Invited Talks

Papers

Preface

These Proceedings contain the papers accepted for publication and presentation at the first International Workshop on Semantic Web for Scientific Heritage (SW4SH) held in conjunction with the 12th ESWC 2015 Conference on June 1, in Portoroz, Slovenia. This workshop aimed at providing a leading international and interdisciplinary forum to disseminate the latest research in the field of Semantic Web for the study of pre-modern scientific texts and of the history of ideas and their transmission.

The program committee members are all involved in this interdisciplinary synergy. They have accepted nine papers (ratio 75%) and invited two keynote speakers. The four workshop organizers belong to the Zoomathia[1] international research network funded by the French National Scientific Research Center (CNRS). This network gathers French, Italian, German and English researchers and aims to study the formation and transmission of ancient zoological knowledge over a long period, with an historical, literary and epistemological approach, and create open knowledge sources on classical zoology to be published on the Web of Linked Data. This workshop was also planned as an opportunity to present the activity of the network, to enlarge it with interested participants of the workshop, and to benefit from the results of related research projects.

This encounter takes place within the general context of Digital Humanities, a research area at the intersection of Humanities and Computer Science which is gaining an ever-increasing momentum and where the Linked Open Data is playing an increasingly prominent role. The opportunity of the workshop was to provide a forum for discussion about the methodological approaches to the specificity of annotating “scientific” texts (in a wide meaning, including disciplines such as history, architecture, or rhetoric), and to support a collaborative reflection on possible guidelines or specific models for building historical ontologies. The iconographic data are also relevant in history of science and arise similar problematics; they offer suggestive insights for a global methodology for diverse media.

The opportunity for a fruitful encounter of knowledge engineers with computer-savvy historians and classicists has come. Since the mid-1970s, classicists and historians have developed textual databases, intending mostly to gather and explore large amounts of primary source materials. For a long time, they mainly focused on text digitization and markup. They only recently decided to try to explore the possibility of transferring some analytical processes they previously thought incompatible with automation to knowledge engineering systems, thus taking advantage of the growing set of tools and techniques based on the languages and standards of the semantic Web, such as linked data, ontologies, and automated reasoning. On the other hand, Semantic Web researchers are willing to take up more ambitious challenges than those arising in the native context of the Web in terms of anthropological complexity, addressing meta-semantic problems of flexible, pluralist or evolutionary ontologies, sources heterogeneity, hermeneutic and rhetoric dimensions.

A key goal of the workshop, focusing on research issues related to pre-modern scientific texts, was to emphasize, through precise projects and up-to-date investigation in Digital Humanities, the benefit of a multidisciplinary research to create interoperable semantic data and reason on them. One of the main interests of the very topic of pre-modern historical data management lies in *historical semantics*, and the opportunity to jointly consider how to identify and express lexical, theoretical and material evolutions. Dealing with historical texts, a major problem is indeed to handle the discrepancy of the historical terminology compared to the modern one, and, in the case of massive, diachronic data, to take into account the contextual and theoretical meaning of words and sentences and their semantics.

Three papers are interconnected to the ZOOMATHIA project. They develop three problematics: extracting knowledge from literary data, linking historical data with Web available information, and assessing theoretical conflicts in the zoological tradition. [1] addresses the problem of extracting zoological knowledge in a text using

[1] http://www.cepam.cnrs.fr/spip.php?rubrique229.

Natural Language Processing (NLP) methods and to publish it as an RDF dataset. [2] focuses on the linking of historical documentation to a reference taxonomical database (TAXREF) and presents a SKOS thesaurus enabling multi-disciplinary studies and approaches. [3] intends to give a general outline of a method combining argumentation theory, Semantic Web languages and techniques to formalize theoretical controversies in scientific texts in an argumentation framework.

Two papers also deal with Medieval knowledge. [4] deals with historical semantics and the epistemological problems arisen by the online *Dictionary of Medieval Scientific French*, and considers how Semantic Web can help to represent and save a semantic complexity and evolution. [5], related to the BIBLISSIMA (*Bibliotheca Bibliothecarum*) project, presents a prototype using open source solutions developed to index and allow complex searches on iconographic databases.

[6], related to the BIBLIMOS project, addresses the problem of digital processing of ancient Arabic manuscripts and present a semantic virtual infrastructure to operate on distributed and heterogeneous sources of digitized manuscripts. [7] reflects on the archaeological knowledge modeling and the methodological issues involved in the description of artefacts, suggesting design perspectives for computer models and tools to address human semiotics.

[8] presents the HPST interdisciplinary project on history and philosophy of science and technology and focuses on the epistemological issues raised by the development of new tools based on Semantic Web and used in historical research. Paper [9] addresses the problem of automatically disambiguating authors' mentions in a corpus of French literary criticism and propose a method based on named-entity linking.

Arnaud Zucker
Isabelle Draelants
Catherine Faron Zucker
Alexandre Monnin

Invited Talk

Lovejoy's Dream – Or How to do History of Ideas Computationally in a Methodologically Sound Way

Arianna Betti
University of Amsterdam

Abstract :

History of ideas is a discipline largely founded by Arthur O. Lovejoy in the early twentieth century. Lovejoy characterized it as being concerned with *unit-ideas*, entities that retain their meaning through time and can therefore be traced in various contexts, that is, periods, intellectual settings, and disciplinary fields (Lovejoy 1936: 3-7, 15).

Now suppose a historian of ideas wants to trace an idea such as a *truth* through two-thousand years. According to WorldCat, 17,843,437 books have been published only between 1700 and 1900. How is a historian of ideas even supposed to think that such quantities of text can be studied with the historian's traditional method of investigation, namely close reading on one's own?

One might think that with today's digital means, such a study is finally possible. However, things are not that simple. First, we are far from the universal corpus we should be able to rely on for such an enterprise. Second, even with a universal corpus at our disposal, generic, simple and shallow bottom-up analyses of lots of diverse, 'long' and complex data is going to fail. For in a field such as this, feeding a computer masses of diverse and complex texts can only yield masses of unorganized details. Third, even if it were possible to make sense of such results computationally, there are fundamental problems with the very method of the history of ideas. The notion of *one idea* traceable through centuries of thought is illusory: for ideas cannot be studied in isolation from their context, and their meaning is in constant flux (Skinner 2002).

In this talk I elucidate a specific proposal for concept modelling to solve in particular the last two problems. I also show in what way my proposal is able to give a theoretical foundation to answering the need for dynamic ontologies, so that a computational turn can be effectively taken to history of ideas and related disciplines.

(based on joint work with Hein van den Berg)

Invited Talk

Linked Data for Digital History

Victor de Boer
Vrije Universiteit, Amsterdam

Abstract :

With the increasing popularity of digital humanities, researchers seek more (international and cross-domain) collaboration. Integrating humanities datasets becomes more important to these researchers. This is very much prevalent in historical research and to further the digital history agenda, sharing data and knowledge is key. Semantic Web technologies present good opportunities to support historians in this aspect.

I will argue that in order for digital history to become a lasting multidisciplinary field of research rather than a popular buzzword, we need to stop building specific tools and visualizations based on requirements provided by historians. Rather, we should develop generic methodologies for modelling, linking and providing access to historical data. Properly represented and accessible data becomes more valuable over time, whereas specific analysis tools are hard to develop, combine maintain.

This means that historians will need to be able to 1) browse heterogeneous datasets in a convenient way to get an intuition of the character and anomalies of the (linked) data; 2) perform arbitrary queries to retrieve results relevant to their research questions; 3) verify the veracity of query results, by following provenance links to original material and 4) analyze the data with their tool of preference.

For historical researchers and computer scientist to successfully co-develop these methodologies computer scientists will need to understand the historical science methods and historians will need to learn how to perform these queries. Under those conditions, the use of Semantic Web technologies presents a real next step for historical research.

I will discuss a number of recent collaborations between computer scientists and historians to investigate their value for digital history. I will present projects that tackle different aspects of the historical science agenda, including the Dutch Ships and Sailors and BiographyNet projects on representing and linking heterogeneous datasets and the DIVE project on browsing linked media collections.

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

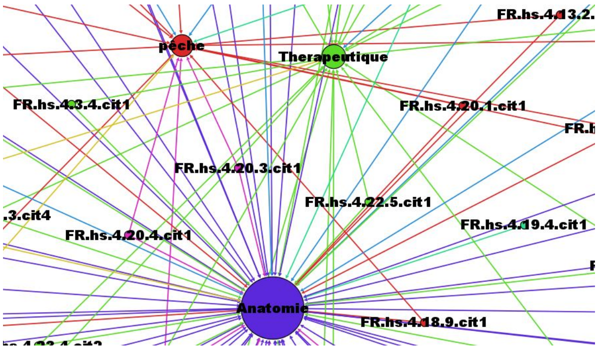


Fig. 1. Relative importance of zoological topics in *Hortus Sanitatis*

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.

Towards a Shared Reference Thesaurus for Studies on History of Zoology, Archaeozoology and Conservation Biology

Cécile Callou[1], Franck Michel[2], Catherine Faron-Zucker[2], Chloé Martin[1], and Johan Montagnat[2]

[1] Archéozoologie et archéobotanique (UMR 7209), BBEES (UMS 3468), Sorbonne Universités, Muséum National d'Histoire Naturelle, CNRS, France
[2] Univ. Nice Sophia Antipolis, CNRS, I3S (UMR 7271), France

Abstract. This paper describes an ongoing work on the construction of a SKOS thesaurus to support multi-disciplinary studies on the transmission of zoological knowledge throughout historical periods, combining the analysis of ancient literature, iconographic and archaeozoological resources. We first describe the I2AF, a national archaeozoological and archaeobotanical inventory database integrating data from archaeological excavation reports. Then we describe the TAXREF taxonomical reference designed to support studies in Conservation Biology, that was enriched with bioarchaeological taxa from I2AF. Finally we describe the TAXREF-based SKOS thesaurus under construction and its intended use within the Zoomathia research network.

Keywords: I2AF, TAXREF, SKOS, History of Zoology

1 Introduction

Animal bones and plant remains from archaeological excavations are a rich and original source of information on the history of biodiversity and its interaction with human societies. When compared with the knowledge about diversity and current locations of human populations, these remains help to figure out the scenarios of past extinction, biological invasions and anthropic impact. This is particularly true during the Holocene, when the influence of human activities overrode that of climatic factors. Therefore, gathering archaeozoological and archaeobotanical data in a sustainable bioarchaeological database, publicly available, represents a major challenge for Natural Sciences and Conservation Biology. *Archaeozoological and Archaeobotanical Inventories of France* database [1] (I2AF) aims to address this challenge.

Historians address a related challenge. Identifying the reported species in ancient literary and iconographic resources, and assessing the documentation is a momentous issue of the History of Zoology. An increasing amount of primary material (such as textual or iconographic resources) is encoded in domain-specific digital formats. For instance, the SourcEncyMe[3] and Ichtya[4] projects aim to encode mediaeval encyclopedias in the XML-TEI standard[5] while adding manual annotations with regards to mediaeval compilers, author sources and taxa. These works succeed in making material about mediaeval scientific knowledge more easily exploited by a broad scientific community, and support researchers studying e.g. the transmission of zoological knowledge throughout historical periods. Yet, the sharing with related scientific communities remains hampered by the lack of formal semantic reference and terminological standards. For instance, the dolphin is a research topic for modern studies on biodiversity, for archaeozoologists, as well as for studies on Greek mythology wherein the dolphin played an important symbolic role [2]. Nevertheless, when the dolphin is identified in the TEI annotation of the *Hortus Sanitatis* mediaeval encyclopedia[6] or in Pliny the Elder work (*Historia Naturalis*), how to know whether this refers to the same animal? How to know which species is targeted, since the Latin word *delphinus* is used in the textual tradition at least for all Mediterranean regular species of Delphinidae, and labels many different modern taxa (Tursiops truncatus, Delphinus delphis, Stenella coeruleoalba, etc.)? How to relevantly relate those terms with the Delphininae subfamily of modern zoological taxonomy, or even upper terms in the classication (family Delphinidae, order Cetacea)? More generally, how to simultaneously query various archaeozoological, zoological and historical data sources, crosscheck the evidences and make sure that concepts share the same meaning across data sources?

Those challenging questions can be addressed through the use of controlled and widely accepted semantic references. A reference thesaurus shared by sibling scientific disciplines would help to clear the many misinterpretations or conflations made by ancient authors and debated at length in modern critic literature referring to Ancient sources (from P. Belon, 1551, to I. Geoffroy Saint-Hilaire, 1841). The Zoomathia research network[7] addresses this challenge, specifically on the study of rich mediaeval compilation literature on Ancient zoological knowledge, supported by archaeological and iconographic knowledge. The Semantic Web provides powerful models and technologies for connecting and sharing pieces of data while making their semantics explicit. RDF facilitates the combination and sharing of different data sets thanks to the underlying Web technologies and the subsequent Linked Data paradigm. Zoomathia intends to leverage those technologies to annotate and link together various medieval compilations such as the *Hortus Sanitatis*[8], archaeozoological data (I2AF database) and iconographic material. In this context, we chose the TAXREF [3] zoological and botanical

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

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

[5] http://www.tei-c.org/index.xml

[6] https://www.unicaen.fr/puc/sources/depiscibus/accueil

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

[8] This very popular text that enjoyed numerous editions and translations between 1491 et 1547 is not only a landmark in the history of encyclopedias, but also, concerning the naturalistic knowledge, representative of the whole medieval tradition. It provides

taxonomy to build a SKOS thesaurus supporting the integration of these heterogeneous data sets.

This paper is organized as follows: Section 2 presents the I2AF project. Section 3 describes the TAXREF taxonomical reference. Then, section 4 presents our ongoing work on the construction of a SKOS thesaurus based on TAXREF. Finally, section 5 concludes and suggests leads for future works.

2 I2AF: Archaeozoological and Archaeobotanical Inventories of France

During the eighties decade, it was acknowledged that the access to archaeological data by researchers was increasingly challenged by the growing amount of data produced, and hampered by its scattering. The risk of permanent loss was even more worrying. Thus, it appeared obvious that data in archaeological reports had to be systematically and sustainably collected and inventoried, in a heritage perspective, while making them available to all potential users. From 2003 on, several programs supported by multiple French institutes designed, deployed and maintained such a national inventory database. Today, the I2AF is a collection of the French *National Museum of Natural History* (MNHN). It is continuously and increasingly populated with data on flora and fauna from reports of all excavations performed in French territories, whether the bioarchaeological material was already studied or not. Since January 2014, the inventory and knowledge dissemination effort has been actively sustained by a national multi-institute network of bioarchaeologists[9]. When data from excavation reports is imported into the I2AF, it is aligned on two thesauri: a chronocultural thesaurus provides temporal terms with regards to cultural periods (the oldest records date back to the Middle Palaeolithic), and a taxonomic thesaurus of zoological and botanical names, namely the TAXREF taxonomical reference (see section 3).

As the national reference for nature and biodiversity, the MNHN is responsible for scientific and technical coordination of the natural heritage inventory. To this end, it develops and distributes the TAXREF taxonomical reference, and maintains the *National Inventory of Natural Heritage*[10] (INPN), an information system that gathers current (contemporary) occurrence data on fauna and flora of metropolitan France and overseas departments and collectivities. To date, INPN gathers data from approximately 800 data sources aligned on TAXREF. In this context, the I2AF was naturally identified as a potential data contributor to the INPN. This was however challenging due to the discrepancies between both databases in terms of temporal periods and inventoried species. Indeed, while the INPN gathers actual environmental data on wild life, the I2AF also

most of the data available between 1260 and 1320 in western Europe, derived from the late antiquity compilations.

[9] GDR 3644 BioArcheoDat, "Societies, biodiversity and environment: archaeozoological and archaeobotanical data and results on the French territory".

[10] Inventaire National du Patrimoine Naturel: http://inpn.mnhn.fr. Muséum National d'Histoire Naturelle [Ed]. 2003-2015.

provides archaeological data on domestic species, exotic species (not inventoried on any French territory, notably imported by menageries as soon as Roman Antiquity) and possibly extinct species. This issue was solved progressively by enriching TAXREF with new taxa along with the integration of I2AF data into the INPN. As examples we can cite extinct species such as the mammoth and the cave bear, domestic species such as the dog and the ox, and exotic species such as the Barbary macaque.

3 TAXREF: a Taxonomic Reference in Conservation Biology

TAXREF[3] is the French national taxonomic reference for fauna, flora and fungus of metropolitan France and overseas departments and collectivities. It is developed and distributed by the MNHN in the context of the Information System on Nature and Landscapes[11]. TAXREF aims to (i) give an unambiguous unique scientific name for each taxon inventoried on the territory, that marks a national and international consensus; (ii) enable interoperability between databases in (archaeo)zoology and (archaeo)botany, to help the study of biodiversity and support strategies of natural heritage conservation; and (iii) manage the taxonomic changes (synonymy, taxonomic hierarchy).

TAXREF can be browsed on the INPN web site, and downloaded in TSV format (tab-separated values). An on-going work aims to set up a Web service allowing to query the taxonomy and retrieve results in XML or JSON formats. TAXREF is organized as a unique, controlled, hierarchical list of scientific names. Conceptually, it consists of a table wherein one row uniquely describes one scientific name. All taxonomical names are presented in the same way, whatever their taxonomical rank. Most salient fields are listed below:

- *CD_NOM*: unique identifier of the scientific name.
- *CD_SUP*: identifier of the upper taxon in the classification.
- *CD_REF*: identifier of the reference taxon. This may be either the same as *CD_NOM* or a different one. In the latter, *CD_NOM* identifies a synonym of the reference name identified by *CD_REF*.
- *Nom*: taxon scientific name.
- *Nom_Vern* and *Nom_Vern_Eng*: French and English vernacular names.
- *Auteur*: taxon authority (author name and publication year).
- *Rang*: taxonomical rank (phylum, class, order, family, gender, species...), represented by a code of two to four letters.
- *HABITAT*: type of habitat in which the taxon usually lives marine, fresh water, terrestrial...) coded as values from 1 to 8.
- A set of biogeographical statuses, one for each geographical region (metropolitan France and overseas departments and collectivities). E.g.: P stands for present, E for endemic, X for extinct, etc.

[11] http://www.naturefrance.fr/sinp/presentation-du-sinp

As an example, Listing 1.1 shows a JSON excerpt describing the common dolphin using its reference scientific name *Delphinus delphis*, and its synonym *Delphinus tropicalis*. Annotation `"HABITAT":1` states that it lives in a marine habitat. Annotation `"Rang":"ES"` states that the taxon belongs to the *species* taxonomical rank (*ESpèce* in French). Annotation `"GUA":"P"` states that its biogeographical status is P (present) in Guadeloupe, a French overseas department. A comprehensive description of allowed values for the habitat, taxonomical rank and biogeographical status is provided in [3].

```
{   "CD_NOM":60878, "CD_REF":60878, "CD_SUP":191591,
    "Nom":"Delphinus delphis",
    "Nom_Vern":"Dauphin commun a bec court",
    "Nom_Vern_Eng":"Short-beaked common dolphin",
    "Auteur":"Linnaeus, 1758",
    "HABITAT":1, "Rang":"ES",
    "FR":"P", "GUA":"P", "REU":"B", (...)
},
{
    "CD_NOM":60881, "CD_REF":60878, "CD_SUP":191591
    "Nom":"Delphinus tropicalis",
    "Nom_Vern":"Dauphin commun d'Arabie",
    "Nom_Vern_Eng":"Arabian common dolphin",
    "Auteur":"Van Bree, 1971",
    "HABITAT":1, "Rang":"ES",
    "FR":"P", "GUA":"P", "REU":"B", (...)
}
```

Listing 1.1. Example of a JSON representation of TAXREF entries

Currently, more than 450.000 taxa are registered, covering the continental and marine environments. From the temporal perspective, all current living beings are considered as well as those of the close natural history, that is, from the Palaeolithic until now. Usage statistics[12] attest the large variety of people using TAXREF, far beyond the research community: botanic conservatories, associations, public institutions and collectivities, private companies, individuals. Given its wide adoption in various communities, we chose it to build a SKOS reference thesaurus that should be published and linked on the Linked Data.

4 A TAXREF-based Thesaurus for the Linked Data

In this section we present our ongoing work on the creation of a SKOS vocabulary faithfully representing the TAXREF taxonomical reference. SKOS[13] is the acronym of Simple Knowledge Organization System; it is a W3C standard designed to represent controlled vocabularies, taxonomies and thesauri. It is used extensively to bridge the gap between existing knowledge organisation systems and the Semantic Web and Linked Data.

[12] TAXREF usage statistics are not published publicly but can be provided on demand.

