aristote>;
    tei:hasReference
      <http://zoomathia.unice.fr/oeuvres/612_a_21-25N_MS>. ];
  tei:hasCitationText "...";
  zoo:hasZoonym <http://zoomathia.unice.fr/Crocodile>;
  dcterms:subject
    <http://zoomathia.unice.fr/subject/therapeutique>,
    <http://zoomathia.unice.fr/subject/anatomie>. ] ].
```

Listing 1.1. RDF annotation of an Aristotle's citation on crocodiles

4.2 Vocabulary

Based on the lexicon initially provided by Historians involved in the Ichtya project, we built a SKOS thesaurus for zoonyms and we aligned it with both the cross-domain DBpedia ontology and the Agrovoc thesaurus specialized for Food and Agriculture[8]. In a near future we intend to align it with the TAXREF taxonomy specialized in Conservation Biology and integrating Archaeozoological data[9]. Listing 1.2 presents an extract of the thesaurus describing taxon *Garfish*.

```
<http://zoomathia.org/Orphie> a skos:Concept ;
  skos:prefLabel "orphie"@fr ;
  skos:closeMatch <http://fr.dbpedia.org/resource/Orphie> ;
  skos:closeMatch <http://dbpedia.org/resource/Garfish> ;
  skos:closeMatch <http://aims.fao.org/aos/agrovoc/c_5102> ;
  skos:altLabel "gwich" .
```

Listing 1.2. Extract of the Zoomathia thesaurus of zoonyms

We built an RDFS ontology of zoology-related sub-disciplines and animal properties, based on the results of interviews with Historians and the properties extracted from texts. This is a preliminary modelisation which has to be further developed.

[8] http://aims.fao.org/vest-registry/vocabularies/agrovoc-multilingual-agricultural-thesaurus

[9] http://inpn.mnhn.fr/programme/referentiel-taxonomique-taxref?lg=en

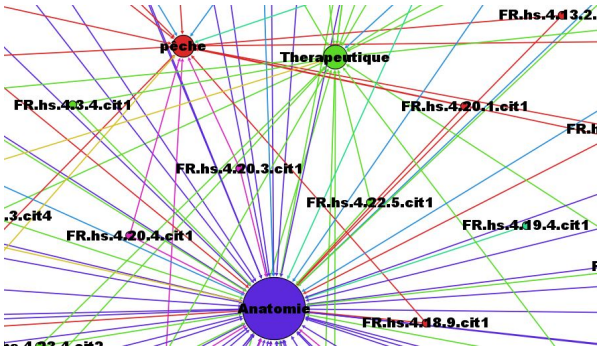

Fig. 1. Relative importance of zoological topics in *Hortus Sanitatis*

4.3 Reasoning on Historical Zoological RDF Data

In order to exploit the extracted RDF knowledge base, we built a set of SPARQL queries enabling to answer questions such as "What are the zoonyms studied in this text?", "What are the topics covered in this text?", "Where can we find these topics?","What are the zoonym properties (in which chapter or paragraph or citation)?". Let us note that it is the semantics captured in the constructed vocabularies which make it possible to answer these queries: multiple labels associated with a taxon in the thesaurus of zoonyms, hierarchy of zoology-related sub-disciplines, denoted by various terms.

We went a step further in the exploitation of the RDF dataset by writing SPARQL queries of the CONSTRUCT form to construct new RDF graphs capturing synthetic knowledge. When graphically visualized, they support the analytical reasoning of historians on texts. For instance, Figure 1 presents the RDF graph capturing the relative importance of zoology-related sub-disciplines in the *Hortus Sanitatis* and their location in it. At a glance, it shows that anatomy occupies a predominant place in this text, far ahead of therapeutics and fishing.

5 Conclusion and Future Work

We presented a preliminary work conducted in the context of the Zoomathia network, on the zoological mediaeval encyclopaedia *Hortus Sanitatis*. This work combines NLP techniques to extract knowledge from texts, and knowledge engineering and semantic web methods to build a linked RDF dataset of zoological annotations of this scientific text. It exploits this dataset to support the analysis of the Ancient zoological knowledge compiled in the encyclopaedia.

The next step will be to apply the presented process on a classical Latin book on fishes (Pliny, *Historia Naturalis*, book 9, 1st century AD), which is a major, though indirect, source of the *Hortus Sanitatis*, to deal with the historical

perspective of zoology, and end up with comparing the data of the two selected works, to appraise the density of the transmission and the evolution of the zoological knowledge on an epistemological point of view. We intend to systematically compare the two texts, with the aim of evaluating the loss, distortion or enrichment of information, and comparing the relative importance in the books of the different zoological perspectives (anatomical, ethological, geographical, etc.) and of the different animal species.

Acknowledgments. Zoomathia is an International Research Group (GDRI) supported by the French National Scientific Research Center (CNRS).

References

1. V. de Boer, J. Wielemaker, J. van Gent, M. Hildebrand, A. Isaac, J. van Ossenbruggen, and G. Schreiber. Supporting Linked Data Production for Cultural Heritage Institutes: The Amsterdam Museum Case Study. In *9th Extended Semantic Web Conference, ESWC 2012, Heraklion, Crete, Greece*, 2012.
2. V. de Boer, J. Wielemaker, J. van Gent, M. Oosterbroek, M. Hildebrand, A. Isaac, J. van Ossenbruggen, and G. Schreiber. Amsterdam Museum Linked Open Data. *Semantic Web*, 4(3), 2013.
3. C. Dijkshoorn, L. Aroyo, G. Schreiber, J. Wielemaker, and L. Jongma. Using Linked Data to Diversify Search Results: a Case Study in Cultural Heritage. In *19th International Conference on Knowledge Engineering and Knowledge Management,EKAW 2014, Linköping, Sweden*, 2014.
4. T. Elliott and S. Gillies. Digital geography and classics. *Digital Humanities Quarterly*, 3(1), 2009.
5. M. Hildebrand. Interactive Exploration of Heterogeneous Cultural Heritage Collections. In *7th International Semantic Web Conference, ISWC 2008, Karlsruhe, Germany*, 2008.
6. L. Isaksen, R. Simon, E. T. E. Barker, and P. de Soto Cañamares. Pelagios and the emerging graph of ancient world data. In *ACM Web Science Conference, WebSci '14, Bloomington, IN, USA*, pages 197–201. ACM, 2014.
7. M. Jackson, M. Antonioletti, A. C. Hume, T. Blanke, G. Bodard, M. Hedges, and S. Rajbhandari. Building bridges between islands of data - an investigation into distributed data management in the humanities. In *Fifth International Conference on e-Science, e-Science 2009, Oxford, UK*, pages 33–39. IEEE Computer Society, 2009.
8. G. Schreiber, A. K. Amin, L. Aroyo, M. van Assem, V. de Boer, L. Hardman, M. Hildebrand, B. Omelayenko, J. van Ossenbruggen, A. Tordai, J. Wielemaker, and B. J. Wielinga. Semantic Annotation and Search of Cultural-Heritage Collections: The MultimediaN E-Culture Demonstrator. *J. Web Sem.*, 6(4), 2008.