[13] http://www.w3.org/2009/08/skos-reference/skos.html

```
@prefix skc: <http://www.w3.org/2004/02/skos/core#>.
@prefix skx: <http://www.w3.org/2008/05/skos-xl#>.
@prefix tc: <http://lod.taxonconcept.org/ontology/txn.owl#>.
@prefix gn: <http://www.geonames.org/ontology#> .
@prefix nt: <http://purl.obolibrary.org/obo/ncbitaxon#> .
@prefix taxr: <http://inpn.mnhn.fr/taxref/>.

<http://inpn.mnhn.fr/taxref/taxon/60878> a skc:Concept;
  skx:altLabel <http://inpn.mnhn.fr/espece/cd_nom/60881>;
  skx:prefLabel <http://inpn.mnhn.fr/espece/cd_nom/60878>.
  skc:broader <http://inpn.mnhn.fr/taxref/taxon/191591>;
  taxr:hasHabitat <http://inpn.mnhn.fr/taxref/habitat#Marine>;
  taxr:bioGeoStatusIn [
   taxr:bioGeoStatus  <http://inpn.mnhn.fr/taxref/bioGeoStat#P>;
   gn:locatedIn <http://sws.geonames.org/3017382/> ];
  taxr:bioGeoStatusIn [
   taxr:bioGeoStatus  <http://inpn.mnhn.fr/taxref/bioGeoStat#P>;
   gn:locatedIn <http://sws.geonames.org/3579143/> ];
  taxr:bioGeoStatusIn [
   taxr:bioGeoStatus  <http://inpn.mnhn.fr/taxref/bioGeoStat#B>;
   gn:locatedIn <http://sws.geonames.org/935317/> ].

<http://inpn.mnhn.fr/espece/cd_nom/60878> a skx:Label;
  taxr:isPrefLabelOf <http://inpn.mnhn.fr/taxref/taxon/60878>:
  skx:literalForm "Delphinus delphis";
  tc:authority "Linnaeus, 1758";
  nt:has_rank <http://inpn.mnhn.fr/taxref/taxrank#Species>;
  taxr:vernacularName "Dauphin commun a bec court"@fr;
  taxr:vernacularName "Short-beaked common dolphin"@en.

<http://inpn.mnhn.fr/espece/cd_nom/60881> a skx:Label;
  taxr:isAltLabelOf <http://inpn.mnhn.fr/taxref/taxon/60878>;
  skx:literalForm "Delphinus tropicalis".
  tc:authority "Van Bree, 1971";
  nt:has_rank <http://inpn.mnhn.fr/taxref/taxrank#Species>;
  taxr:vernacularName "Dauphin commun d’Arabie"@fr;
  taxr:vernacularName "Arabian common dolphin"@en.

<http://inpn.mnhn.fr/taxref/taxrank#Species> a skc:Concept;
 skc:prefLabel "Species"@en;
 skc:exactMatch
   <http://purl.obolibrary.org/obo/NCBITaxon_species>;
 skc:exactMatch
   <http://rdf.geospecies.org/ont/geospecies#TaxonRank_species>.

<http://inpn.mnhn.fr/taxref/habitat#Marine> a skc:Concept;
 skc:prefLabel "Marine habitat"@en;
 skc:relatedMatch
   <http://lod.taxonconcept.org/ontology/txn.owl#MarineHabitat>;
 skc:exactMatch
   <http://purl.obolibrary.org/obo/ENVO_00002227>.
```

Listing 1.2. Example SKOS representation of TAXREF entries

Listing 1.2 shows the proposed SKOS representation of the taxon presented in Listing 1.1, using the Turtle RDF syntax. The keystone of our modelling of TAXREF in SKOS is as follows. Each taxon is represented by a SKOS concept (line 8); its URI is in namespace `http://inpn.mnhn.fr/taxref/taxon/`, which local name is CD_NOM, the TAXREF taxon identifier (see section 3). The `skc:broader` property is used to model the relationships between a taxon and the upper taxon in the classification (CD_SUP). The reference scientific name of a taxon and its synonyms are defined as values of properties `skx:prefLabel` and `skx:altLabel` respectively (lines 9 and 10). They are URIs in namespace `http://inpn.mnhn.fr/espece/cd_nom/`. These URIs have been defined by INPN; today they are dereferenced to a Web page providing a HTML description of the taxon. The label literal values themselves are defined with property `skx:literalForm` (lines 25 and 33). The habitat and biogeographical status are represented by a property value which subject is the URI representing the taxon (lines 12 to 21), while the authorities, taxonomical rank, and vernacular names are attached to labels (lines 26 to 29 and 34 to 37).

We identified existing vocabularies that can be reused in our context, keeping in mind that we wish to link the TAXREF thesaurus with existing, well-adopted data sets, in particular within the Linking Open Data cloud. We first focussed on classes and properties that represent taxon characteristics (habitat, taxonomical rank, name authority, etc.). For example, taxonomical ranks are defined in various ontologies such as the NCBI taxonomic classification[14] and the GeoSpecies ontology[15]. Similarly, the type of habitat is commonly defined in several ontologies such as the ENVO[16] environment ontology. To keep full control over the TAXREF vocabulary, we chose to define terms (SKOS concepts) for the taxonomical ranks (lines 39 to 44), types of habitat (lines 46 to 51) and biogeographical statuses in a specific TAXREF namespace (`http://inpn.mnhn.fr/taxref/`), and align them with concepts of existing vocabularies using the `skc:exactMatch` or `skc:closeMatch` properties. In future works, we intend to align the TAXREF taxa themselves with taxa in other well-adopted taxonomies.

To perform the translation of TAXREF into a SKOS vocabulary, we use xR2RML [4], a declarative mapping language designed to address the mapping of a large and extensible scope of databases (RDB, NoSQL, XML native database, object oriented, etc.) into RDF, by flexibly adapting to various data models and query languages. The produced RDF graph can reuse existing domain vocabularies. A prototype implementation of the xR2RML mapping language, Morph-xR2RML, supports the translation of data from relational databases and from the MongoDB[17] NoSQL document store. To deal with TAXREF, we import its JSON version into a MongoDB instance. Then, we write the xR2RML mapping that describes how to map the result of queries to the MongoDB instance into RDF triples. Finally, the Morph-xR2RML tool coordinates the whole process: it

[14] http://www.ontobee.org/browser/index.php?o=NCBITaxon

[15] http://datahub.io/dataset/geospecies

[16] http://www.ontobee.org/browser/index.php?o=ENVO

[17] http://www.mongodb.org/

parses the mapping description, performs the queries against the database and produces the resulting target SKOS vocabulary according to the mapping.

5 Conclusion and Future Works

In this paper, through a few simple example questions, we have highlighted today's needs of some scientific disciplines, as diverse as Conservation Biology, Bioarchaeology, and Ancient literature, to gather and make sense of heterogeneous data and material. Then, we have described I2AF, a national archaeozoological and archaeobotanical inventory database integrating data from archaeological excavation reports. We have presented the TAXREF taxonomical reference designed to support studies in Conservation Biology. To meet the needs of Archaeozoology and Archaeobotany, TAXREF was progressively extended with taxa from I2AF. It is the first taxonomical reference used to integrate data from Bioarchaeology and Conservation Biology[5].

Then we have presented our ongoing work on the construction of a SKOS thesaurus based on TAXREF. In the context of the Zoomathia research network, we aim to use this thesaurus to support multi-disciplinary studies on the history and transmission of zoological knowledge throughout historical periods, combining the analysis of ancient and mediaeval literature, iconographic and archaeozoological resources. This will require the enrichment of the TAXREF-based thesaurus with philological and cultural information and its geographical extension to other Mediterranean areas (Greece, Italy, etc.). Besides, in order for a large community to benefit from this work, and to spur its adoption by linked-data based applications, future works target the automatic creation of links with other well-adopted open data sets and thesaurus, may they be non-specialized like DBpedia, or domain-specific like the NCBI taxonomical reference.

References

1. C. Callou, I. Baly, C. Martin, and E. Landais, "Base de données I2AF: Inventaires archéozoologiques et archéobotaniques de France," *Archéopages*, vol. 26, 2009.
2. E. Voultsiadou and A. Tatolas, "The fauna of Greece and adjacent areas in the Age of Homer: evidence from the first written documents of Greek literature," *Journal of Biogeography*, vol. 32, no. 11, 2005.
3. P. Gargominy, S. Tercerie, C. Régnier, T. Ramage, C. Schoelinck, P. Dupont, E. Vandel, P. Daszkiewicz, and L. Poncet, "TAXREF v8.0, référentiel taxonomique pour la France: Méthodologie, mise en oeuvre et diffusion," in *Rapport SPN 2014 - 42*, 2014.
4. F. Michel, L. Djimenou, C. Faron-Zucker, and J. Montagnat, "Translation of relational and non-relational databases into RDF with xR2RML," in *Proc. of 11th International Conference on Web Information Systems and Technologies (WEBIST)*, 2015.
5. C. Callou, I. Baly, O. Gargominy, and E. Rieb, "National Inventory of Natural Heritage website : recent, historical and archaeological data," *The SAA Archaeological Record*, vol. 11, no. 1, 2011.

Towards a Better Understanding of Critiques about Ancient Texts using Argumentation

Serena Villata[1] and Arnaud Zucker[2]

[1] INRIA Sophia Antipolis, France serena.villata@inria.it
[2] Univ. Nice Sophia Antipolis, France zucker@unice.fr

Abstract. Ancient texts are interpreted by critics in order to assign them a given semantics. However, the semantics to be associated to these texts is not unique and different critics may have different conflicting opinions about their "correct" interpretation. In this paper, we propose to adopt argumentation theory, a technique to manage conflicting information, together with Semantic Web languages and techniques to provide an overall view of such conflicting critiques, detect what are the different competing viewpoints and what are the strongest arguments emerging from the debate. An ontology for argumentative documents is used to annotate ancient texts, and an example of such annotation is provided about the topic of the *Eternity of the species in Aristotle*.

1 Introduction

Ancient texts are subject to different interpretations depending on the historical context of the text, the personal interpretation of the critic writing the critique, and the literal sense that is associated to the sentences composing the text. In general, apart from the ecdotic aspects (that is textual criticism), the primary goal of a critique is to ascertain the text's primitive or original meaning in its literal sense and its original historical context. In order to have a better understanding of the ancient text and the associated critiques, the following methodologies have been proposed in the literature:

Genre critique : the literary form of the text is analyzed with special attention to genre requirements and tradition (e.g., prose vs verse, letters, epics, dialog, scientific text, etc.);

Source critique : the search for intertextuality, especially directed to the sources which lie behind a canonical text or compilation literature, such as encyclopedias;

Cultural critique : the study of the historical, social, and intellectual context of the text, used to reconstruct the cultural issues at stake and the historical meaning of the work;

These forms of criticism can be adopted or combined to have a clearer understanding of an ancient text, but one step that is missing is *how to deal with situations where different critics have viewpoints that are in contrast with each* *other?* This is the research question we address in this paper, with the aim to detect which critiques could be considered compatible with others and to let emerge competing viewpoints. More precisely, we propose to adopt argumentation theory, a reasoning technique designed to infer non conflicting conclusions starting from a set of heterogeneous possibly conflicting arguments. Our proposal consists in merging argumentation theory as reasoning engine and Semantic Web languages and techniques to represent such data and extract further interesting information.

The combination of these two techniques can actually help in having a better comprehension of a set of critiques from different sources, supporting in such a way an informed choice about the kind of interpretation we aim to back up or to adopt (e.g., in a learning scenario, the fact of providing a clear overall view of a set of different critiques about a specific ancient text can support students in constructing a better grasp of such a text).

The reminder of the paper is as follows: the overall framework we are in introducing is presented in Section 2, and then some conclusions are drawn together with a comparison with the related work.

2 The proposed framework

An abstract argumentation framework [4] aims at representing conflicts among elements called *arguments* through a binary *attack* relation. It allows to reason about these conflicts in order to detect, starting by a set of arguments and the conflicts among them, which are the so called *accepted arguments*. The accepted arguments are those arguments which are considered as believable by an external evaluator, who has a full knowledge of the argumentation framework. A Dung-style framework is based on a binary *attack* relation among arguments, whose role is determined only by their relation to other arguments.

Dung [4] presents several acceptability semantics that produce zero, one, or several sets of accepted arguments. The set of accepted arguments of an argumentation framework consists of a set of arguments that does not contain an argument attacking another argument in the set. Roughly, an argument is *accepted* if all the arguments attacking it are rejected, and it is *rejected* if it has at least an argument attacking it which is accepted. In Figure 1.a, an example of abstract argumentation framework is shown. The arguments are visualized as circles, and the attack relation is visualized as edges in the graph. Grey arguments are the accepted ones. We have that argument a attacks argument b, and argument b attacks argument c. Using Dung's acceptability semantics [4], the set of accepted arguments of this argumentation framework is $\{a, c\}$. The need of introducing also a positive relation among the arguments, i.e., a *support* relation, leads to the emergence of the so called *bipolar* argumentation frameworks [3]. An example of bipolar argumentation framework is visualized in Figure 1.b where the dashed edge represents the support relation.

Our idea consists in *i)* exploiting argumentation to provide an overall view of the set of critiques about an ancient text, and *ii)* to provide a semantic machine

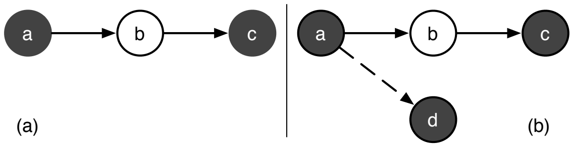


Fig. 1. Example of (a) an abstract argumentation framework, and (b) a bipolar argumentation framework.

readable representation of such argumentative set of possibly conflicting critiques. In order to address the former step, we adopt bipolar abstract argumentation theory such that two possible relations among the critiques are highlighted, i.e., a positive support relation, and a negative conflict relation. Concerning the second step, in order to not introduce yet another argumentation vocabulary, we reuse the SIOC Argumentation module [5], focused on the fine-grained representation of discussions and argumentations in online communities.[1] The SIOC Argumentation model is grounded on DILIGENT [2] and IBIS[2] models. More precisely, we adopt the extension proposed by Cabrio et al. [1] of the SIOC Argumentation vocabulary where two new properties `sioc_arg:challengesArg` and `sioc_arg:supportsArg` whose range and domain are `sioc_arg:Argument`. These properties represent challenges and supports from arguments to arguments, as required in abstract argumentation theory.[3] This needs to be done since in SIOC Argumentation challenges and supports are addressed from arguments towards `sioc_arg:Statement` only.

The following example shows a real instance of ancient text, i.e., *the eternity of the species in Aristotle*, how it is annotated using argumentation theory, and what are the winning arguments we detect.

Example 1. Consider the following five arguments proposed by the critiques about Aristotelian interpretation of the eternity of the species:

Argument 1 : The biological species are eternal.

This argument relies on the following assumptions taken out of the Aristotelian works:

- The general Aristotelian conception is that the world is eternal and uncreated, and so are the parts of the world, either in number or in another way.
- A form, consisting logically in the prior cause of everything, can neither be created nor destroyed.
- The species although not being eternal in number are eternal in form.
- Through generation each organism is reproduced one in form and replicates the same form it has received.
- The final cause of animal is to obtain eternity through reproduction.
- Any kind of generation presupposes the preexistence of a form that has to be transmitted, and this form is transcendent to the individuals.

[1] For an overview of the argumentation models in the Social Semantic Web, see [6].

[2] `http://purl.org/ibis`

[3] The extended vocabulary can be downloaded at `http://bit.ly/SIOC_Argumentation`

  - Even without being fathered (in spontaneous generation) creatures display the form of a regular species.

Argument 2 : The species are not eternal.

This argument relies on the following assumptions:

  - The existence of the form characteristic of members of a kind is contingent on members of that species.
  - The form is not fixed since the individuals constantly differs, because the form (given by the male) has to struggle with the matter-principle, which is the contribution of the female, and it often turns out that the movements of the male are dominated and the form damaged and altered by the power of the matter-principle.
  - Hybrids are fertile, and the offspring has necessarily a form ; yet they are produced by individuals of different species.
  - The species is not a universal type, but a series of historical individuals which are the same in form.

Argument 3 : The species do not exist at all as entities or *ousiai*.

This argument relies on the following assumptions:

  - Aristotle never gives a definition of an animal, whereas definition is an ontological requirement for all substances (*ousiai*).
  - He uses always the word *eidos* (form/species) relative to something else (and not independently).
  - An animal *eidos* is not a substance (*ousia*) according to the definition provided by Aristotle in *Posterior Analytics*, where he states that it should have predicates ranked in correct order (which is impossible in the case of animal, the predicate being simultaneously coordinate and not strictly subordinate).
  - The animal *ousia* in the biological realm is the concrete individual animal.

Argument 4 : Aristotelian zoology tolerates evolutionary mechanisms.

This argument relies on the following assumptions:

  - As Aristotle puts it, new kinds arise from fertile hybrids.
  - Continuance of species does not entail fixity.
  - Individuals are generated in an approximation to a "form" of that species but never reach the perfect form of a species.
  - There are dualizing organisms such as seals, bats, ostriches, ...
  - The offspring offers many differences with its parent.

Argument 5 : In the conceptual frame of Aristotelian biology, the species are fixed.

This argument relies on the following assumptions:

  - The species exists as such only if it has a hereditary form (*genos*).
  - The theory of form and formal cause entails that the species coincide with a fixed pattern.
  - The only reason (or formal cause) of generation is the replication of a form granting living creatures existence.
  - If species were not fixed there would be no possible science of living creatures, since science requires permanency and only deals with firm realities.

These arguments are annotated as follows using the extended SIOC-Argumentation vocabulary, where due to space constraints we show only two assumptions for each of the two main arguments of the example:

```
<http://example.org/aristotle/arg1> a sioc_arg:Argument ;
                 sioc:content "The species are eternal." ;
                 sioc_arg:challengesArg <http://example.org/aristotle/arg2> .
```

```
<http://example.org/aristotle/arg2> a sioc_arg:Argument ;
                 sioc:content "The species are not eternal." ;
                 sioc_arg:challengesArg <http://example.org/aristotle/arg1> .

<http://example.org/aristotle/stat1arg1> a sioc_arg:Statement ;
                 sioc_arg:argues_on <http://example.org/aristotle/arg1> ;
                 sioc:content "The general Aristotelian conception is that the
                 world is eternal and uncreated, and so are the parts of the
                 world, either in number or in another way." .

<http://example.org/aristotle/stat1arg2> a sioc_arg:Statement ;
                 sioc_arg:argues_on <http://example.org/aristotle/arg2> ;
                 sioc:content "The existence of the form characteristic of
                 members of a kind is contingent on members of that species." .
```

More precisely, the arguments (i.e., the general claims that are raised) are expressed as `sioc_arg:Argument`, and the statements (i.e., the statements on which the argument is built) are expressed as `sioc_arg:Statement`. Statements are linked to their related arguments by the property `sioc_arg:argues_on`. The advantage of using RDF is that the stored information can then be queried using SPARQL to retrieve further insightful information from the available data. Finally, Figure 2 shows how the arguments are linked to each other by support and conflict relations. Using acceptability semantics, we have that different set of arguments can be accepted together, i.e., either $\{A_1, A_3, A_5\}$ or $\{A_2, A_3, A_4\}$.

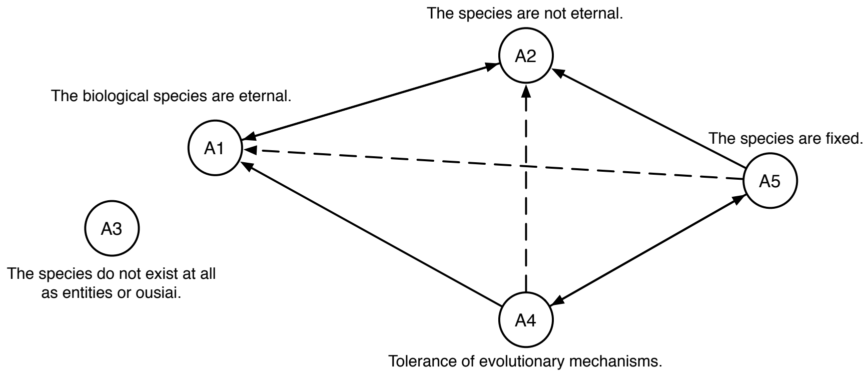

The species are not eternal.
A2
The biological species are eternal.
A1
The species are fixed.
A5
A3
The species do not exist at all as entities or ousiai.
A4
Tolerance of evolutionary mechanisms.

Fig. 2. The complete bipolar argumentation framework resulting from the arguments proposed in Example 1.

Note that an argumentation model where only "challenges" and "supports" are used to represent the relations among the arguments is not sufficient to handle a fine-grained analysis of metaphysical controversies. An example is provided in Figure 2 where argument A_3 is not linked with any of the other arguments because its relation with them cannot be casted in an attack/support relation. For this reason, we plan to consider more complex argumentation structures, namely *argumentation schemes*, to capture finer grained argument patterns in controversies.

3 Conclusions

In this paper, we have proposed a framework to have a better understanding of the critiques about an ancient text by combining argumentation theory and Semantic Web languages and techniques. There are few works with purposes similar to our one. Note that the problem here is that what we call an ancient text is a set of several works (*Posterior Analytics*, *On Generation of Animals*, *Metaphysics* . . .). One of them has been proposed by Vasilopoulou-Spitha and Bikakis [7]. They propose to use argumentation as a tool for the natural representation of claims about cultural artifacts and the arguments they are associated with. This point is shared with our present work. On the other side, they propose to extend ontology-based models like CIDOC-CRM to integrate information about the sources of cultural information (e.g. bibliographic data) enabling users to assess the validity of this information. So, the goal of the two papers is different even if similar, and the adopted methodology differs as well.

There are several open issues to be addressed. First of all, we are currently annotating a dataset of argumentative critiques using the extended SIOC-Argumentation ontology, so that we can use query languages like SPARQL to retrieve further interesting information. Second, we will apply our approach to learning scenarios, where the argumentation graphs of the critiques are used to detect the winning opinions and analyse them, improving their comprehension by students. This methodology could be applied also to internal controversies displayed in ancient texts (such as the question debated by Aristotle in *On Respiration* wether fish breath or not, with conflicting arguments). Third, we need to adopt natural language processing techniques to automatically extract such arguments from texts and to detect the relations among them, starting from the approach presented in [1] and adapting it to such a kind of specific texts.

References

1. E. Cabrio, S. Villata, and F. Gandon. A support framework for argumentative discussions management in the web. In *Proc. of ESWC*, pages 412–426, 2013.
2. A. G. Castro, A. Norena, A. Betancourt, and M. A. Ragan. Cognitive support for an argumentative structure during the ontology development process. In *Proc. of Intl. Protege Conference*, 2006.
3. C. Cayrol and M.-C. Lagasquie-Schiex. Bipolarity in argumentation graphs: Towards a better understanding. In *Proc. of SUM, LNCS 6929*, pages 137–148, 2011.
4. P. Dung. On the acceptability of arguments and its fundamental role in nonmonotonic reasoning, logic programming and n-person games. *Artif. Intell.*, 77(2):321–358, 1995.
5. C. Lange, U. Bojars, T. Groza, J. Breslin, and S. Handschuh. Expressing argumentative discussions in social media sites. In *Proc. of SDoW*, 2008.
6. J. Schneider, T. Groza, and A. Passant. A review of argumentation for the social semantic web. *Semantic Web J.*, 2011.
7. A. Vasilopoulou-Spitha and A. Bikakis. Semantics-based models for the representation of claims about cultural artifacts and their sources. In *Proc. of AImWD*, volume 981. CEUR, 2013.

What terms to express the categories of natural sciences in the *Dictionary of Medieval Scientific French*?

Cécile Le Cornec Rochelois

Centre de Recherche en Poétique, Histoire Littéraire et Linguistique
Université de Pau et des Pays de l'Adour, France

cecile.rochelois@univ-pau.fr

Fabrice Issac

Université Paris 13

fabrice@ldi.univ-paris13.fr

Abstract. This paper deals with several lexicographic problems arising from the choice of hypernyms in the *Dictionary of Medieval Scientific French* (Créalscience database). The generic names used by medieval scholars indicate categorizations which differ from modern classifications. We examine some cases selected from the medieval domain of natural sciences where we can notice a conceptual discrepancy between medieval and modern taxonomies. We first explain the kind of difficulties the editors encounter when they choose a generic name to give a medieval definition, by taking examples in the fields of botany (*arbre, herbe, courge*), mineralogy (*mineral, pierre, metal*) and above all, zoology (*poisson, coquille, ver, mouche, bête*). In order to avoid anachronisms and allow the users to understand conceptual gaps, several means are used by the editors of the *Dictionary*: indicating by an asterisk the words whose meaning has changed and specifying the identification corresponding to the modern categorizations at the end of definition or in an encyclopedic note. After presenting these lexicographic choices, we will wonder how Semantic Web can help to represent and understand this variable lexicon.

Keywords: anachronism; botanic; categorization; classification; conceptual discrepancy; generic names; hypernyms; lexicography; mineralogy; taxonomy; zoology.

Introduction

The *Dictionary of Medieval Scientific French* is a corpus based dictionary which aims to study the genesis of the scientific lexicon between the XIIth and the XVth centuries to reveal the lexical creations and semantic evolutions during this period.[1] It has been published on-line since 2014, with a first version of the letter C. The Crealscience website is based on a XML database, managed by the Basex engine of the University of Konstanz. The DTD, mostly TEI complient, is the result of joint work of the team, which includes linguists specialized in the Middle Ages language, historians of science and lexicographists. This project is behind a CMS dedicated to lexicography, Isilex, which has a consultation/edition interface but also a dictionary data validation module. Indeed, each writer of this collaborative platform is formally identified and a role is assigned. Before being published, any changes must be approved by an administrator. The tool developed uses standard technologies like (X)HTML/CSS/JavaScript on the client side and PHP on the server side. An XQuery engine allows database queries.

The corpus based on scientific texts in old French does not lend itself easily to lexicographic processing for linguistic and epistemological reasons: the identification of scientific terms and of the fields of knowledge themselves; the considerable diachronic variations which affect nomenclatures and taxonomies during the medieval period; the place to be granted to Greek, Latin and Arabic words and to their relations with this vernacular language which builds itself in a situation of diglossia such as scholars often have a Latin vocabulary larger than their French scientific vocabulary; the choice of lemmas among multiple variants. So many difficulties complicate the representation of the medieval scientific and technical lexicon, but the automatic data processing may bring more adapted solutions than a printed dictionary.

We have chosen to focus in this paper on a recurring problem in every scientific field: the choice of generic names to be used in the definitions in order to avoid anachronism.[2] Resorting to a category which does not belong to medieval scientific culture appears as inconsistent. Is it possible to use generic names which did not yet exist in French between the XIIth and the XVth centuries without betraying medieval scientific paradigms? While it is tempting to borrow the words used in the texts of the corpus, we soon realize the limits of this solution: the meaning of many words has changed. What terms should we choose then to express the categories through which the clerics of the Middle Ages saw the world? The choice made by the team was to avoid any reference to later classifications and to express medieval definition, even if formulations sound strange.[3]

[1] About the history of the project and its stakes, see Ducos, J., Salvador, X.L., « Pour un dictionnaire de français médiéval: le projet Crealscience », *Langages*, n° 183, septembre 2011, pp. 63-74.

[2] This difficulty was underlined in the introduction of the Lexicon of scientific language, the starting point of this lexicographical project, about medicine. *Lexique de la langue scientifique (Astrologie, Mathématiques, Médecine…). Matériaux pour le Dictionnaire du Moyen Français*, Jacquart, D., Thomasset, C. (dir.), Klincksieck, Paris (1997), p. IV.

[3] http://www.crealscience.fr/DFSM/fr/Projet [consulté le 01/03/2015]

Our purpose in this paper is to deal with these choices and their implications by considering cases which involve gaps between medieval and modern taxonomies, in domains which are a matter of natural science: botany, mineralogy and above all, zoology. We will first explain the kind of difficulties the editors encounter when they choose a generic name to give a medieval definition. Then we will set out the lexicographic solutions that have been proposed and the way Semantic Web can help to represent and understand this variable lexicon.

1. Lexical anachronisms, cultural anachronisms: the stumbling blocks of definition

The definitions including unattested terms in medieval French and referring to anachronistic notions are obviously to be excluded. So, for plants or animals, Linnaean definitions from the dictionaries of modern French are not suitable. The *coloquinte* was certainly not for the medieval clerics a "plante grimpante de la famille des Cucurbitacées, originaire de la Méditerranée orientale et dont le nom savant est *Citrullus colocynthis*" (TLFi). It does not seem wise to define fly as an *insecte*, oyster as a *testacé*, crab as a *crustacé*, dolphin as a *mammifère* or frog as a *batracien*: not only because these terms were not a part of the French scientific lexicon before the XVIth century[4], but furthermore, the corresponding notions do not seem to be relevant taxonomic criteria in the Middle Ages. It does not mean that the information which allows us to define these zoological categories today was ignored. So, the ancient knowledge on viviparity in dolphins and whales or on their udders was available in the medieval encyclopedia, because their authors did not ignore the aristotelian heritage. Nevertheless, these criteria which lead us today to distinguish fish from marine mammals did not define a special class.[5] If viviparity as well as the presence of udders is mentioned among other characterizations, they do not appear at the beginning of encyclopedic articles and give rise only to occasional links. The medieval texts in Latin as in French leave no doubt on this matter: dolphins and whales are fishes.

The modern distinction between *crustacés*, *testacés* and *céphalopodes* does not seem more relevant to define the concerned aquatic creatures in the *Dictionary*. The outlines of these categories seem nevertheless clearly drawn by Aristote: among the

[4] The first attestation is found in 1542 for *insecte* (FEW, vol. 4, p. 710a, *insecta*) and it does not yet correspond to the current sense because it includes gastropods, amphibians and lizards, as it will be the case until XVIIth century. *Testacé* in its zoological sense is used for the first time in French in 1578 (FEW, vol. 13, p. 282b, *testaceus*). Finally, it is necessary to wait until 1713 for *crustacé* (TLFi), 1791 for *mammifère* (FEW, vol. 6,1, p. 134b, *mamma*) and 1806 for *batracien* (TLFi).

[5] As Aristote never supplies normative definition of the genre of *cètes*, he does not introduce any exclusion between *cètes* and fishes. The cetaceans of the modern science do not exist as class different from fishes before XVIIIth century. See Zucker, A., « Étude épistémologique du mot κητος », *Les Zoonymes : actes du colloque international tenu à Nice les 23, 24 et 25 janvier 1997*, Publications de la Faculté des lettres, arts et sciences humaines de Nice (Centre de recherches comparatives sur les langues de la Méditerranée, 38), Nice (1997), pp. 425-454.

animals without blood, as opposed to animals with blood such as fish, the Greek scholar distinguishes soft animals (μαλακόι), close to our cephalopods, from animals with a soft shell as our *crustacés* (μαλακοστράκοι) and animals with scaly shell (ὀ στρακόδερμα) which correspond to our *testacés*. This antique taxonomy has left marks in the learned works of the XIIIth century. We find the expressions *omnis piscis et animalia mollis teste* used by Thomas of Cantimpré[6] or *animalia durae testae marina* by Vincent of Beauvais.[7] The group of the animals without blood is explicitly mentioned in the *Speculum naturale*, which copies Aristote, but also Pline who had already compiled the zoological books of Stagirite.[8] However, following the example of Pline, the medieval encyclopedists do not use this distinction.[9] It occurs in a heterogeneous chapter of Vincent of Beauvais about the diversity of fish, on the same level as the opposition between marine and freshwater fish or the category of fish which carry a stone inside their head. Blood animals are the subject of no peculiar chapter and do not constitute a visible group in catalogs of species. Even if the Dominican encyclopedists occasionally remind us that certain species are included in one of the three categories of bloodless animals, this information has no consequence on their classification: they are placed most of the time between two spindle-shaped fish. Encyclopedists generally include soft animals with *pisces*. For example, chapter 18 of the book XVII of the *Speculum naturale*, where bloodless animals are evoked, is entitled *De diversis generibus piscium* and *malaciae* are called a *genus piscium*. When we turn to French texts, the names used for these categories of animals seem to move further away from the Aristotelian model, especially as the uses vary. In 1267, Brunet Latin provides the following definition for *coquille*:

> *Coquille est un poissons de mer enclos en charsoiz come une escavris, et est toute raonde; mes ele l'ovre et clot quant ele viaut, et son manoir est au fon[t] de la mer.*[10]

[6] Thomas of Cantimpré, *Liber de natura rerum*, Boese, H. (éd.), Walter de Gruyter, Berlin-New York (1973), VII, 1, p. 251 : *Omnis piscis et animalia mollis teste modicum dormiunt.* "All fishes and animals with soft head sleep little."

[7] Vincent of Beauvais, *Speculum naturale, Bibliotheca Mundi Vincentii Burgundi… Speculum quadruplex*, éd. de Douai 1624 (repr. Anastatique Akademische Druck- und Verlagsanstalt, Graz, 1965), 4 vol., XVII, 18, col. 1262.

[8] *Ibid.*: *Aristoteles* [...] *In quibusdam marinis non est sanguis, ut est saepia, et karabos, et omnia quae plures habent quatuor pedibus.* Plinius ubi supra. *Tria genera sunt aquatilium sanguine carentium. Primum, scilicet quae appellantur mollia. Deinde crustis tenuibus contecta, postremo testis duris conclusa.* "*Aristote* […] Some animals have no blood as the cuttlefish, the spiny lobster and all those who have more than four legs. *Pline*: There are three genres of aquatic animals without blood. At first those who are called soft animals, then those protected by thin crusts, finally those locked in hard shells."

See Pline, *Histoire naturelle*, IX, 83, de Saint-Denis, E. (éd.), Les Belles Lettres (CUF), Paris (1955), p. 64.

[9] About the warping of the Aristotelian classification of *crustacés* and *testacés* in Pline's works, see *Hortus sanitatis: Livre IV*, Fishes, chapter 16, notes 1. https://www.unicaen.fr/puc/sources/depiscibus [Site consulted on 02/03/2015].

[10] Brunet Latin, *Tresor*, Beltrami, P. G., Squillacioti, P., Torri, P., Vatteroni S. (éd.), Giulio Einaudi, Turin (2007), I, 133, p. 236.

As a matter of fact, this excerpt deals with the oyster. We nevertheless understand throughout the text that the word *coquille* is also suitable as hypernym for the *pourpre* (*Murex* Linnaeus, 1758) and the crab.[11] *Testacés* and *crustacés* are thus mixed-up under this metonymical name, which gives the animal the name of the shell. In 1372, Jean Corbechon's translation of the *Liber of proprietatibus rerum* gives two more precise generic expressions for *testacés*: "oistres, molles et aultres poissons qui ont forte escaille" translates the latin *ostreae et alii quidam pisces in conchis degentes* and "une manière de poissons en ostree ainsi comme oystre" is used as a substitute for the two names of species *murices* and *conchylia* which are found in the text of Bartholomaeus Anglicus. In spite of a generalization attempt, the translator does not seem to worry about naming a defined class.[12] While both excerpts deal with oyster and similar creatures, the formulation varies: within a few lines, "poissons qui ont forte escaille" become "poissons en ostree ainsi comme oystre," which seems to indicate the absence of stable terminology in French for this category.

The modern zoological terms of classification which we have just evoked are linked to scientific data which were often collected in the learned works of the Middle Ages without being classification criteria. The lexical anachronism then involves abstract anachronism.

The choice of the chronological border between the relevant terms and the terms considered as anachronistic should be questioned. First of all, why would terms not attested in French before 1500 not have their place in the *Dictionary*? For instance, the concept corresponding to the Linnaean family of *Cucurbitaceae* has obviously existed since medieval times. Nothing challenges the equivalence between species grouped under the hypernym *courge* and the *Cucurbitaceae* of the modern botany, also named *courge* in modern French. However, the simple mention in the definition of a term referring to the Linnaean classification would be enough to distort things and *courge*, which still has a generic value in French, seems more adequate. A word like *amphibien* is undeniably anachronistic from a lexical point of view: it is Rabelais who introduces this Hellenism in French in 1553. It does not prevent medieval scholars from mentioning the customs of the animals which, according to the expression of Jean Corbechon, "vivent partie en eaue et partie en la terre" and "nagent et vont sus la terre si comme font les cocodriles et les chevauls d'yaue et mout d'autres qui vivent en terre et en yaue". The phrase compensates for the absence of an adjective; by excluding on principle *amphibien* from the elements of definition, the editors of the *Dictionary* condemn themselves to a circumlocution which can seem curious to readers who know it as a common term today. The precautionary principle consisting in excluding the unattested words before 1500 is certainly an inconvenience: common terms like *amphibien*, *carnivore* or *migrateur* are excluded even though, conceptually, they do not seem anachronistic.

[11] *Ibid.*: "Une autre coquille est en mer qui a nom morique, et li plusor l'apellent oistre, por ce que quant ele est taillie environ lui il en issent larmes, de quoi l'en taint les porpres; et cele tainture est de son charcois. Une autre coquille est que l'en apele cancre, por ce que il a jambes et est raonde ;"

[12] We notice moreover that this process is not systematic: the Latin *cancri et huiusmodi* is translated by "escrevices, escrevices de mer".

On the other hand, we can wonder if certain words attested before 1500 are suited to the state of the knowledge during the four centuries which preceded? Is it advisable for example to use the adjective *mineral* in the *Dictionary*? According to etymological dictionaries, the first occurrence of the noun meaning “corps inorganique qui se trouve dans l’intérieur de la terre ou à sa surface” would be in 1538 and the use of the adjective in 1516. But an earlier occurrence is mentioned by the *Dictionary*, in the works of Nicolas Panis in 1478 about arsenic:

> *Arcenic et orpigment, ce sont mineralx et sont sublimes et sont chaulx ou tiers, secs ou second et oultre [...]. (Nicolas Panis, Guidon,1478, tr.VII, doct.1, chap.7)*

Thus, "matière minerale" seems appropriate to define arsenic, whose classification is all the more delicate as its modern definition requires knowledge of chemistry.[13] The reference to Nicolas Panis provides a contemporary scientific guarantee. But does that justify extending the use of the noun or the adjective to definitions of the other terms referring to stones or metals? A systematic use of the expression "matière minerale" is not relevant because the categories "pierre" and "métal" are more precise. Furthermore they seem to correspond better to the representations of medieval scholars. The distinction between stones and metals are very clearly formulated by Jean Corbechon.[14] We might be tempted to choose "minéral" for materials which, following the example of arsenic, are neither metals nor stones, such as antimony. However, antimony is mentioned in medical works which do not propose definitions or classifications. The question is whether antimony is part of the class of minerals such as medieval scholars conceived it. When the word *antimoine* was used on 1256 by Aldebrandin of Siena in his *Régime du corps*, was it already considered by learned contemporaries to be a mineral? In the time of Nicolas Panis, was there a common scientific idea of what constitutes a mineral? The use of "minéral" in the definitions supposes a lexicological work on the meaning of this term when it is used by scholars like Nicolas Panis and, more widely, on the meanings of the Latin word *mineralis* and their evolution throughout the Middle Ages.

Even if the chronological limit is of course debatable, by excluding as a matter of principle terms unattested before 1500, we limit to a certain extent the references to paradigms later than the Middle Ages. As regards zoology, the date of 1500 allows us in particular to not include terms which appear in French through naturalists like Pierre Belon du Mans or Guillaume Rondelet, in a pivotal period when zoological knowledge fundamentally evolves. We do not claim to define terms as medieval scholars would have and this choice creates difficulties in the formulation of definitions. But it seems essential to adopt these chronological restrictions in order to avoid anachronism and to help the editors to harmonize their definitions. Nevertheless this

[13] This is the définition found in TLFi (*Trésor de la langue française informatisé*): « Corps simple solide, de symbole As, d'aspect métallique, de couleur gris acier possédant à la fois des propriétés de métal et de métalloïde. »

[14] Jean Corbechon, *Livre des propriétés des choses*, Paris, BnF, fr. 16993, XVI, 75, f. 236ra.

precaution is not enough to avoid notional anachronisms because numerous terms were left in French with important semantic evolutions.

2. Lexicographical expressions of the conceptual discrepancy

One of the main goals of the *Dictionary* is to show the discrepancy between medieval and modern taxonomies. Three primary means were selected to indicate semantic evolutions and relationships with current terminologies: the asterisk behind the terms used in a medieval sense, the addition of a common name in modern French at the end of definitions and encyclopedic notes. The asterisk marks the terms whose meaning has evolved in comparison with modern language. It allows formulating definitions by means of taxemes which changed extension. Such is the case of *poisson** (fish), used in the broad sense of "créature qui vit dans l'eau" throughout the Middle Ages, and which appears for this reason in the definitions of dolphin or crab. The *coquille* mentioned by Brunet Latin will thus be defined as a "poisson* dont le corps est protégé par une coquille ou une enveloppe rigide."[15] Because of the formulations of certain authors of the corpus, the attention of the editors of the *Dictionary* is often drawn to taxonomic discrepancies, which incite them to use asterisks. We hesitate for instance to follow our first idea by defining the *artemisia* or the *aloe* as plants when Jean Corbechon presents them as *herbes*. Out of caution, we thus prefer *herbe*, accompanied with an asterisk which shifts to the article dedicated to this generic name the question of the relationship between the medieval meaning of *herbe* and the modern meanings of *plante* and *herbe*. If we refrain from using the name *insecte*, anachronistic, we can turn to the terms *mouche* (fly) or *ver* (worm) used by the clerics, since their generic value is well specified in the *Dictionary*.

This process which consists in tracing the medieval lexicon of the scholars not to deform their concepts has something reassuring; but what to make when all the authors of the corpus do not use the same hypernym? Is the medieval crab rather a *poisson**, as suggested by a majority of texts, or a *coquille**, as used by Brunet Latin? Yet this is only a hesitation on the extension of the hypernym: since a *coquille** is a *poisson**, there is no contradiction. Another more complex case reveals the difficulties that the editor can encounter choosing a generic name because of the instability of medieval terminology: the crocodile. Jean Corbechon presents it – rather logically from our point of view – as a *poisson*, next to the *cheval d'eau*, or, in other words, the hippopotamus. But Brunet Latin, who uses the term *poisson* to introduce the hippopotamus and mentions the crocodile within his inventory of aquatic creatures, prefers for the latter the more general term of *animal*. It is the word *beste* that appears in the quotation extracted from the *Chirurgie* of Henri de Mondeville and at the beginning of the note "La cocodrille" in the long version of the *Bestiaire* attributed to Pierre de

[15] The word *coquille* meaning "stiff shell" is already used in the Middle Ages and is thus in the definition without an asterisk.

Beauvais.[16] Further, we read in this last note that the crocodile is a "serpent marage". According to the texts of the Créalscience base, the medieval crocodile is thus at the same time an *animal*, a *beste*, a *poisson* and a *serpent d'eau.*[17] What category is it then advisable to select in the definition? Is it better to favor the majority use, by taking the risk of seeing this choice questioned by new reports, or to adopt the nomenclature of an author like Jean Corbechon, who defines most of the hypernyms and strives to be consistent? Another parameter must be taken into account, that of the coherence among entries. To define the hippopotamus as a *poisson** and the crocodile as a *serpent** would be inconsistent. The lexicographer cannot adopt a terminology as elastic and heterogeneous as his diverse sources. In this particular case, it seems reasonable to explain the taxonomic variations due to the hybrid nature of this animal in the note and to choose in the definition the term *animal* because it is the most neutral and most general, and the closest to modern use.

The asterisk actually raises another issue: its multiplication harms the legibility of the definition. Yet its presence could be justified after a large number of terms if we use it as soon as a discrepancy exists between medieval and modern knowledge. As regards gourds, as the list of the species quoted by medieval texts (*concombre*, *citrule*, *courge sauvage*, that is *courge d'Alexandrie*, that is *coloquinte*) does not correspond exactly to the species a modern botanist would recognize as gourds, and in the absence of an explicit medieval definition, the asterisk is imperative by caution: how can we be certain that medieval clerics were referring the same gourds that we are? Should we then put an asterisk after *corbeau*, *cheval* or *chien* because the medieval representation of these animals differs from the modern definition? To avoid the proliferation that the application of this principle could lead to, we prefer to limit the use of the asterisk to the terms which involve taxonomic gaps or whose semantic evolution can be confusing. So, besides *mouche*, *ver*, *poisson*, *herbe*, *arbre* and other generic names borrowed from medieval scholars, we append asterisks to terms like *lièvre* (which can be a rabbit) or *ongle* (used for the hooves of ungulate mammals).

The medieval definition can make certain familiar animals unrecognizable. It then proves useful to specify the identification of the animal defined by its current name. Would the reader be able to recognize the animal named *boterel* hidden under the definition "Ver* venimeux aux yeux rouges qui fréquente les lieux humides et subit une mue", if we do not mention that it is the *crapaud* (toad)? To make the consultation easier, this identification is added at the end of definition. However, it is not al-

[16] Henri de Mondeville*, Chirurgie,* 1314, éd. Bos, ch. 247, p. 71 ; *Bestiaire version longue, attribuée à Pierre de Beauvais),* 42, p. 194.

[17] Latin sources use hypernyms *animal*, *bestia* and *belua*. In Vincent of Beauvais's work, the crocodile appears in book XVII, that of the fishes, but in the section dedicated to the marine monsters, at the end, after fishes "qui pure naturam and speciem piscis habent". The crocodile is also found in Thomas of Cantimpré's book 6 of *De Natura rerum*, "De monstris marinis". Albert the Great places the crocodile in book 24 of *De Animalibus*, the inventory of aquatic species, among other *pisces*. He compares it with the lizard, without using any other generic name than *aquatici* or *belua*. It seems that the three Dominicans classify the crocodile among the aquatic species because it lives in water, even if their sources do not present it explicitly as a *piscis*.

ways convenient or possible. So the equivalence between the *cète* and the animal which we name today *baleine* is not at all obvious. At least this notation presents the advantage of making an identification possible for a reader knowing nothing of the antique or medieval *cète*. Editors of the *Dictionary* can use the encyclopedic note to explain the variable relationship between the textual creature and the real animal. Concerning *chamel leopard*, to assimilate the animal to the giraffe would be meaningless. Here is the definition which we can develop by compiling the medieval information: “Animal doux et beau, qui vit en Éthiopie et présente une tête semblable à celle du chameau, un cou de cheval, des pattes de buffle et un pelage tacheté comme celui du léopard”. If this animal has given rise to such a strange description, as far as to look like a hybrid, a textual fancy, it is exactly because the connection with the real animal known in old French from the XIIIth century under the names *girafe* or *orafle*, inherited from the Arabic, was ignored. As it has been showed by Thierry Buquet, it will be necessary to wait until the end of the XVth century for the *camelopardalis* bequeathed by the antique knowledge and Deuteronomy to be identified with the giraffe.[18]

The *chamel leopard* raises another recurring problem. Its definition looks more like a description than a definition in compliance with lexicographical uses. It is due to the nature of the zoological knowledge (and to a certain extent botany) passed on by medieval works. The longest notes of bestiaries and encyclopedia proceed by accumulation of natural properties according to the compiled sources, so that heterogeneous elements from our point of views are mixed without explicit hierarchy. Let us use as an example the text dedicated by Brunet Latin to the crocodile in his *Tresor*, where we can find the following information: the crocodile is a four-legged animal of yellow color born in the Nile; it is twenty feet long and armed with big teeth and long claws; its skin is so resistant that it would not feel a blow from a stone; it lives on the ground in the daytime, in the river at night; it has no tongue and it is the only animal in the world able to move its upper jaw while keeping lower jaw immobile; it is a rival to the hydra. What should be selected in the definition? Collecting the elements in the corpus necessary to reconstitute a modern definition would mean deforming the representation expressed by the medieval text. If it is possible to distinguish striking properties, to select certain data either because they appear at the beginning of the notes, or because they are obviously recurring, we can understand the nature of a given animal, plant or stone in medieval culture and formulate organized definitions, reflecting the particularity of a medieval system of representations.

Thus for the eel, the link with the snake appears as essential information. Indeed the Isidorian etymology which connects *anguilla* to *anguis* is systematically mentioned and this comparison enlightens most of the medieval representations attached to this fish. This characteristic will thus have its place in the medieval scientific definition. However deciding which characteristics are striking is far from obvious. For instance, we are tempted to select elements which allow us to represent the crocodile such as we know it (“Animal vivant à la fois dans l'eau et sur la terre, au cuir jaune

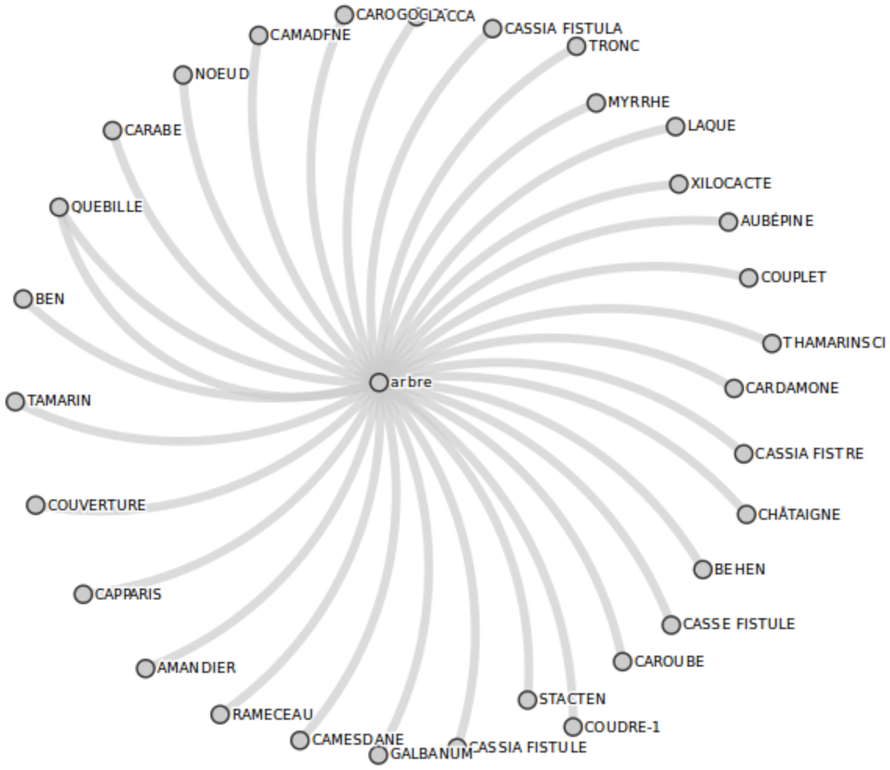
[18] Buquet, T., “La girafe, belle inconnue des bibles médiévales. Camelopardalis : un animal philologique”, *Anthopozoologica* 43 (2), 2008, pp. 47-68.

résistant, dangereux pour l'homme") and to relegate the properties ruled out of the definition to the encyclopedic note. Nevertheless, this solution is not satisfactory because the medieval point of view is falsified to a certain extent: in bestiaries and in the iconographic tradition, the fight against the hydra which devours it from within appears as a striking property of the crocodile. We can thus wonder if this characteristic, which may have been prominent for a medieval cleric, would not have its place in a zoological definition. And how might we justify the exclusion of the crocodile's tears or its opposing jaw which are so peculiar to this animal? It would be useless to look in these encyclopedias for an organization of the knowledge comparable to the one that will be proposed by naturalistic doctors of the XVIth century. Encyclopedists and medieval translators did not try to find the "marques" by which the scholars as Fuchs, Belon or Rondelet will structure their descriptions of plants and animals to allow their identification.[19] The medieval taxonomies are rather organized around prototypes from which the various species of the category are more or less distant. In order to give an exact definition, we need to know what is the best example of *poisson**, *ver** or *herbe**. The selection of the striking characteristics depends thus essentially on the appreciation of the editors of the *Dictionary*, that is on their representation of medieval culture.

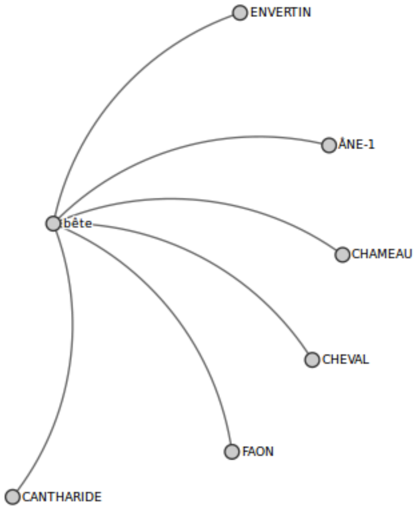
3. The opportunities of Semantic Web

What advantage do they provide data Web resources to better represent and understand these taxonomic and/or lexical differences? As part of work on the definitions, the main advantage is in our opinion the opportunity to create links between the entries to build semantic networks. They can be established by different ways: create explicit links (through the synonym tag for instance); create implicit links based on the content of the entries (keywords in the definitions or in the encyclopedic notes); use the navigation history that shows the interests of users and target certain playback modes; or, finally, combine the previous strategies. The Créalscience already provides access to networks constructed from the contents of definitions and a graph of synonyms. It is then possible for someone who has no access to the vocabulary of the ancient language to navigate through the *Dictionary* and to find, for instance, kinds of tree without knowing their ancient name.

[19] Philippe Glardon, *L'Histoire naturelle au* XVI^{e} *siècle. Introduction, étude et édition critique de* La nature et diversité des poissons *de Pierre Belon (1555)*, Droz (Travaux d'Humanisme et Renaissance, n° CDLXXXIII), Genève, 2011, especially pp. 207-240. According to Philippe Glardon, the "marque" or "note" is a decisive cultural tool in the emergence of natural history in the XVIth century, "l'indispensable trait d'union entre le texte et la représentation graphique, l'élément discursif qui doit emporter la décision de l'identification, qui s'est resserrée autour de l'espèce réelle".



Besides the interest of this kind of representation for users, it definitely speeds up the work of editors: it helps to verify and coordinate formulations. The use of graphs allows for example to instantly measure the degree of polysemy of a term, to see what terms is assigned a hypernym, to cross-check in order to replace uniformly anachronistic terms and to ensure that the same generic term was used by all editors for specific terms relating to the same category.



In this case, the graph allowed us to notice a mistake: *cantharide* in a zoological sense was defined as a *ver**(which was right), but in a medical sense the generic name *bête** had been retained.

However any automatic processing applied to the definitions seems risky for the moment. Imposing such as hypernym *poisson** in all definitions which deals with an animal that lives in the water would be a contradiction. As we have said, the medieval *crocodile* is not only a *poisson** and other animals related to water as the *crapaud* clearly fall into the category of *vers**. At this stage of research, we cannot do without a trial and error linked to both the body and the lexicographical process itself.

Other applications of graphs can be considered for the future. As synonyms are tagged by the editors, their linking can provide a representation of onomastic fields. It will be interesting then to compare the networks of different related words: *arbre** and *herbe**, or on the other hand *poisson**, *bête**, *ver** and **coquille*. The addition of a "comparison" tag would integrate an important criterion of identification and classification in the field of natural sciences. Certain plants or animals are referred to as prototypes. For example, it could be interesting to study the place of the word *pomme* among the other fruits. Contrary to the generic Latin word *pomum*, the French word has a specific sense. But a comparison with the apple can still be used to describe another fruit[20]. A query with the word *serpent* would probably show a set of several related animals. Some associations could be unexpected because medieval sciences do not involve the same criteria as ours. Furthermore, the explanation of the relationship

[20] Lemon (*citron*) is described as a kind of apple (“une manière de pomme”) for instance by Olivier de la Haye (*Poème sur la grande peste de 1348,* 1426, éd. H. Georg, p. 187).

with the Latin scientific lexicon could give interesting results. Indeed, a significant portion of corpus texts are translations of Latin works or take nomenclatures expressed at the same time in Latin. A specific tag under which one would notice the Latin equivalent would refine the data on lexical creation process.

Finally, the *Dictionary*, which already includes scientific senses identified in the AND (Anglo-Norman Dictionary), the DEAF (Dictionnaire Étymologique de l'Ancien Français) and DMF (Dictionnaire du Moyen Français), if it continues its expansion, should allow clear observation of semantic changes from the XIth to the XVth century. Concerning the confrontation with modern scientific terminology, it is already possible even if the operation is in its early trials. At more or less long term, it should be possible to compare the dictionary systematically to modern French nomenclature of specialized vocabularies. It is likely that some modern terms will not find resonance in the database, even including among the association criteria text notes. It will then be necessary to ask whether this absence is caused by a wrong description to correct in the database, a lexical creation that appeared after 1500 or an epistemological discrepancy that could then be explained in the note. The *Dictionary* would then allow queries from modern terms, by a non medievalist user curious about ancient representations and evolution of knowledge. It would thus give users the opportunity to question their own conceptual tools. There is much to be done before we reach a satisfactory model of scientific neologisms and semantic relationships between categorizing terms during medieval times, but we can assume it is worth meeting the challenge for linguists and historians of science.

Conclusion

Writing articles in the *Dictionary of Medieval Scientific French* raises permanent questions about the choice of generic names and the nature of definitions. Indeed, the lexicographic project leads us to forget the categories through which we think of nature and our reflexes of definition. The difficulty lies, on one hand, in the nature of the inherited knowledge, the fact that data are not selected yet and not organized into a hierarchy according to stable and homogeneous criteria from the XIIth to the XVth centuries, and on the other hand, in the semantic evolution of the lexicon. The meaning of generic words evolved at the same time as taxonomies, which entailed a lexical vacancy for certain concepts: there is no word in modern French for *coquilles** of Brunet Latin, for *poissons** in the medieval sense or for *herbes** as Jean Corbechon understands them. Furthermore, the contrast is great between the lexical freedom and the conceptual flexibility that the French corpus natural sciences lets us perceive and the necessary coherence of a dictionary, especially an electronic one. Far from the monosemic ambition of modern scientific language, medieval scientific works in French are characterized by a linguistic variety that their authors obviously did not try to reduce.[21] They allow us to see the evolution of thought or at least practices of trans-

[21] See, for instance, about the multiplicity of names for the same concept, Ducos, J., Salvador, X.L., « Pour un dictionnaire de français médiéval : le projet Crealscience », *Langages*, n° 183, septembre 2011, p. 65.

lation. Categorization is organized around a prototype, an exemplary species. The aim of the project is to show evolutions which are often not linear and which we cannot understand well without comparing them to the lexical uses of Latin and without placing it in a wide cultural context, by being careful to take into account modern knowledge which is the prism through which the user reads the *Dictionary*. To find the balance between allegiance to medieval scientific culture and coherence, homogeneity and legibility which are expected from a lexicographical database, the encyclopedic note offers a useful space to collect the essential philological, linguistic and scientific information to explain the relationship with modern nomenclature.

The use of a collaborative platform has already taken forward the project. We can hope that browsing by graphs, taking into account specific tags (such as keywords in definitions and notes, fields, synonyms, Latin equivalents or comparisons) will allow to improve the analysis of the birth of French scientific terminology and to connect this work with other web projects concerning the reconstruction of former scientific paradigms.

Biblissima's Prototype on Medieval Manuscript Illuminations and their Context

Stefanie Gehrke, Eduard Frunzeanu, Pauline Charbonnier, and Marie Muffat

Equipex Biblissima, Campus Condorcet, Paris, France
`{stefanie.gehrke,eduard.frunzeanu,pauline.charbonnier,marie.muffat}@`
`biblissima-condorcet.fr`
`http://www.biblissima-condorcet.fr`

Abstract. Biblissima is an online digital library, which provides easy and coordinated access to a huge and complex mass of documentation on manuscripts and early printed books, the texts contained therein, their circulation and their readers, from the 8th to 18th centuries. This workshop presentation will give an overview of the steps followed and decisions made along the way to releasing a first prototype of the Biblissima portal: from mapping data to a common ontology, via the establishment of a thesaurus, to the technical development of a single interface and a common triple store for data deriving from different iconographic databases on medieval manuscripts.

Keywords: cidoc crm · frbroo · medieval manuscript · illumination · interoperability · descriptors · thesaurus · historical place names · semantic web · linked data · library · Middle Ages · Humanism · Renaissance

1 Objectives of the Biblissima Observatory

Biblissima - Bibliotheca bibliothecarum novissima - is an observatory for the written cultural heritage of the Middle Ages and the Renaissance, developed through the French government programme *Équipements d'excellence*, part of the *Investissements d'avenir* [1]. The observatory focusses on documents written in the main languages of culture in Medieval and Renaissance Europe (Arabic, French, Greek, Hebrew, Latin, etc.) and contributes to a better understanding of the circulation of texts, the evolution of libraries and the transmission of knowledge in Europe from the 8th to the 18th century.

In addition to its contributions to research, Biblissima plays an important role in disseminating knowledge about the written cultural heritage of the Middle Ages and the Renaissance to the widest possible audience.

Led by the Campus Condorcet, the Biblissima project brings together eight French partner institutions in the fields of research, teaching and cultural heritage, including the BnF (National Library of France) and the IRHT (Institut de recherche et d'histoire des textes).

The two main components of the observatory are a cluster of the project's data on manuscripts and early printed books currently found in as many as 40 databases in different formats and with different research interests (including illuminated manuscripts, history of the transmission of texts and history of collections) and a digital image repository. The databases will be interconnected using semantic web technologies and linked to a platform for digital editions and to the project's digital image repository.

1.1 Semantic Web Solutions for Historical Data

In order to handle the heterogeneity of the database formats (MySQL, EAD, TEI P5, UNIMARC, etc.) and the variety of Biblissima's data (manuscript cataloguing databases, textual editions, iconographic databases) we have chosen to use the *CIDOC Conceptual Reference Model* (Comité International pour la Documentation Conceptual Reference Model [2]) and *FRBRoo* (Functional Requirements for Bibliographic Records object oriented [3]) as framework for a project-specific extension of those ontologies that facilitates the internal mapping to a single common model and allows the partners to expose their data in RDF compliant to a globally established standard.

CIDOC CRM is an accepted ISO standard (ISO 21127). As an event-centric ontology it covers different phenomena in space and time like provenance, copying of texts, creation of works and expressions, as well as the production of information carriers and attribute assignments. As CIDOC CRM and FRBRoo (which combines the CIDOC CRM approach with the common vocabulary for the transmission of works (WEMI) that is provided by the FRBR model) are generic models for the museum and library domains, it was decided to define a few more specific classes and properties related to manuscripts, early printed books and illuminations. For example, within the scope of the Biblissima project a medieval manuscript is an instance of the class bibma:Manuscript, which is a subclass of `frbroo:F4_Manifestation_Singleton` ("This class comprises physical objects that each carry an instance of F2 Expression, and that were produced as unique objects, with no siblings intended in the course of their production"). An instance of a `bibma:Manuscript` might be composed of several parts (`bibma:Component`) and might carry both text and illustrations.

As regards the illustrations, there are several possible modelling solutions in CIDOC CRM, such as E38 Image ("This class comprises distributions of form, tone and colour that may be found on surfaces such as photos, paintings, prints and sculptures or directly on electronic media") or its subclass E36 Visual Item ("This class comprises the intellectual or conceptual aspects of recognisable marks and images"). These solutions have been adopted both for book illustrations by the "*Illustrations of Goethe's Faust*" project [4] and for maps by the "*Carte de la nouvelle frontire Turco-Grecque*" project [5]. In order to model the illumination genre, we decided instead to define an illumination as an instance of a class called `bibma:Illumination`, which is a subclass of E26 Physical Feature

('This class comprises identifiable features that are physically attached in an integral way to particular physical objects"). The following RDF triple expresses this relationship.

```
:c a bibma:Component .
   :i a bibma:Illumination .
   :c crm:P56_bears_feature :i .
```

This is a shortcut for the fully developed path:

```
:folio a crm:E53_Place .
   :c a bibma:Component ;
   crm:P59_has_section :folio .
   :i a bibma:Illumination ;
   crm:P53_has_former_or_current_location :folio .
```

Instances of E53 Place are a folio or a particular zone on a folio, for example.

The ontology interacts with a thesaurus of technical terms used in medieval studies (codicology, palaeography, iconography etc.) and descriptors used for indexing medieval illuminations in the project's databases. The data is structured in a thesaurus compliant with the international standard for thesauri and interoperability with other vocabularies (ISO 25964). The different lexical and semantic relationships that can be defined between the descriptors will have an intrinsic (semantic) role, in that they will help to show the relationships of hyponymy or synonymy, as well as an extrinsic (technical) function for the search engine. In addition, the project's data on people, corporate bodies, places and titles are aligned with existing authority files and linked data repositories, such as Rameau, VIAF, and GeoNames.

1.2 The Historical Dimension of Biblissima's Data

The majority of Biblissima's databases contain descriptive and structural metadata for medieval manuscripts and early printed books, issued from the cataloguing of these documents or scientific research, but the project also includes digital editions in TEI P5 of library inventories and texts and records on illuminations. Metadata like the date of creation and place of origin of a manuscript and its illuminations, the identification of the scribe, translator or commentator of the copied text, former owners (people and corporate bodies) of a manuscript throughout the centuries and lists of books kept in libraries at a given moment in time can be used to study the history of the texts and manuscripts, as well as reading and collecting practices.

In order to develop the portal step by step we have chosen to begin by creating a unified access point to two iconographic databases.

2 Objectives of the Biblissima Prototype

Using Semantic Web technologies, the Biblissima prototype aims to demonstrate the potential of the available metadata produced by the Biblissima project. It was developed using open source solutions and all the data is publicly available under an open licence in order to facilitate reuse.

The prototype is built on a subset of two iconographic databases: *Mandragore* [6], the database of the Department of Manuscripts of the National Library of France (BnF) and *Initiale* [7], the database of the IRHT.

It provides federated access to a subset of data present in the two databases, such as illumination related data: caption, descriptor, folio carrying the illumination, illumination record, digital surrogate of the illumination, artist, context of the illumination (author and title of the textual work per artistic unit), date of origin and place of origin. The data set also contains manuscript related data such as shelfmark, common name, grouping, repository, digital surrogate, manuscript record.

The subset is limited, in the case of both databases, to records on illuminations indexed with at least one geographical descriptor, which equates to almost 5 000 descriptors for approximately 20 000 illuminations.

A SPARQL endpoint and a federated search engine make it possible to search all the data in the cluster. Users can search by descriptor, artist, date of origin, place of origin, author or work title, and can refine their search with a series of facet filters. The results are displayed in a user-friendly manner by grouping them in lists. Pages about manuscripts, their units and illuminations include frames that display the corresponding digital surrogate using *IIIF manifests* [8], relating text and images. Other visualisation features are available to the user, such as timelines and maps. The data from *Initiale* and *Mandragore* are augmented with data on the digital surrogates of illuminations in their context, extracted from other manuscript catalogues (*Medium* [9] - the IRHT manuscript repository, Gallica - the digital library of the BnF, and BnF archives et manuscrits [10] - the catalogue of the Department of Manuscripts of the National Library of France). Each illumination record in the prototype links back to the original record from one of the two databases as well as to the full digitisation of the manuscript when available. The search capabilities currently do not include manuscript genealogies. This can be achieved by including more databases and classes like `bibma:Type_of_Use_Manuscript` and `bibma:Source` in the future, when implementing solutions deriving from lessons learned about texts and their transmission.

Both databases have been used to index manuscript illuminations for the last 25 years and different systems were chosen for structuring the descriptors. A polyhierarchical classification of Biblissima's thesaurus may make it possible to retain the original descriptor classifications while reordering them in a new systematisation. However, these classifications do not reflect the medieval practices of organising knowledge. The identification of the iconographic elements is sometimes based on internal information such as heading titles, chapter titles, inscriptions or notes present in the manuscript; in the absence of this kind of information, identification is dependent on the scholar's culture, especially in the case of living things and artefacts. This means that when using these descriptors to study medieval illuminations, we must keep in mind that their identification has a disparate chronological and cultural origin. Another feature of the indexing practices specific to these databases is that the data does not provide co-textual information, and the contextual information is not always available. This makes quite difficult to trace the diachronic evolution of the meaning of a word and of its iconographic representation.

The technical solutions adopted by Biblissima open new avenues and yield new ways of searching through data that could contribute to the analysis of iconographic representations. On the basis of the geographic descriptors, one could attempt to answer several questions regarding the status of cities in artistic imagery and define the notion of the city through iconographic choices: what are the criteria which confer an urban identity to a community and what makes the difference between an urban and a rural environment? From what point in time do cities begin to be represented and what cities are the most represented over the centuries? How could one explain the cases of single occurrence? Could one analyse the anachronistic representations of places, be they cities or battlefields? What are the most common descriptors associated with toponyms?

3 Conclusion

The semantic web solutions that Biblissima has chosen could be adapted in order to provide answers to other kinds of research topics. As such, Biblissima's polyhierarchical thesaurus makes it possible to establish new classifications of the descriptors that already exist in the databases and to recreate a medieval taxonomy of living species as it was conceived by an encyclopedist or a physician, for example. One might also connect the thesaurus to the digital edition of exegetical texts such as the biblical Glossa [11], one of Biblissima's partner projects, and try to study the semantic relations between the four senses of the Scripture (historical, allegorical, tropological and anagogical) and the iconographic representation of the biblical words and scenes, for example.

By adopting common standards for the ontology and for the thesaurus, Biblissima's data might also be aggregated with and used by other semantic web projects in the future.

References

1. Investissements d'avenir (CGI), ANR-11-EQPX-0007 `http://investissement-avenir.gouvernement.fr`
2. CIDOC CRM 6.0, `http://cidoc-crm.org/docs/cidoc_crm_version_6.0.pdf`

3. FRBRoo 2.2, http://www.ifla.org/files/assets/cataloguing/frbr/frbroo_v2.2.pdf
4. Abrami, G., Freiberg, M., Warner, P.: Managing and Annotating Historical Multimodal Corpora with the eHumanities Desktop - An outline of the current state of the LOEWE project Illustrations of Goethe's Faust. In: Historical Corpora, pp. 353 – 363. Narr Francke Attempto, Tübingen (2015)
5. Gkadolou, E., Stefanakis, E.: A formal ontology for historical maps, http://galaxy.hua.gr/~heraclitus/images/gkadolou/3_Gkadolou_ICA.pdf (2013)
6. Mandragore, http://mandragore.bnf.fr
7. Initiale, http://initiale.irht.cnrs.fr/accueil/index.php
8. IIIF, http://iiif.io/
9. Medium, http://medium.irht.cnrs.fr/
10. BnF, Archives et Manuscrits, http://archivesetmanuscrits.bnf.fr
11. Biblical Glossa, http://www.glossae.net/

Semantic Web for BIBLIMOS (position paper)

BÉATRICE BOUCHOU MARKHOFF[1], SOPHIE CARATINI[2], FRANCESCO COREALE[2],
MOHAMED LAMINE DIAKITÉ[3] and ADEL GHAMNIA[1]

[1] Université François Rabelais Tours - Laboratoire d'Informatique LI (EA 6300)
beatrice.bouchou@univ-tours.fr, adel.ghamnia@univ-tours.fr
[2] Université François Rabelais Tours - Laboratoire CITERES (UMR 7324)- EMAM team
sophie.caratini@univ-tours.fr, francesco.coreale@univ-tours.fr
[3] Université des Sciences, de Technologie et de Médecine - DMI, Nouakchott, Mauritanie
diakite@ustm.mr

Abstract. We present the BIBLIMOS project, which aims to address the Western Saharan culture and history, by considering both local ancient Arabic manuscripts and European colonial archives. We describe the project's context and objectives before focusing on ancient Mauritanian manuscripts, the content of which covers many scientific fields. We assess the current state of such ancient manuscripts' digital processing and we analyse what the semantic web can bring for their use by scholars, from North and South: the ability for *applications* to operate jointly on several distributed and heterogeneous sources of digitized manuscripts and other kinds of archives, to support collaborative reflection.

Keywords: Ancient Arabic Manuscripts; Data Integration; Semantic Virtual Infrastructure; Western Saharan Cultural Heritage

1 Introduction

BIBLIMOS is a long standing programme, led by the CITERES laboratory[4], that proposes to collect information, and facilitate the constitution of thematic corpora, from public and private archives pertaining to the history of the Western Saharan region. Its first goal was to provide to local students and researchers the ability to study their history, through a digital remote access to original materials (through images and descriptions), and also the ability to collaborate more easily with foreign teams, on these materials. Moreover, in the long run, it is also planned to deal with both primary sources (original material created at the time under study) and secondary sources (material written by scholars). In parallel, it is intended to address colonial archives about this geographical area, from European countries (mainly France and Spain), in order to cross complementary points of views, and thus, to discover new knowledge.

Involving an international and cross-disciplinary team of researchers in the humanities and, more recently, in computer science, BIBLIMOS aims to renew the knowledge and analysis of Western Sahara's societies, by making available to researchers from the North and the South an open and interactive tool for searching and comparing local

[4] http://international.univ-tours.fr/centre-for-cities-territories-environment-and-societies-citeres--283347.kjsp?RH=INTER

archive funds, including the manuscripts of the desert, and European archives related to these regions. There is also an important multilingual challenge, as we plan to perform cross-referencing of Arabic, Pulaar, Soninke, Wolof, French, Spanish, Portuguese, Italian, Dutch, German, English sources relating to the political, military, economic, legal, social, scientific and religious history of the territories of the Western Saharan region, from the modern era to the end of the Cold War.

Concerning computer science, the BIBLIMOS programme is just getting started: it aims to create an e-infrastructure based on a network of information around the history of the Western Sahara. This open tool will offer (i) an access to sets of archival sources and original manuscripts, (ii) a guide to navigate this knowledge network, (iii) an automatic registration of new sources and (iv) new tools for knowledge creation and visualizations. It will also be interfaced with various useful existing applications for research, such as electronic publishing platforms, collaborative editing tools, bibliography management tools, etc. To achieve this goal, three lines of work have been initiated. First, to instigate, assist and sustain the creation of quality digital resources from the original sources, second, to develop partnerships with providers of already existing digital resources, and third, to incrementally build the target distributed e-infrastructure, including a web portal as mediator, relying on semantic web resources and technologies.

In the first line of work, BIBLIMOS stakeholders in Social Sciences and Humanities (SSH) are engaged in actions aimed at discovering new local sources and convincing their owners to join the programme. Concerning the second line of work, today manuscript sources concealed in the Western Sahara are already partly inventoried, and many European archive funds are now available to the public. As shown in Table 1, on the one hand, online digitized full-text manuscripts exist, duly indexed and catalogued, and on the other hand, institutions or associations offer to collaborate in order to index digitized materials from many sources (cf. last lines in Table 1). Clearly *the Web*, that provides information *exploitable by humans*, well supports all those very useful initiatives. However, the query, the analysis, the combination and the overlapping of these multiple funds, still represents a major challenge for every interested person. This paper is dedicated to the third line of work in the BIBLIMOS programme, which addresses the field of the automatic data-processing of such sources, in order to better assist humans in these tasks. This is a field in which almost everything has to be designed and built. *The Semantic Web*, i.e., *the web knowledge exploitable automatically by computers*, is the way to cope with these challenges, as we argue in Section 3, after having presented the state of the art of digital processing of Ancient Arabic Manuscripts in Section 2.

2 Digital processing of Mauritanian Ancient Arabic Manuscripts

2.1 Mauritanian Ancient Arabic Manuscripts

We focus on Mauritania's manuscripts because Sophie Caratini, the instigator of the BIBLIMOS programme, is an anthropologist specialist of Mauritania and she built strong collaborations with scholars in Nouakchott, in particular through the IMRS[5].

[5] Institut Mauritanien de la Recherche Scientifique, see http://www.imrs.mr/spip.php?page=sommaire_fr

Site	Description
http: //www.westafricanmanuscripts.org/	**University of Illinois, Urbana-Champaign**. Online catalogue, references about 22500 manuscripts from eleven different collections, including Northwestern Univ.
http://digital.library.northwestern.edu/arbmss/index.html	**Northwestern University, Chicago**. Online catalogue, entries from four separate collections.
http://memory.loc.gov/intldl/malihtml/malihome.html	**Library of Congress**. Online catalogue, with access to images of 32 manuscripts from Timbuktu, Mali.
http://gallica.bnf.fr/	**French National Library (BnF)**. Online access to 35 manuscripts from Timbuktu, Mali.
http://www.tombouctoumanuscripts.org	**University of Cape Town**. Tombouctou Manuscripts Project; access to primary sources upon registration.
http://omar.ub.uni-freiburg.de/	**Universities of Freiburg and Tübingen** (Germany). Online images of approx. 2.500 Arabic manuscripts (134.000 images) from Mauritania, with bibliographical metadata.
http://wamcp.bibalex.org/	**Bibliotheca Alexandrina** (Egypt). Online collection of Arabic manuscripts related to classical medicine, around 1000 books and fragments.
http://www.qdl.qa/en	**Qatar Digital Library** (with the British Library). Archives, maps, manuscripts, sound recordings, photographs with explanatory notes and links, in both English and Arabic.
makrim.org	**IMRS** (Mauritanian Islamic Republic). Catalog of Mauritanian manuscripts, in both French and Arabic.
http://www.islamicmanuscript.org/extresources/manuscriptcatalogues.aspx	**The Islamic Manuscript Association** (Cambridge, stakeholders from 25 countries). List of Islamic manuscripts catalogues.
http://openlibrary.org/	**Open Library** (world wide **open access** project). List of resources on Arabic manuscripts (catalogues, books, etc.).
http://www.archive.org/	**The Internet Archive** (USA non profit association). A search on *Arabic manuscripts* gives some digitized books.

Table 1: Web sites about Western Saharan, or more generally, Arabic manuscripts.

Mauritania is known [...] for its enormously rich heritage of Arab manuscripts, many brought from the Arab East by pilgrims returning from Makkah, some recopied from those imported sources by students in the Qur'an schools [...], and others composed by Mauritania's own jurists, poets and historians[6] [16]. According to researchers, some Mauritanian manuscripts were written as early as in the 10^{th} century, and their forms and subjects are very diverse, including law, science and religion. To have access to this legacy, the first step is to build up a precise survey of all manuscript repositories in existence in the territories of the Western Saharan region. This has been the goal of long term projects: for instance, the West African Arabic Manuscripts Database Project, from the University of Illinois at Urbana-Champaign, started in 1987, provides a catalogue (first line of Table 1) that references more than two thousand manuscripts. Currently, it references eleven collections, which still is far from representing the actual reality of family libraries. This is one of the web resources we plan to exploit in the BIBLIMOS programme, in parallel of completing the repositories survey work performed by the SSH teams. Several other websites provide information on Western Saharan or, more generally, on Arabic manuscripts: the list presented in Table 1 shows that there is already a lot of knowledge available on the web, but this knowledge still is exploitable only through human labour.

2.2 Digital Processing of Ancient Arabic Manuscripts

Concerning manuscripts, many different descriptions may be stored in computer memories: (i) seeing the manuscript as an archaeological object, i.e. starting from its external aspect, a set of features may be evaluated, for instance the material it is made with, the colour of ink, etc. This is called codicology [4] and a well-established vocabulary for such a set of descriptors is provided by the IRHT[7]; (ii) a numerical image of the manuscript can be taken; (iii) a transcription of the manuscript's textual content can be created, either manually or automatically from its numerical image (with OCR tools); (iv) both the image and the transcription may be annotated, this is the case for many European manuscripts, whose textual contents are encoded using the TEI standard; (v) the manuscript can be catalogued, i.e. classified and described by librarians or archivists, so it could be found again among collections: this supposes to define and identify descriptors, including the location, and some general information about the content.

For each of these descriptions, active research is conducted and, in some cases, they converge to well established standards. Specifically for ancient Arabic manuscripts, in [15] the authors present the problem of cataloguing, assessing the difficulties involved in identifying the metadata used by different schools (those dealing with specimen and those addressing whole volumes). The solution proposed for enhancing interoperability is to rely on the DCMI[8] vocabulary. The TEI[9], aimed at helping libraries, publishers, museums and universities to encode texts in order to facilitate information retrieval from

[6] http://www.saudiaramcoworld.com/issue/200306/mauritania.s.manuscripts.htm

[7] Institut de Recherche et d'Histoire de Textes, see http://codicologia.irht.cnrs.fr

[8] World widely used, simple and generic, digitized resources' description, see http://dublincore.org/

[9] Text Encoding Initiative: http://www.tei-c.org/index.xml

textual contents, is another important medium for interoperability [14]. Nevertheless we cannot hope to use it in the short term because for now the only way *to get transcriptions of Mauritanian manuscripts* is to manually enter the text. Indeed, automatic character recognition algorithms hardly apply to these kinds of manuscripts, written with Arabic graphemes but very often actually in many other languages (e.g. Pulaar, Wolof, etc.). In [1], the authors recall the existing difficulties for applying OCR to ancient Arabic manuscripts and, although recent advances are reported in [3] and [11], they need to be further developed. Manuscript image analysis is not reduced to OCR: for instance, word spotting may be a useful alternative to character recognition. This is why several works propose to build ontological descriptions (or sets of metadata) of graphical image features, in order to index and retrieve manuscripts' digital images on this descriptive basis [7, 6]. But to the best of our knowledge, such proposals have never been applied to ancient Arabic manuscripts.

When it comes to ontological representation of ancient manuscripts, the work described in [10], about the SAWS[10] project (Sharing Ancient WisdomS), is clearly an example of what we target in the BIBLIMOS framework. It deals with collections of moral and social advice and/or philosophical ideas from Greek and Arab wisdom literatures. Many of the concerned manuscripts have been transcribed and annotated using TEI, and an extension of the FRBRoo ontology [9] has been developed to describe the transmission of information (from one copyist to another and from one language to another). The authors extract the relationships defined in the ontology from the TEI annotations, to generate a conceptual network expressed in RDF[11]. This network allows researchers to explore links between the different documents' contents. This is an example of how semantic web technologies contribute to the building of new means of knowledge, by opening up and linking various sources for research which would otherwise remain isolated and unused.

3 Semantic Web Architecture for BIBLIMOS

For humans, carrying out some scientific work by using the resources listed in Table 1 is still difficult, as there are no means to perform cross-references, comparisons, or to analyse the different points of view they provide, etc. Regarding BIBLIMOS' aims, other kinds of sources than manuscripts (e.g. European archives) should also be exploited, which increases again these difficulties. Fortunately, while *the web* allowed sources' owners (or depositaries) to publish their resources through websites, *the semantic web* now supports the development of softwares that help humans to cope with these difficulties. Indeed, the semantic web is a network of semantic representations of web-published information that relies on the same technical principles as the websites' network, but allows *programs* to operate on data at this semantic level. Main semantic web concepts are (i) web ontologies and (ii) linked (open) data; they provide *a global space of interoperability*, thus they are important components for BIBLIMOS' aims.

Figure 1 illustrates the intended general architecture for the BIBLIMOS programme. The novelties brought by the semantic web obviously start at the DATA level: to benefit

[10] `http://www.ancientwisdoms.ac.uk/`

[11] Data model standard: `http://www.w3.org/RDF/`

from these novelties, beyond all the work that has to be done to obtain results presented in the previous section, digital sources should also be *pushed up to the semantic level*. To this aim, the sources' concepts and their relationships must be specified, from the bottom-up (starting from the source contents), top-down (from already well defined consensual ontologies), or both. The source's content should be related to this conceptual level, which may be done by using tools called *Mapping Frameworks* in Figure 1. Some of those tools propose to export the source data into a set of RDF triples (the standard data warehouse approach in data integration systems), and some of them propose to access data through the conceptual level, based on the ontology-based data access (OBDA) principles [5] (the mediation approach, which is provided by, e.g., *ontop*[12]). Whatever the chosen approach, the source's content is then searchable at the semantic level, with SPARQL. Those contents may be combined using reference thesauri and ontologies.

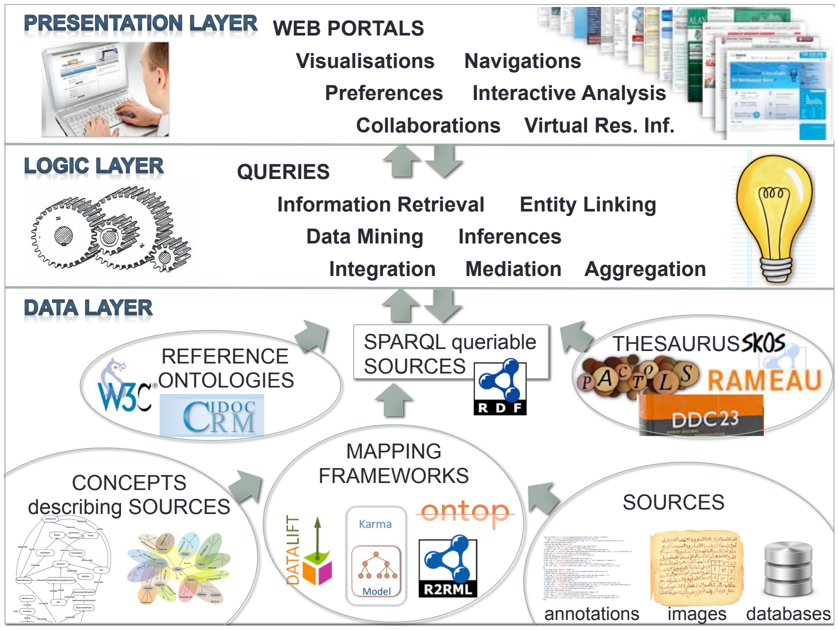


Fig. 1: Global BIBLIMOS' Virtual Infrastructure.

Querying the semantic web through its linked data sets is still in its infancy. Public well-established reference knowledge resources play the important role of hubs in this linked data network. The most visible are resources of facts, e.g. DBpedia, but at the conceptual level, reference domain ontologies also act as fundamental integration means. This is the case for CIDOC CRM [13] for cultural heritage, with its extension

[12] http://ontop.inf.unibz.it/

FRBRoo for libraries. These reference domain ontologies are the product of a long, international collaborative work, reflecting a consensus among the domain experts. These distributed and collaborative dimensions of the web are naturally inherited by the semantic web. In the context of BIBLIMOS, this is extremely powerful because these two features mirror the local structural organization of the Mauritanian family libraries, open to communities but distributed in the country rather than centralized in only one authoritative place.

The semantic web resources also promote *multilingualism*, particularly in vocabulary resources such as *thesaurus*, as evidenced by multilingual ones, e.g. VIAF[13] or RAMEAU[14], the French national library thesaurus now accessible on the semantic web (in SKOS), which is fully interlinked with a German (SWD) and an American (LCSH) thesaurus (thanks to the *Multilingual ACcess to Subjects* project).

Above the DATA layer is the LOGIC layer, in which all the well-known successful inventions in the field of data operation (some of which are listed in Figure 1) may be revisited to take into account the semantic dimension of data. A corner stone for most of them is to access multiple sources conjointly, which supposes interoperability: one of the solutions provided by the semantic web is to align the local lightweight ontologies that describe the sources' content to the reference ontologies, allowing mediator systems to aggregate local data sets, for instance following the principles described in [12, 2]. Very active researches are conducted in the semantic web community to develop this LOGIC level, based on efforts to produce a strong semantic data layer. Lastly comes the PRESENTATION layer, whose innovative potential is also greatly boosted by the possibilities issued from the semantic web.

4 Conclusion

We first drew a state of the art concerning the ways ancient Arabic manuscripts are processed and made available to the public nowadays, considering that the picture is not so different in the area of European archives (except that OCR tools are more usable). Once digitized, sources must be pushed up to the semantic level, for the query, the analysis, the combination and the intersection of these multiple funds to be supported by automatic data-processing of sources. We presented the semantic-web Virtual Infrastructure designed to cope with these challenges within the BIBLIMOS programme.

We are aware that BIBLIMOS is a very ambitious programme - we are not aware of the existence of a similar enterprise anywhere else - as semantic web applications in this field are just beginning to emerge. For now, agreements are signed between our universities (Tours and Nouakchott), both in the computer science side and the social science side.AFD[15] currently funds a training campaign for librarians of the IMRS[16] on cataloguing documents, and the Mauritanian government is going to support all the needed local actions. Concerning the semantic web level, we are building an ontology

[13] Virtual International name Authority File: `http://viaf.org/`

[14] `http://data.bnf.fr/en/semanticweb`

[15] Agence Française de Développement: `http://www.afd.fr/lang/en/home`

[16] Institut Mauritanien de recherches scientifiques: `http://www.imrs.mr/spip.php?page=sommaire_fr`

for the IMRS' manuscripts [8], a part of which is already digitized, and we plan to work on designing and building an annotation tool based on this ontology. In order to include the European side (archives on these countries), we are thinking about a MSC Action (deadline in january, 2016). The campaign of partnerships with already existing materials is still to be done, as we must first build the semantic web tools that we should propose to them.

References

1. Abdel Belaïd and Nazih Ouwayed. Segmentation of ancient arabic documents. *Guide to OCR for Arabic Scripts*, pages 2–16, 2011.
2. Beatrice Bouchou and Cheikh Niang. Semantic mediator querying. In *International Database Engineering and Applications Symposium (IDEAS)*, pages 29–38. ACM, 2014.
3. W. Boussellaa, A. Zahour, H. El Abed, A. Benabdelhafid, and A. Alimi. Unsupervised block covering analysis for text-line segmentation of arabic ancient handwritten document images. In *20th International Conference on Pattern Recognition (ICPR)*, pages 1929–1932, 2010.
4. Stefanie Brinkmann and Beate Wiesmüller, editors. *From Codicology to Technology: Islamic Manuscripts and Their Place in Scholarship*. Frank and Timme GmbH, 2009.
5. Diego Calvanese, Giuseppe De Giacomo, Domenico Lembo, Maurizio Lenzerini, Antonella Poggi, Mariano Rodriguez-Muro, Riccardo Rosati, Marco Ruzzi, and Domenico Fabio Savo. The MASTRO system for ontology-based data access. *Semantic Web*, 2(1):43–53, 2011.
6. M. Coustaty, R. Pareti, N. Vincent, and J.M. Ogier. Towards historical document indexing: extraction of drop cap letters. *IJDAR*, 14(3):243–254, 2011.
7. B. Coüasnonet, J. Camillerapp, and I. Leplumey. Access by content to handwritten archive documents: Generic document recognition method and platform for annotations. *IJDAR*, 9(2):223–242, 2007.
8. Mohamed Lamine Diakité and Beatrice Bouchou Markhoff. OMOS: Ontology for Western Saharan Manuscripts. Technical Report 313, Université François Rabelais Tours, Laboratoire d'Informatique (available in HAL: https://hal.archives-ouvertes.fr/hal-01134010), 2015.
9. Martin Doerr and Patrick Le Boeuf. Modelling intellectual processes: The frbr - crm harmonization. In *Digital Libraries: Research and Development, volume 4877 of Lecture Notes in Computer Science*, pages 114–123. Springer, Berlin / Heidelberg, 2007.
10. A. Jordanous, K. F. Lawrence, M. Hedges, and C. Tupman. Exploring manuscripts: Sharing ancient wisdoms across the semantic web. In *2nd International Conference on Web Intelligence, Mining and Semantics (WIMS)*, pages 678–683. ACM, New York, 2012.
11. A. Khemiri, A. Kacem, and Belaid A. Towards arabic handwritten word recognition via probabilistic graphical models. In *Frontiers in Handwriting Recognition (ICFHR)*, pages 678–683, 2014.
12. Cheikh Niang, Béatrice Bouchou, Yacine Sam, and Moussa Lo. A Semi-Automatic approach For Global-Schema Construction in Data Integration Systems. *IJARAS*, 4(2):35–53, 2013.
13. Dominic Oldman. *The CIDOC Conceptual Reference Model (CIDOC-CRM): A Primer, Version 1*. CIDOC CRM (http://www.cidoc-crm.org/docs/CRMPrimer_v1.1.pdf), 2014.
14. Desmond Schmidt. Towards an interoperable digital scholarly edition. *Journal of the Text Encoding Initiative [http://jtei.revues.org/979]*, 7, 2014.
15. M. O. Soulah and M. Hassoun. Which metadata for ancient arabic manuscripts cataloguing? In *International Conference on Dublin Core and Metadata Applications, The Hague, Netherlands*, 2011.
16. L. Werner. Mauritania's manuscripts. *Saudi Aramco World*, 54(6):2–16, 2003.

Semiotic Issues and Perspectives on Modeling Cultural Artifacts

Revisiting 1970's French criticisms on 'New archaeologies'

Aurélien Bénel

ICD/Tech-CICO, Troyes University of Technology (France)
aurelien.benel@utt.fr

Abstract. This paper looks back at 1970's modeling initiatives in archaeology in order to draw parallels with current initiatives on applying Semantic Web techniques to cultural artifacts. At those times, epistemological criticisms were raised on the lack of consideration by these models of the semiotic value of cultural artifacts. Based on these arguments, we propose several design perspectives for computer models and tools to aim at human semiotics rather than formal semantics. Last, those models and tools are seen in action as well as how art historians make sense of them.

1 Introduction

"Contrary to natural sciences, human sciences formalize an already formalized object"[1] (J. Gagnepain, as quoted in Bruneau, 1976). As an example, "archeologists are not the first ones to describe or classify artifacts"[2] (Bruneau, 1976). Indeed, the people of ancient times, as designers or users, had already their own theory of their technical universe.

These statements on the very nature of human sciences in general and archaeology in particular were published in the 1970's, a time when archaeologists and art historians wanted to modernize their disciplines by using 'models' (inspired from other disciplines) and large 'databanks' to store the 'graph of facts' formalized with a 'universal documentary language'. As stimulating as the modeling initiatives could have been, they were identified by critics as leading to an epistemological dead-end.

To begin with, we will look back at 1970's modeling initiatives in archaeology in order to draw parallels with current initiatives on applying Semantic Web techniques to cultural artifacts. Then, by studying epistemological criticisms raised at those times, we will see how those models failed at taking into account the semiotic value of these objects. Based on these arguments, we will propose

[1] "À la différence des sciences de la nature, les sciences humaines formalisent un objet déjà formalisé".

[2] "Les archéologues ne sont pas les premiers à décrire ou à classer le matériel dont ils traitent".

several design perspectives for computer models and tools to aim at human semiotics rather than formal semantics. Last but not least, we will see those models and tools in action, and how art historians make sense of them.

2 Back to the future

In the field of archaeological knowledge modeling, 1972 was a decisive year with two major collective works published: *Models in archeology* (Clarke, 1972), and *Les banques de données archéologiques* (Borillo & Gardin, 1972). In both books, number of authors proposed to record the description of artifacts as well as their relationships in space or time, using statistical but also logical models on computers. Notably, set theory was adopted (Litvak King & García Moll, 1972) to formalize artifacts *taxonomies* (*e.g.* Every *kantharos* is a *wine vase*) and spatial *meronomies* (*e.g. Paestum* is a part of *Magna Graecia*).

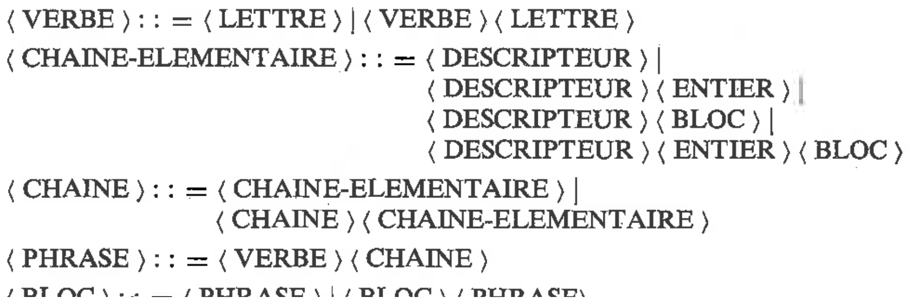
For Semantic Web researchers, the more interesting works of these times are probably those that used 'SATIN 1' (Chouraqui, 1972), a system first designed for the French general inventory of cultural heritage. It was composed of an *analysis language* to represent artifacts descriptions, and a *query language* to retrieve or aggregate those descriptions.

As RDF today, SATIN 1 analysis language (see Fig. 1) was expressive enough to tackle with complex descriptions. For example, figure 2 shows the formal description of a small (25 × 15 *mm*) amygdaloidal object made of carnelian, found in Vaphio, dated from Late Helladic II and depicting a man on a chariot leading two horses (Ginouvès & Guimier-Sorbets, 1978).

Similarly to what is done nowadays in Web ontologies, every *descriptor* ('LENGTH', 'HORSE', 'WHEELS', 'LEAD') had to be defined in a *domain* (material, finding location, description, etc.), and in several domains, the lexicon could be hierarchically structured ('LACONIA / VAPHIO', 'STONE / CARNELIAN', 'LATE HELLADIC / LHII').

In a very contemporary way, SATIN 1 inventor pointed out that because descriptors from different domains can be mixed in the same description, the addition of new descriptors (*e.g.* related to decor) can be done in several ways: either adding it to every impacted domains (*e.g.* sculpture, furniture, architecture, etc.) or creating a new domain usable on any kind of objects (Chouraqui, 1972). Beyond the formal benefit of combinatory expressivity, the world-wide reuse of domains and descriptions was advocated by one of the promoter of these projects as a 'necessity' and a 'duty', to go from an 'egoist and closed possession' of information to a 'common good' (Ginouvès, 1972).

A nice illustration of this trend was provided in 1975 by Anne-Marie Guimier-Sorbets in her thesis on the analysis and formalization of geometric ornaments in Greco-Roman mosaics for automatic processing. In a very formal and logical way, she defined every attribute she used, and described the process one should follow to set the right value to the right segment of the artifact. Philippe Bruneau, who was on the examining board of this iconic thesis, wrote a subsequent article



```
⟨ VERBE ⟩ : : = ⟨ LETTRE ⟩ | ⟨ VERBE ⟩ ⟨ LETTRE ⟩
⟨ CHAINE-ELEMENTAIRE ⟩ : : = ⟨ DESCRIPTEUR ⟩ |
                             ⟨ DESCRIPTEUR ⟩ ⟨ ENTIER ⟩ |
                             ⟨ DESCRIPTEUR ⟩ ⟨ BLOC ⟩ |
                             ⟨ DESCRIPTEUR ⟩ ⟨ ENTIER ⟩ ⟨ BLOC ⟩
⟨ CHAINE ⟩ : : = ⟨ CHAINE-ELEMENTAIRE ⟩ |
                 ⟨ CHAINE ⟩ ⟨ CHAINE-ELEMENTAIRE ⟩
⟨ PHRASE ⟩ : : = ⟨ VERBE ⟩ ⟨ CHAINE ⟩
⟨ BLOC ⟩ : : = ⟨ PHRASE ⟩ | ⟨ BLOC ⟩ ⟨ PHRASE⟩
```

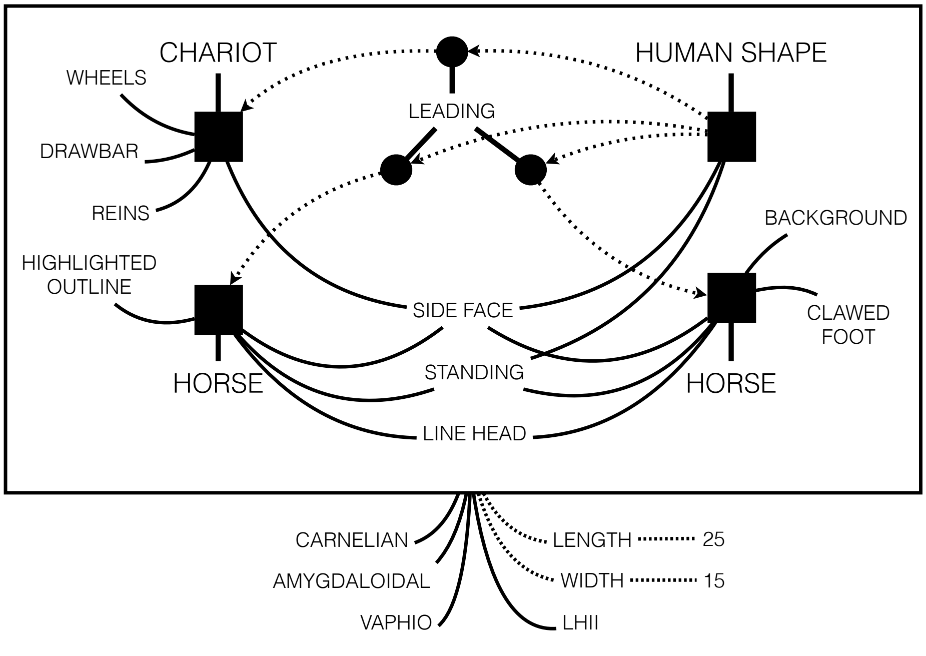


Fig. 2. Modeling a Creto-mycenian seal and its iconography with SATIN 1 (Ginouvès & Guimier-Sorbets, 1978)

(Bruneau, 1976), which spawned an unprecedented polemic in the refined world of the French School of Archaeology in Athens.

The director of publication felt the need to write a prologue (Amandry, 1977) in the following issue of the journal to give the "definition of what is and what is not [the journal]",[3] stating that "The journal does not seem to be an appropriate place for doctrinal lectures or handbooks of methods".[4] And even fourteen years later, the author of the thesis felt the need to reply to Philippe Bruneau's arguments in the introduction of her own handbook (Guimier-Sorbets, 1990).

3 Criticisms from the past

Philippe Bruneau's criticisms on the way cultural artifacts are modeled was focused on the notion of 'descriptor' (i.e. the element of an ontology – class, individual, property, etc.).

He argued, first, that descriptors such as 'foreground' or 'background' are usually chosen in order to be universal, independent of era and geography, which should be itself quite surprising in a historical science.

Secondly, he asked, what could be the validity of 'foreground' and 'background' in a case like Greek frets, where every black fret on white has a complementary white fret on black. Anne-Marie Guimier-Sorbets answered: "by convention, the fret to be analyzed is the outer one of the mosaic. The other complementary part is analyzed as background". Philippe Bruneau noted that it was a shame to decide *by convention* that the whole description would be from the border to the centre whereas mosaics were built from the central panel to the border.

Lastly, the very term 'descriptor', connoting agency, would lead one to think that it is not the archaeologist but the device that describes an artifact. And to forget the archaeologist as the describer leads one to forget that the first to describe and classify the artefact was indeed the ancient user himself.

Beyond this sole example and even beyond modeling issues, Philippe Bruneau tried to formulate the very nature of archaeology as a human science. As an *arte factum* (*i.e.* done by human skills), the artifact is indeed a semiotic object. As the two sides of Saussure's sign, one cannot split its material *configuration* from the *program* assigned by its designer (and by its users too).

Therefore, contrary to a common misconception, it would be meaningless to describe it factually first and to interpret it later. Moreover, as a semiotic object, its meaning depends on the other objects in the surrounding neighborhood. In a normal desktop setting, the important feature of a pen is that its writings can be erased contrary to a pencil's. But in the absence of a pencil its main feature would be that you can write with it. And in the absence of a pipe tool, its main feature would be its form. You cannot say anything of a pencil, neither its use nor

[3] "la définition de ce qu'est et de ce que n'est pas le *Bulletin de correspondance hellénique*".

[4] "[Le *Bulletin*] ne paraît pas être un endroit approprié pour des exposés de doctrine ou des traités de méthode".

its features, without knowing anything about the *state of things*, the state of its technical context. For this reason, the description of an artifact is never finished: it will be revised and revised again in a spiral approach. As rationally structured as a linguistic universe, the state of things is although always idiomatic: it can in no way be universal. Furthermore, it would not occur to anyone to describe a foreign language without 'getting in' the system of its users.

4 From issues to perspectives

Though these semiotic objections highlight rough issues in formal semantic descriptions, they also bring very promising perspectives on how knowledge modeling could serve cultural artifacts sense making. First, instead of looking for the universality of the description language, the latter should be tightly tied to the coherent *state of things* it was created for. Because an artifact is part of an indefinite number of states of things, we should strive for maintaining the identity of the artifact in overlapping states of things and corresponding analyses. Second, instead of overfocusing on inferences based on out-of-context definitions (*type*, *subClassOf*, *partOf*), one should be able to browse the different states of things and see how a feature activate or deactivate others, both at the artifact level and at the fragment level, in other words to provide interactive multi-level co-occurences visualization. Here follows two examples of the use of our semiotic-centric tools and methods by art historians[5].

The first one deals with the iconography of Dionysos and banquets on vases from the area of Paestum (Italy). To do this, the team gathered more than 600 photographs about those vases from museums all over the world. Years after years, each master or PhD student has tried to make sense about a given puzzling feature (bearded/unbearded Dionysos, face and feet in different directions, etc.) laying out their analysis on the top of the other. Our tools and methods are especially suited to the case where the meaning a feature can be discovered through the co-occurrence with another (Bénel, 2006). For example, it appeared that Dionysos was bearded when he was depicted in presence of a kantharos (a vase used in rituals). The interpretation by the PhD student (Pouyadou, 2001) was that the beard was significant of the fact that Dionysos was the god receiving offerings rather than the character of mythological stories.

The second one relates to the typology and chronology of Iron Age vases discovered in the excavations of the cemetery of Athens called 'Kerameikos'. A recent monograph analyzed features of each vase, and then gathered them into new coherent stylistic groups. In order to review this research work, a professor used our software to model "how [the author] himself, classified it". Then, in order to initiate Master students to research, he asked each of them to analyze the stylistic features of one type of vases. On evaluation day, he asked them to "combine features to get groups as coherent as possible" (Bénel *et al.*, 2010).

[5] Pr. Jean-Marc Luce, his colleagues and students (PLH-CRATA Research team, University of Toulouse II, France).

Even if the analysis by the student was incomplete and perfectible, the vases she described as having a flat paunch ('panse plate') and a short lip ('lèvre courte') appeared to be exactly what the specialist considered to be the oldest group (see Fig. 3).

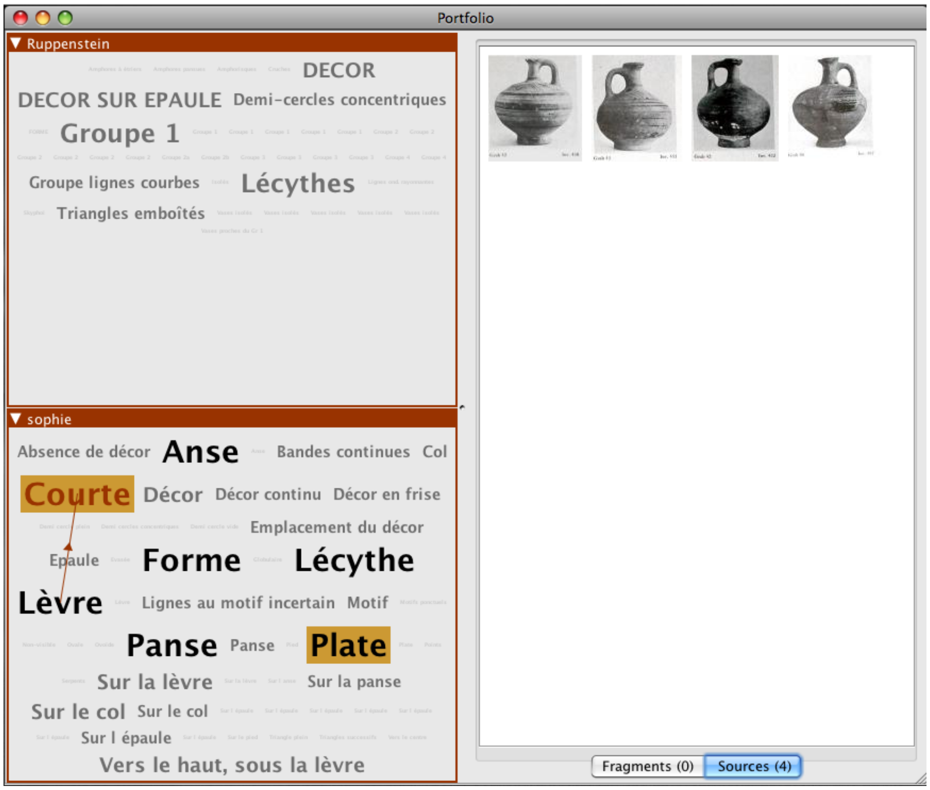


Fig. 3. "Do your stylistic features define a group?" (*Porphyry* screenshot)

5 Conclusion

In order not to do the same error twice, one should learn from errors in the past. In the 1970's, criticisms on modelling cultural artifacts were about the forgotten semiotic value of cultural artifacts, and therefore the tight link we should preserve between its definitions and its *states of things*, its contexts. We then illustrated how an interactive multi-level co-occurences visualization can be used by art historians to make sense of cultural artifacts.

References

Amandry, P.: Avant-propos. Bulletin de Correspondance Hellénique 101(1), 1–3 (1977)

Bénel, A.: Porphyry au pays des Paestans : Usages d'un outil d'analyse qualitative de documents par des étudiantes de maîtrise en iconographie grecque. Suppl. à Texto! 11(2), 173–179 (2006)

Bénel, A., Zhou, C., Cahier, J.-P.: Beyond Web 2.0... and beyond the Semantic Web. In: Randall, David; Salembier, Pascal (eds.). From CSCW to Web 2.0: European Developments in Collaborative Design, pp. 155–171. Springer Verlag (2010)

Borillo, M., Gardin, J.-C. (eds.): Les banques de données archéologiques. Marseille, 12-14 juin 1972. CNRS (1974)

Bruneau, Ph.: Quatre propos sur l'archéologie nouvelle. Bulletin de Correspondance Hellénique 100(1), 103–135 (1976)

Clarke, D.L. (ed.): Models in Archæology. Methuen, London (1972)

Chouraqui, E.: Le système d'exploitation automatique de l'inventaire général des monuments et richesses artistiques de la France (Formalisation du langage d'analyse). In: Borillo, M., Gardin, J.-C. (eds.): Les banques de données archéologiques. Marseille, 12-14 juin 1972, pp.147–159. CNRS (1974)

Ginouvès, R., Projets de banques de données archéologiques à l'Université de Paris X In: Borillo, M., Gardin, J.-C. (eds.): Les banques de données archéologiques. Marseille, 12-14 juin 1972, pp.221–226. CNRS (1974)

Ginouvès, R., Guimier-Sorbets, A.-M. La constitution des données en archéologie classique. CNRS (1978)

Guimier-Sorbets, A.-M.: Les bases de données en archéologie : Conception et mise en œuvre. CNRS (1990)

Litvak King, J., García Moll, R. Set Theory Models: An approach to taxonomic and locational relationships. In: Clarke, D.L. (ed.): Models in Archæology, pp. 735–755. Methuen, London (1972)

Pouyadou, V. Dionysos barbu : Le sens du poil. Pallas 57, 169–183 (2001)

SemanticHPST: Applying Semantic Web Principles and Technologies to the History and Philosophy of Science and Technology

Olivier Bruneau[1], Serge Garlatti[2], Muriel Guedj[3],
Sylvain Laubé[4], and Jean Lieber[5]

[1] University of Lorraine, LHSP-AHP, 91 avenue de la Libération, BP 454, F-54001 Nancy cedex, France
[2] Telecom-Bretagne, LabSTICC, CS 83818, F-29238 Brest Cedex 3, France
[3] University of Montpellier 2, LIRDEF, 2 place Marcel Godechot, BP 4152, F-34092 Montpellier Cedex 5, France
[4] University of Bretagne Occidentale, Centre François Viète (EA 1161), 20, rue Duquesne, CS 98 837, F-29 238 Brest Cedex 3, France
[5] University of Lorraine, LORIA, Campus scientifique, BP 239, F-54506 Vandoeuvre-lès-Nancy Cedex, France

Abstract SemanticHPST is a project in which interacts ICT (especially Semantic Web) with history and philosophy of science and technology (HPST). Main difficulties in HPST are the large diversity of sources and points of view and a large volume of data. So, HPST scholars need to use new tools devoted to digital humanities based on semantic web. To ensure a certain level of genericity, this project is initially based on three sub-projects: the first one to the port-arsenal of Brest, the second one is dedicated to the correspondence of Henri Poincaré and the third one to the concept of energy. The aim of this paper is to present the project, its issues and goals and the first results and objectives in the field of harvesting distributed corpora, in advanced search in HPST corpora. Finally, we want to point out some issues about epistemological aspects about this project.

Keywords: HPST (history and philosophy of science and technology), modern history, Semantic web, RDFS annotations, HPST ontologies, exact search, approximate search, harvesting distributed corpora, epistemology

1 Introduction

The application of computer science to research in history has existed for a long time [1],[2] though it can be noticed that the recent research domain of "Digital Humanities" (DH) is growing as result of a digital "revolution" at work that impacts the whole society at the international level. In France, tools and utilities dedicated to DH like the very large facility Huma-Num (http://www.huma-num.fr) have been created in order to favor "the coordination of the collective production of corpora of sources (scientific recommendations, technological best practices).” It also provides research teams in the human and social sciences with a range of utilities to facilitate the processing, access, storage and interoperability of various types of digital data.” The *Dacos and Mounier report* [3] shows that the French research is active, however the authors recommend the creation of “Centers of Digital Humanities”. The research network SemanticHPST is based on a strong coupling of laboratories in History and Philosophy of Science and Technology (HPST) and in Computer Science (LHSP–AHP, LORIA in Nancy) and (CFV, LabSTIIC in Brest) with research questions about the use of semantic web for HPST. The SemanticHPST project takes part in the emerging issues at the French and international levels in the domain of HPST.[1] Actually, the Semantic Web technology appears as efficient in order to generate tools adapted to the need of production and diffusion of distributed “intelligent digital” corpus in history [4].The objectives of the project are: (i) to integrate the existing technologies to manipulate digital contents of large volume by modeling knowledge as ontologies (annotation, request) for History and Philosophy of Science and Technology; (ii) to extent these technologies. The goal of this paper is to present the SemanticHPST project: its history, its objectives, the first results according to the information retrieval aspect and some epistemological issues. Because the methods in History of science and Technology are covering some elements of others domains in humanities (for example in history or in archeology), another goal of the SemanticHPST group is to share questions and results with the scientific community.

The paper is organized as follows. Section 2 presents the main goals of the SemanticHPST project and its three French HPST sub-projects for which semantic web technologies are useful. Section 3 presents some requirements and corresponding tools supporting different resource retrieval processes according to the researchers’ practices. Section 4 presents some issues from an epistemological viewpoint. Section 5 concludes the paper.

2 The SemanticHPST Project

In November 2010, the main topic of a European workshop was the uses of ICT and history of science and technology in education.[2] To improve research in HPST on one hand, and to promote dissemination of the HPST in the field of education on the other hand, some participants were convinced by the necessity to use new ICT tools [6], [7], [8], [9].

[1] See the 18^{th} session organised by some authors of this paper during the last meeting of SFHST (French society for history of science and technology), April 2014 (http://sfhst2014lyon.sciencesconf.org/resource/page/id/5), and the last meeting of the international consortium DigitalHPS at Nancy, September 2014, (http://dhps2014.sciencesconf.org).

[2] After this workshop, an extensive book written by participants and others has been published in 2012 [5].

In 2012, some historians of science and technology and computer scientists have created a consortium called SemanticHPST.[3]

The main goal of SemanticHPST project is to enrich the practices of researchers and communities in HPST. According to the specificity of the practice as historians of science, three main issues were tackled:

1. The management of large quantities of data especially for the most recent periods (XIXth, XXth centuries up to the present day). Knowing that the historical approach involves to integrate relevant elements from the context of production of these data into metadata.
2. The heterogeneity of sources and corpora constituted from these sources.
3. The production of new relevant digital corpora from several available digital historical collections.

To address our main goal and the three previous issues, our project is based on the Semantic Web principles and technologies. Thus, it has three main sub-goals: (i) Building intelligent digital corpora, that is to say corpora with primary and secondary sources having semantic metadata and their corresponding ontologies; (ii) Designing tools to access and enrich existing corpora and to create new ones; (iii) Evaluating the resulting practices and building an epistemological viewpoint about the use of TIC in HPST.

To achieve these goals, it is necessary to ensure a certain level of genericity for metadata, ontology, computer-based tools and practices.

To deal with genericity and the diversity of sources, the project is applied in three different use cases or sub-projects with the aim to cover different methods and approaches that are typical in the domain of HPST. Those approaches are covering only partially the methods used in history and archaeology. These sub-projects are described in the following paragraph.

2.1 The port-arsenal of Brest

This sub-project takes part in the research programs “History of marine science and technology” and “Digital Humanities for History of Science and Technology” developed in Brest in the Centre F. Viète. One topic concerns the comprehension of the scientific and technological evolution of the port-arsenal in Brest (France) on a large period (XVIIth to XXth century) with a methodological approach considering this military-industrial complex dedicated to shipbuilding as a large technological system [10]. The objectives are:

1. To compose and publish a digital library (based on semantic web) about the material culture of the port-arsenal of Brest associated to several projects

[3] Participants at this consortium came initially from LaB-STICC (Telecom Bretagne, Brest), Centre François Viète (University of Brest), LIRDEF (University of Montpellier), LHSP-Archives Poincaré (University of Lorraine, Nancy) and later LORIA (University of Lorraine, Nancy). During the years 2012-2014, the INSHS (a French national institute of human and social sciences), the national network of Maisons des Sciences de l’Homme and University of Lorraine supported this consortium.

about 3D replications of artifacts and to cultural mediations dedicated to science and technology heritage.

2. To develop digital tools (based on semantic web) dedicated to a comparative history of science and technology of the port on a large area and a large period (since ancient times until now).

The hypothesis is to consider the large technological system of the port-arsenal as a large spatiotemporal and multi-scale artifact which is possible to decompose in elements of smaller scale (which are also artifacts) like industrial workshops, shipbuilding areas, storage areas, etc. Each of these elements are themselves composed by elements/artifact of smaller scale. The system has to be seen as the sum of all these artifacts and of all the relationships between them. The research in Brest [11], [12] has shown the interest to propose an historical evolution model of the port (inspired by works in geography [13]) where "simple" artifact like cranes, quays, dry docks are efficient indicators to characterize the cycle of evolution of the port-arsenal during a large period. This method is used in a comparative research [14] between Brest (France) and Mar del Plata (Argentine) in a thesis in progress by B. Rohou (directed by S. Garlatti and S. Laubé).[4] From these works, the contribution in the SemanticHPST group is to produce a methodology and a knowledge model efficient to produce a generic ontology where an artifact is a material object (made by human beings) associated to a "life cycle" with at least three steps:

1 design and construction of the artifact;
2 the artifact in use;
3 the disappearance of the artifact.

That "life cycle" involves the elaboration of fives categories of entities: time entities, actors (individuals or social groups), concepts/theories, location and artifacts. The analysis of the important ontology in the domain of cultural heritage named CIDOC-CRM (that "provides definitions and a formal structure for describing the implicit and explicit concepts and relationships used in cultural heritage documentation")[5] shows that this ontology could be a first reference to help and build our own ontologies because some concepts and relationships about "temporal entities" and "actors" can be reused. But if the concept of "Thing" exists in the CIDOC-CRM, we consider that the concept of "Artifact" and the associated relationships have to be elaborated first from our historical model and by considering of course the possibility of equivalent concepts in the CIDOC-CRM. A work is in progress in Brest about this topic from concrete examples of artifact as crane, quays and seawalls. A second step will be to examine others methods to produce ontologies well-adapted to our HPST problems in the domain of marine history [15].

This work is coupled with examples of typical requests (when and where were positioned all the cranes in the port of Brest since 1650 until 1970? In the

[4] See http://brmdp.hypotheses.org/.

[5] http://www.cidoc-crm.org/.

port of Mar del Plata? Which firms were in charge of the construction of the quays/cranes in the port of Brest since 1800 until 1900? What are the engine power of all cranes in the world since 1850 until 1970? Etc.).

2.2 Henri Poincaré's correspondence

The platform Henri Poincaré papers. In 1992, the laboratory of history of science and philosophy Archives Henri Poincaré was created to promote Henri Poincaré's manuscripts and to publish his correspondence. For more than 20 years, this long-term project has produced three volumes of letters: the first one is devoted to the Poincaré - Mittag-Leffler letters [16], the second one is on the correspondence with physicists, chemists and engineers [17], the third one is with astronomers and, in particular, geodesists [18]. Two other volumes are in preparation, one devoted to the letters from or of mathematicians and the other one consists of administrative and personnal correspondences.[6]

The corpus consists of more than 2000 letters, 1046 sent by Henri Poincaré and 949 received by him.[7] All known letters are digitalized[8] and around 50% of them are in plain text (in LaTeX and XML versions). Lots of letters contain mathematical and physical formulae. In Henri Poincaré Papers website,[9] the correspondence is available. In this platform, each known letter is indexed with Dublin Core extended metadata.[10] This enables to query the corpus by e.g.

$Q_1 =$ "Letters sent by Henri Poincaré in 1885"

$Q_2 =$ "Letters received by Eugénie Launois between 1882 and 1894"

There is also the possibility of plain text search for the letters already transcribed.

Towards more HPST-adapted search. Now, consider the following queries:

$Q_3 =$ "Letters from an astronomer"

$Q_4 =$ "Letters in reply to a letter of Mittag-Leffler"

$Q_5 =$ "Letters about the n-body problem"

$Q_6 =$ "Letters of the late XIXth century"

These queries cannot be executed in the current platform. They require additional data and knowledge:

[6] This correspondence is partly online http://henripoincarepapers.univ-lorraine.fr.

[7] About 50% of this letters are with scientists. Original letters come from 63 different archive centers and libraries from 14 countries.

[8] Due to copyright laws, some are not available online.

[9] http://henripoincarepapers.univ-lorraine.fr.

[10] It exists different projects devoted to scientific correspondences for example the CKCC project (http://ckcc.huygens.knaw.nl) [19] or Mapping the Republic of Letters (http://republicofletters.stanford.edu).

- Q_3 requires to know that an individual is an astronomer, possibly using deduction (for instance, Rodolphe Radau was a geodesist and every geodesist is an astronomer).
- Q_4 requires to know relationships between letters (including lost letters).
- Q_5 requires semantic annotations about the content of the letters (Poincaré worked on the three-boby problem).
- Q_6 raises the problem of modeling "late XIXth century": the boundaries of interval of time are imprecise.

The possibility to take into account such queries using semantic web principles and technologies, are examined in the SemanticHPST consortium.

2.3 The concept of energy

One part of the SemanticHPST project is dedicated to the concept of energy. Our aim is to create an ontology of energy for researchers working in the field of HPST as well as for science teachers.

For researchers, the ontology aims at making available a methodical body of knowledge that allows previously unseen connections to be made. For example, correspondence between two authors or the presence of a specific term or concept in a text will allow researchers to put forward hypotheses regarding the emergence of an idea or the cross-fertilization of ideas.

For teachers, the ontology aims at acting as a resource, allowing educators to find historical information relevant to school curricula as well as ideas for specific activities to carry out in the classroom.

The content consists of reference texts in the field of HPST, contemporary scientific texts and a database of historic scientific instruments and documents. This content is currently being selected and developed and will be enhanced as the research progresses.

To date, the following three steps have been undertaken on the project:

- The first step was to identify the presumed ways the ontology will be used, for example, the type of requests that a researcher or teacher might make in a search. To this end, one 'persona' for a researcher and one for a teacher have been created. Analyzing the theoretical queries from these two personas helps in the selection of a relevant body of work and is also a useful guide for indexing.
- The second step was to begin indexing the reference texts. Duhem, Poincaré, Mach and Meyerson have been selected for a first approach in order to produce keywords and common references and to outline an embryonic model. Using the shared scientific knowledge of the physicists involved in the project, a sort of 'cloud' of concepts related to describing energy was defined and classified. These elements led to the structure of an initial mind map.
- Finally, based on this mind map (created with Docear), we used Protégé software to create a first draft overview of the project. The next steps require documenting these three steps in detail to refine the data and then build the ontology.

During the stages of the project carried out so far, various problems have been identified that must be resolved. One of the main problems concerns the modeling of time. How can an event be modeled? Moreover, how can knowledge be modeled in a way that avoids immobilizing the knowledge? How should knowledge be contextualized? What approach should be adopted when modeling concerns a concept or an object? How can a coherent and logical body of content be created and how can its coherence be assessed? It is clear that the question of time as well as how to approach the treatment of objects and works are issues to be investigated in the semanticHPST project.

3 The SemanticHPST tools and requirements

According to the three described sub-projects, the main goals of researchers in HPST are to access and retrieve relevant resources in existing primary and secondary sources or corpora, to produce new resources in existing corpora, to enrich existing digital corpora or to create new ones, for answering research questions in the history of science and technology. Existing digital corpora come from libraries, information holdings, digital libraries or others like Gallica (http://gallica.bnf.fr), Internet Archive (http://archive.org), Google Books (http://books.google.com), etc., and CMS (Content Management System) (blogs, wikis, Drupal, Omeka, etc. more generally social media tools) have been used by the community[11] and digital AHP (http://www.ahp-numerique.fr/). Some heritage and bibliographic resources have already been described by several institutions, associations and/or project (BNF, Gallica, British Museum, Europeana, Amsterdam Museum, LODLAM, ...). The creation of new corpora or resources can be made on social media tools distributed on Internet (as well as other digital corpora).

The design of tools for HPST researchers has to integrate and/or aggregate the existing heterogeneous tools and to ensure interoperability among them. Thus, the goal is not to build a single new environment, but to design a platform which integrates existing tools selected for their relevance according to the practices of researchers and provide an agile architecture able to model and/or support the processes involved in the research work and enrichment.

This platform will be mainly based on the Semantic Web and Linked Data approaches (RDF Triple Store, ontologies, OWL 2, RDFS, SPARQL, etc.). Nevertheless, the platform will also provide access to non-semantic resources. A network of ontologies dedicated to HPST will be designed to meet the interoperability and open access requirements for corpora. Some existing ontologies and standards will be reused and integrated in the ontology network, like CIDOC-CRM, FRBRoo, FRSAD, Dublin Core, etc. and those available at LOV (http://lov.okfn.org/dataset/lov/).

In this paper, we focus our attention on the resource retrieval problem that we can divide into two different aspects : advanced search in HPST corpora and harvesting distributed corpora. The former focuses on advanced search function-

[11] The *alambic numérique* (http://alambic.hypotheses.org/4924) is based on Omeka.

alities in a single corpus. The latter studies the resource retrieval on distributed corpora. These two aspects will be integrated.

3.1 Advanced search in HPST corpora

In order to perform advanced searches in a HPST corpus, we have to build intelligent digital corpus: corpus with primary and secondary sources having semantic metadata (RDF Triples) and their corresponding ontologies using a fragment of OWL (actually, RDFS will be sufficient for the following examples). These ontologies are domain ontologies related to the corpus. A domain ontology for Henri Poincaré letters has already been designed. Finally, some tools will have to be developed for answering some of the queries.

This section presents the advanced search using the query examples Q_3-Q_6 introduced in Section 2.2.

Q_3 requires some additional data and knowledge to get satisfactory answers, as stated in Section 2.2. In particular, if the annotation file contains the following RDFS triples:

$$\begin{array}{l}(\texttt{letter1 isSentBy rodolphe\_radau})\\(\texttt{rodolphe\_radau rdf:type Geodesist})\\(\texttt{Geodesist rdfs:subClassOf Astronomer})\end{array}$$

then the execution of the following SPARQL query on an engine supporting RDFS

$$Q_3 = \text{SELECT } \texttt{?}\ell \text{ WHERE } \{\texttt{?}\ell\ \texttt{isSentBy ?a . ?a rdf:type Astronomer}\}$$

will return `letter1`.

Q_4, similarly, can be answered by a SPARQL engine supporting RDFS with the following query:

$$Q_4 = \text{SELECT } \texttt{?}\ell \text{ WHERE } \left\{\begin{array}{l}\texttt{?}\ell\ \texttt{isAnAnswerTo ?}\ell\texttt{2 .}\\\texttt{?}\ell\texttt{2 isSentBy mittag-leffler}\end{array}\right\}$$

It can be noticed that this query can give a letter of the corpus that answers a lost letter: the missing letter cannot be found, but its answer can.

Q_5, for being executed, requires the use of annotations about the scientific content of the letter:

$$Q_5 = \text{SELECT } \texttt{?}\ell \text{ WHERE } \left\{\begin{array}{l}\texttt{?}\ell\ \texttt{hasForTopic ?t .}\\\texttt{?t rdf:type N-body-problem}\end{array}\right\}$$

The n-body problem is a topic having sub-topics, in particular, the 3-body problem is a problem more specific than the n-body problem. For this reason, we have chosen to model these two problems by two classes, the former being more general than the latter. Therefore, a letter of the corpus about the 3-body problem will be returned by the execution of this query.[12]

[12] We could also have chosen to model the 3-body problem as an instance of the n-body problem, but first, it is more homogeneous to consider every topic as a class,

Q_6 can be modeled by a SPARQL query based on the assumption that "the late $\textsc{xix}^{th}$ century" corresponds to the interval $1881 - 1900$:

$$Q_6 = \textsc{select}\ ?\ell\ \textsc{where} \left\{ \begin{array}{l} ?\ell\ \texttt{sentDuringYear ?y .} \\ \textsc{filter}(\texttt{?y >= } 1881\ \texttt{\&\& ?y <= } 1900) \end{array} \right\}$$

However, this solution is debatable: the modeling of the fuzzy period of time by a crisp interval raises the problem of the choice of the boundaries. Indeed, some events before 1881 or after 1900 can be considered by historians to be related to the end of the $\textsc{xix}^{th}$ century. In order to address this issue, some approximate search is planned. How to put this idea in practice is an ongoing work.

3.2 Harvesting distributed corpora

Harvesting distributed corpora at semantic level (according to Linked Data principles) require to solve two different problems. The first one is to queries several triple store by means of federated queries to linked distributed sources. The second one is to get RDF triples from social media tools.

Most of social media applications are data silos. In other words, data are unavailable on the web. Only people may have access to data, not computers. Reuse and exchange of data among social media tools are only possible by means of API – that is to say manually by mean of one API per tool. Some social media tools like Drupal, Semantic media wiki may have their own triple store exposing data to others.

A toolkit, called SMOOPLE for Semantic Massive Open Online Pervasive Learning Environment, has been designed to solve these two problems. It was firstly dedicated to the technology-enhanced learning domain [20]. The core part of the toolkit can be reused for HPST. It fulfills the needs of researchers in HPST, that is to say it enables us to federate distributed sources and tools.

SMOOPLE has semantic services which are in charge of managing incorporated semantic models, extracting and storing the data produced on social media tools, making and answering to semantic queries against one or several distributed sources (federated queries). The Semantic Web server (semantic services) is based on Jena 2. When the social media tools do not have a triple store and a SPARQL endpoint, content and corresponding semantic metadata can be extracted on the fly from social media applications, by means of plugin (similar to sioc_export) and stored in a RDF repository. Several light ontologies (SIOC, FOAF, DC, RDF, RDFS, etc.) are used to acquire semantic metadata automatically. It will be necessary to define the interlinkage among distributed sources (triple stores) to support federated queries.

second, this way, it is always possible to consider a more specific topic, e.g., the restricted 3-body problem for which the mass of one of the 3 bodies in considered to be negligible.

4 Epistemological aspects

An aim of the SemanticHPST project is to focus on the epistemological issues raised by the development of these new tools based on semantic web. This work in progress takes part to epistemological questions in the domain of Digital Humanities.[13] A first series of questions concerns the modeling of knowledge, the main step in building ontologies so that researchers can easily identify and apprehend knowledge. Therefore the creation of effective ontologies requires defining concepts and elucidating certain tacit or implicit knowledge. So the initial questions are: How to approach these definitions? How to ensure that indexing does not immobilize knowledge? How can the modeling anticipate how it will be used in order to ensure that the knowledge generated is contextualized to avoid anachronism and misinterpretation? Moreover, the wide range of works in the collection, including texts (manuscripts, books, letters, web pages), multimedia documents, 3D archaeological or historical objects and media from a variety of sources (photographs, original texts, maps, etc.), necessitate different approaches. This raises the question: How to approach a photograph, a scientific instrument or a text and still obtain a unified ontology? How can the modeling enable relationships between objects yet avoid the pitfalls described above?

In the field of HPST, the issue of modeling time is central and particularly tricky. Modeling a long period of time, an event, a succession of events or events that are juxtaposed requires making decisions that should be taken collectively. Indeed, this emerging issue is shared by historians [23], [24], [25] and should serve to feed into theoretical discussions between researchers from different disciplines.

A second series of questions concerns the researcher's environment, which has significantly changed with the rise of digitized data. Whatever the works considered or their origin (libraries, archives, etc.), the massive volume of data, its diversity and location are all part of this change. Yet this radical shift is not exclusively the result of the accumulation of a large amount of data. The fact that data can be 'analyzed as well as communicated, represented, reused – in short, mobilized for research – in a quantity and with an ease incomparable with previous periods' [3] is a major transformation that needs to be taken into account. This raises new questions for researchers:

- How does one build and define a body of content that is coherent and complete? Whereas 'traditional' methods created collections using identified, bounded, localized archives, with the question of consistency limited in most cases to the cross-fertilization of archives as regards the historical context, the accessibility of multiple documents today requires a reexamination of the very concept of a collection of works.
- How does one evaluate a body of work; in other words, how does one recognize its relevance?
- In this context, the type of source and its references must be specified. Does the wide range of sources used require more refined classification than the

[13] See thematical issue "la numérisation du patrimoine" of [21] or the issue "Le métier d'historien à l'ère numérique : nouveaux outils, nouvelle épistémologie ?" of [22].

standard usage of primary and secondary sources? Would a new typology be pertinent given this broad diversity? Should the references to these sources, particularly information concerning digital archives, lead to new codification that allows, for example, multiple identifications for the considered source, improving its accessibility?

5 Conclusion

The aim of this proposal is to contribute to the development of the research in the domain of digital humanities. Based on the Semantic Web principles and technologies, the SemanticHPST group proposes new methodologies in History and Philosophy of Science and Technology in the framework of a strong collaboration between labs working in the area of computer science and humanities (here HPST). The main goal is to enrich the practices of researcher and communities in HPST as well in science and technology heritage. To deal with such a goal, the project has to: i) Build intelligent digital corpora, that is to say corpora with primary and secondary sources having semantic metadata and their corresponding ontologies; ii) Design tools to access and enrich existing corpora and to create new ones; iii) Evaluate the resulting evolution of practices in historical science and build an epistemological viewpoint about the impact of new tools and practices in humanities based on knowledge modeling and semantic web.

Another important issue is to deal with the reuse of intelligent digital corpora. Thus, it is necessary to build representations of the entities, people and processes involved in producing the digital corpora. The "PROV Model Primer" from W3C (http://www.w3.org/TR/prov-primer/) can be used to address this issue.

References

1. V. A. Ustinov, "Les calculateurs électroniques appliqués à la science historique," *Annales. Économies, Sociétés, Civilisations*, vol. 18, no. 2, pp. 263–294, 1963.
2. O. Boonstra, L. Breure, and P. Doorn, "Past, Present and Future of Historical Information Science," *Historical Information Science*, vol. 29, no. 2, pp. 4–132, 2004.
3. M. Dacos and P. Mounier, "Humanités numériques," rapport commandé, Institut Français, Ministère des Affaires étrangères, Paris, 2014.
4. A. Meroño-Peñuela, A. Ashkpour, M. van Erp, K. Mandemakers, L. Breure, A. Scharnhorst, S. Schlobach, and F. van Harmelen, "Semantic technologies for historical research: A survey," *Semantic Web Journal*, pp. 1–27, 2015.
5. O. Bruneau, P. Grapi, H. Peter, S. Laubé, M.-R. Massa-Esteve, and T. De Vittori, *History of Science and Technology, ICT and Inquiry Based Science Teaching*. Berlin: Frank-Timme, 2012.
6. O. Bruneau, S. Laubé, and T. de Vittori, "ICT and History of mathematics in the case of IBST," in *[5]*, pp. 145–160, 2012.
7. O. Bruneau and S. Laubé, "Inquiry based Science teaching and History of Science," in *[5]*, pp. 13–28, 2012.
8. J. M. Gilliot, N. C. Pham, S. Garlatti, I. Rebaï, and S. Laubé, "Tackling Mobile & Pervasive Learning in IBST," in *[5]*, pp. 181–201, 2012.

9. M. Guedj and M. Bachtold, "Towards a new strategy for teaching energy based on the history and philosophy of the concept of energy," in *[5]*, 2012.
10. T. P. Hughes, "The Evolution of Large Technological Systems," in *The Social Construction of Technological Systems* (W. Bijker, T. P. Hughes, and T. J. Pinch, eds.), pp. 51–82, Cambridge, Massachusetts: MIT Press, 1987.
11. S. Laubé, "Les grues de l'arsenal en tant que marqueurs de l'évolution scientifique et technologique du port arsenal de Brest," in *[?]*, To be published in 2015.
12. S. Laubé, "Culture matérielle du port arsenal de Brest au XVIIIème siècle : approche systémique," in *[?]*, To be published in 2015.
13. J. Bird, *The Major Seaports of the United Kingdom.* London: Hutchinson, 1963.
14. S. Laubé, B. Rohou, and S. Garlatti, "Humanités numériques et web sémantique. De l'intérêt de la modélisation des connaissances en histoire des sciences et des techniques pour une histoire comparée des ports de Brest (France) et Mar del Plata (Argentine)," in *Digital Intelligence 2014*, September 17-19, 2014.
15. V. de Boer, M. van Rossum, J. Leinenga, and R. Hoekstra, "Dutch ships and sailors linked data," in *The Semantic Web – ISWC 2014* (P. Mika, T. Tudorache, A. Bernstein, C. Welty, C. Knoblock, D. Vrandečić, P. Groth, N. Noy, K. Janowicz, and C. Goble, eds.), vol. 8796 of *Lecture Notes in Computer Science*, pp. 229–244, Springer International Publishing, 2014.
16. P. Nabonnand, ed., *La correspondance entre Henri Poincaré et Gösta Mittag-Leffler.* Basel: Birkhäuser, 1998.
17. S. Walter, E. Bolmont, and A. Coré, eds., *La correspondance entre Henri Poincaré et les physiciens, chimistes et ingénieurs.* Basel: Birkhäuser, 2007.
18. S. Walter, R. Krömer, and M. Schiavon, eds., *La correspondance entre Henri Poincaré avec les astronomes et les géodésiens.* Basel: Birkhäuser, 2014.
19. P. Wittek and W. Ravenek, "Supporting the Exploration of a Corpus of 17th-Century Scholarly Correspondences by Topic Modeling," in *Supporting Digital Humanities 2011: Answering the unaskable* (B. Maegaard, ed.), 2011.
20. J.-M. Gilliot, S. Garlatti, I. Rebaï, and C. Pham Nguyen, "A Mobile Learning Scenario improvement for HST Inquiry Based learning," in *Workshop Emerging Web Technologies, Facing the Future of Education* (, ed.), 2012. Workshop in conjunction with www2012 conference.
21. *Documents pour l'Histoire des Techniques*, vol. 18-2. 2009.
22. *Revue d'histoire moderne et contemporaine*, vol. 58-4bis. 2011.
23. A. Neelameghan and G. J. Narayana, "Concept and expression of time: Cultural variations and impact on knowledge organization: PART 7: Ontology and representation of time in knowledge organization tools used in information systems1," *Information Studies*, vol. 19, no. 2, p. 105–131, 2013.
24. I. Corda, B. Bennett, and V. Dimitrova, "A logical model of an event ontology for exploring connections in historical domains," in *Workshop on Detection, Representation and Exploitation of Events in Semantic Web (Derive 2011), Tenth International Semantic Web Conference (ISWC)*, 2011.
25. E. Hyvönen, T. Lindquist, J. Törnroos, and E. Mäkelä, "History on the semantic web as linked data–an event gazetteer and timeline for the world war i," in *Proceedings of CIDOC*, 2012.

Semantic Web Based Named Entity Linking for Digital Humanities and Heritage Texts

Francesca Frontini[1,2], Carmen Brando[1], and Jean-Gabriel Ganascia[1]

[1] Labex OBVIL. LiP6. CNRS, 4 place Jussieu, 75005, Paris,
{Francesca.Frontini,Carmen.Brando,Jean-Gabriel.Ganascia}@lip6.fr
[2] Istituto di Linguistica Computazionale CNR, Pisa, Italy,
{Francesca.Frontini}@ilc.cnr.it

Abstract. This paper proposes a graph based methodology for automatically disambiguating authors' mentions in a corpus of French literary criticism. Candidate referents are identified and evaluated using a graph based named entity linking algorithm, which exploits a knowledge-base built out of two different resources (DBpedia and the BnF linked data). The algorithm expands previous ones applied for word sense disambiguation and entity linking, with good results. Its novelty resides in the fact that it successfully combines a generic knowledge base such as DBpedia with a domain specific one, thus enabling the efficient annotation of minor authors. This will help specialists to follow mentions of the same author in different works of literary criticism, and thus to investigate their literary appreciation over time.

Keywords: named-entity linking, linked data, digital humanities

1 Introduction

Named Entities (NE) are linguistic expressions that stand like rigid designators for referents; such entities normally include names of persons, geographical places, organizations, but also temporal references such as dates. Enriching mentions with a link to its referent by means of a unique identifier is crucial for the semantic annotation of texts. This is done by pointing to an external resource, such as a Universal Resource Identifier (URI) in the Linked Open Data (LOD) cloud. Segments in text referring to a Named Entity are known as entity mentions.

Named Entity Linking (NEL) [9] is a sub task of Named Entity Recognition and Disambiguation (NERD). NERD algorithms automatically detect entities in texts and assign them to a given class[3]. The NEL module assigns a unique identifier to the detected entities, thus disambiguating them by pointing to their referent. Linking is crucial since the same mention can represent different entities in different contexts and at the same time one entity can be mentioned in the text in different forms. So for instance the mention "Goncourt" can refer

[3] See [8] for a survey on NER.

to any of the two Goncourt brothers, Edmond or Jules. At the same time Jules de Goncourt can be referred to in the text as “Goncourt”, “J. Goncourt”, “J. de Goncourt”, ... This means that, in order to automatically retrieve all passages in a text where Jules de Goncourt is mentioned, it is necessary not only to annotate all these mentions as a Named Entity of the class person, but to provide them with a unique key that distinguishes them from those of other people, in this case those of Edmond. The bibliographic identifier “Goncourt, Jules de (1830-1870)”, as well as the links <http://www.idref.fr/027835995> and <http://fr.dbpedia.org/page/Jules_de_Goncourt> are examples of such an identifier.

Besides ensuring disambiguation, linking also performs an important additional task, namely textual enrichment, in that it connects the mention with sources of additional information - such as DBpedia in the previous example - that needs not be stored in the text but can be accessed when required. In the case of Edmond de Goncourt, additional information from DBpedia can tell us what books he authored, where he was born, ...

The main issue with NEL in digital humanities is that mentions of persons often refer to individuals that are not listed in general ontologies such as Yago or DBpedia, that constitute the typical knowledge base for linking in other domains. Such individuals are often present in other knowledge bases, notably bibliographical linked data repositories (such as the French National Library BnF linked data repository). On the other hand, linking requires access to ontological knowledge, in that choosing between two individuals having the same name may requires comparing the context of the mention with a priori knowledge. In this respect, knowledge bases such as DBpedia remain an important source of general knowledge of the World. Thus the ideal linking algorithm for literary criticism texts combines general and domain specific sources. The experiment here described goes in this direction.

The paper will first present previous approaches to NEL, then the proposed graph based disambiguation algorithm based on the notion of centrality, finally describe the experiment carried out on the corpus and the results. Some conclusions and suggestions for further improvement of the algorithm are finally given.

2 Previous approaches

Previous approaches for NEL can be divided in two main families. Those using text similarity and those using graph based methods. Both these methods are unsupervised, and they do not rely on pre-annotated corpora for training.

The best known tool of the first group is DBpedia Spotlight [7], that performs NER and DBpedia linking at the same time. Spotlight identifies the candidates for each mention by performing string similarity between the mention and the DBpedia labels, then it decides which entry is the most likely by comparing the text surrounding the mention with the textual description of each candidate. The referent whose description is more similar to the context of the mention in terms of TF/IDF is chosen. This method is known to be very efficient, but it can only provide linking towards resources such as DBpedia, whose entries come with a description in the form of unstructured text. Other knowledge bases do not provide a textual description for their entries, such is the case of the bibliographical databases that constitute the ideal linking for mentions of authors.

Graph-based approaches rely on formalised knowledge described in graph form that is built from a Knowledge Base (KB) (e.g. the Wikipedia article network, Freebase, DBpedia, etc.). Reasoning can be performed through graph analysis operations. It is thereby possible to at least partially reproduce the actual decision process with which humans disambiguate mentions. A reader may decide that the mention "James" refers to philosopher "William James" and not to writer "Henry James" because it occurs in the same context as "Hume" and "Kant". In the same way such algorithms build a graph out of the candidates available for each possible referent in a given context and use the relative position of each referent within the graph to choose the correct referent for each mention. The graph is built for a context (such as a paragraph) containing possibly more than one mention, so that the disambiguation of one mention is helped by the other ones.

This kind of approach is similar to the one used in Word Sense Disambiguation [11], where a set of words in a given sentence needs to be labeled with the appropriate sense label by using the information contained in a lexical database such as WordNet. The key idea of this approach is that for all ambiguous words in the context, senses that belong to the same semantic space should be selected, and that in this way two ambiguous words can mutually disambiguate each other. More specifically, a subgraph is built, constituted only of the relevant links between the possible senses of the different words, and then for each alternative sense labeling, the most central is chosen. This procedure, when applied to such context specific subgraphs, ensures that in the end the chosen senses for each word will be the one better connected to each other.

Centrality is an abstract concept, and it can be calculated by using different algorithms[4]. In [11] the experiment was carried out using the following algorithms: *Indegree*, *Betweenness*, *Closeness*, *PageRank*, as well as with a combination of all these metrics using a voting system. Results showed the advantage of using centrality with respect to other similarity measures. While the combination of all centrality algorithms scores the best, Indegree centrality seems to be the better performing when compared to the other ones in terms of precision.

This graph based approach has been applied to NEL, where mentions take the place of words and Wikipedia articles that of WordNet synsets. Here too centrality measures are performed on the Wikipedia structure in order to use the rich set of relations to disambiguate mentions. More specifically in [4] English texts were disambiguated using a graph that relies only on English Wikipedia, and was constituted of the links and of the categories found in Wikipedia articles. So for instance the edges of the graph represent whether ArticleA links

[4] For a discussion of the notion of centrality see also [10]

to ArticleB or whether ArticleA has CategoryC. Here too "local" centrality is then used to assign the correct link to the ambiguous mention. We have chosen a graph-based approach to NEL that will be described in the next section.

3 Our approach

Our approach to disambiguate NE mentions is a graph-based one. Vertices are represented by URIs of mention candidates (e.g. dbpedia:Victor_Hugo) as well as URIs of concepts (e.g. foaf:Person) or individuals connected to at least two different candidates. Edges are semantic relations defined explicitly between URIs (e.g. "type"). The graph is undirected and their vertices and edges are *a priori* unweighted. We take advantage of the notion of centrality in Graph Theory to link a NE mention with the URI of the most probable candidate for that mention. In other words, we want to find the subset of vertices of different candidates having the greatest number of edges among them. The edges and vertices of the graph are built leveraging knowledge from different LOD sources whose nature is graph-based.

We illustrate the proposed approach with an example. Let us consider the following phrase of a French text of literary criticism written by Albert Thibaudet (1936) :

Quant au rythme, si ***Victor Hugo*** *a dépassé* ***Lamartine****, il n'a pas été plus loin que* ***Vigny****.*

In bold there are three mentions automatically recognized by a NER algorithm, that need now be linked to an identifier.

For each mention, the NEL algorithm selects possible candidates by exact string matching of the current mention and dictionary entries (e.g. Hugo, M. Hugo) and retrieves the corresponding URIs of the listed LOD sources. An excerpt of the candidates of the three named-entities from the example is listed below by distinguishable personal information instead of URI for readability sake.

Candidates (Victor Hugo) = Hugo, Victor (1802-1885)

Candidates (Lamartine) = Lamartine, Alix de (1766-1829), Lamartine, Alphonse de (1790-1869), Lamartine, Elisa de (1790-1863)

Candidates (Vigny) = Vigny, Joseph Pierre de (1742-1812), Vigny, Benno (1889-1965), Vigny, Alfred de (1797-1863)

Thanks to the URIs, it is possible to retrieve from the Web of Data the associated RDF graph for each candidate and combine them into a single graph. It should contain only those predicates involving at least two candidates of different mentions because we only want the predicates that play an important role in the disambiguation process. Calculating the centrality for every candidate will then give us the best candidates for the three mentions. Figure 1 shows an excerpt of the resulting graph where the chosen mention candidates are marked in bold. We can notice that the vertex yago:RomanticPoets is the one that influences the centrality measure the most because it is shared by the three chosen candidates. Likewise, other vertices connected to the chosen nodes, such as dbpedia:romanticisme and dbpedia:Alexandru_Macedonski, are influential.

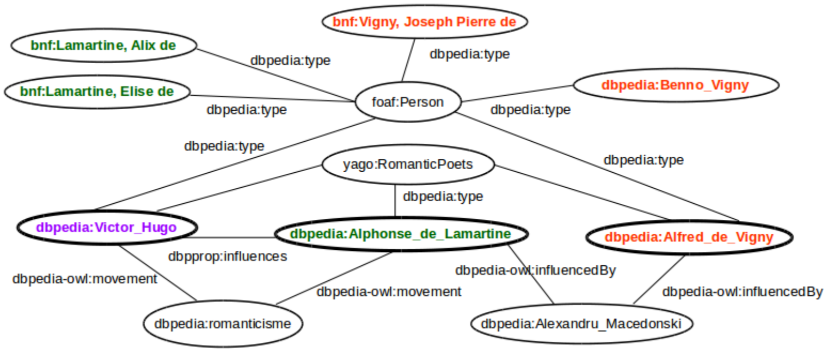


Fig. 1. Excerpt of the chosen URIs (in bold) for three candidates; a color designates all candidates of a single mention.

Named Entities are disambiguated and referenced within the context of a paragraph, so in principle two (identical) mentions of the same author within one paragraph will always receive the same link, while the same mention in different paragraphs might be assigned a different referent, depending on the other mentions it occurs with.

The NEL task is commonly defined in such a way that it does not assume the existence of the correct referent among the candidates in the knowledge base [5]. This is due to the fact that Wikipedia/DBpedia can hardly be a complete knowledge base even for textual genres such as contemporary newspapers articles. This seems even less true for the corpus that constitutes the object of our experiment. French literary criticism texts contain references not only to famous authors, but also to other minor figures that are not listed in Wikipedia. Therefore our proposal is to aim for a quasi complete reference base for the task of referencing authors.

Our approach relies importantly on a lookup dictionary; this is the subject of the following section.

4 LOD-based lookup dictionary

Linked data [1] is an important way of publishing knowledge in the Semantic Web. Such data is easily available via web services; LOD is composed of triplets of the form (subject, predicate, object) where subjects designate URIs, objects may be URIs or data-typed literals, and predicates represents binary relations. Queries can be run in the SPARQL language and data is provided with a dereferenciable and persistent identifier called URI (Uniform Resource Identifier). Many of the available linked data are of great interest for digital humanities [12], and for the domain of literary criticism in particular. More specifically, information on authors for French texts can be found in the French version of DBpedia on the one hand, and in the catalogue of the Bibliothque Nationale de France (BnF) on the other.

The French DBpedia is constituted of the articles of the French version of Wikipedia. In DBpedia entries are classified one or more of the types of the DBpedia ontology. So for instance the author known as Stendhal[5] is classified as *Person*, *Artist*, *Writer*, and at the top level, as *Thing*. Moreover, authors are linked to each other by horizontal relations such as *InfluencedBy*, and, indirectly, by being linked to the same concept, such as *Romanticism*. BnF entries list all authors of books ever published in France; their entries contain information on date of birth and death, gender, alternative names, works authored. For instance the BnF entry for Voltaire[6] gives several alternative names such as Franois-Marie Arouet (Voltaire's real name), Wolter, Good Naturd Wellwisher, ...

Most crucially, BnF links its entry to the DBpedia one when existing, thus making it very easy to connect the two resources in one knowledge graph. Moreover, BnF entries also list the author's Idref, which is the official identification system used by French universities and higher education establishments to identify, track and manage the documents in their possession. The combination of these two sources was considered able to grant a sufficient coverage for a corpus of French literary criticism, thus the BnF and the DBpedia SPARQL endpoints were queried for all authors, retrieving their biographic information (name, surname, alternative names, dates of birth and death, title, ...) in structured form.

In order to be able to retrieve all possible mentions of an author, this information was processed into a dictionary of authors, that contains all alternative names of an author, plus a series of alternative forms automatically generated, with the links to BnF and DBpedia entries. Automatically generated alternative names are of the form:

- surname only (Rousseau)
- initials + surname (J.J. Rousseau, JJ Rousseau, ...)
- title + surname (M. Rousseau, M Rousseau)

Gīvėn̈ the domain (French literature) this procedure ensures that the retrieval of at least one candidate URI for most mentions. At the same time, the mass of information present in the BnF repository will generate several homonyms and make most mentions ambiguous; thus good disambiguation becomes crucial.

5 Implementation of the NEL algorithm

The NEL algorithm processes a file in XML-TEI format[7]; NE mentions are annotated with NER annotations (e.g. tag <persName>) for every paragraph; the

[5] http://fr.dbpedia.org/page/Stendhal

[6] http://data.bnf.fr/11928669/voltaire/

[7] http://www.tei-c.org/index.xml

algorithm is devised to processes one single class at a time (here Person). It uses a lookup dictionary per class listing superficial forms and their associated URIs from LOD sources, as described in the previous section. The algorithm produces an enriched version of the input file indicating the chosen candidate for each mention. We developed our implementation in Java ; RDF data is processed thanks to the Jena API[8]; graphs are manipulated by the JgraphT API[9] and implementation of centrality measures are available in the Social Network analysis tool, JgraphT-SNA[10]. In particular, the algorithm performs the following steps for every paragraph of the XML-TEI file:

1. look for URIs of mention candidates in the dictionary
2. retrieve the RDF graphs of those URIs
3. simplify and combine graphs then compute the selected centrality measure
4. choose URI of candidate with the higher score per mention then write results in TEI file

The algorithm searches for (1) possible candidates of mentions by exact string matching the mentions of the current paragraph and superficial forms in the dictionary; there must be at least one ambiguous mention to continue. It retrieves URIs (BnF, DBpedia) of mention candidates from dictionary entries. Next, the RDF graph is retrieved (2) for every URI and converted to a JgraphT-compatible graph, where RDF objects and subjects are vertices and RDF predicates are edges. Irrelevant edges and vertices are removed from graphs. We keep edges which involve at least two vertices representing URIs candidates. Information coming from different sources is combined into a single graph (3); the way we combine graphs is straightforward. The fusion is implicitly done thanks to one of the main LOD principles which consists of reusing vocabularies published in the LOD vocabulary cloud. In other words, edges (predicates) and vertices (URI nodes) should be shared by at least two graphs associated to candidates of different mentions. The selected centrality measure (e.g. closeness) is calculated for the resulting graph. Finally, the algorithm chooses (4) the URI of the mention candidate with the higher centrality score and annotates the input XML-TEI file with this information.

Furthermore, simplification of graphs and calculation of centrality measures in the combined graph are crucial parts of the algorithm (3). This step is detailed in the Algorithm 1. It essentially removes edges which are irrelevant to calculate a centrality measure, in other words, it deletes those edges which involve at most one vertex of a non-candidate URI.

6 Experiments and results

This section describes the experiments settings used to test our proposal as well as preliminary results which are encouraging. In this experiment, in order to

[8] https://jena.apache.org/

[9] http://jgrapht.org

[10] https://bitbucket.org/sorend/jgrapht-sna

Algorithm 1 NEL: simplify and combine graphs, compute centrality

```
Require: graphs: graphs of candidates per mention, measure: centrality measure
  for graph in graphs do
    initialize vertexToDelete
    for vertex in graph do
      if vertex is not a candidate then
        initialize vertexCheck
        for edges of vertex do
          if vertex1 notEqual vertex AND vertex1 is candidate then
            vertexCheck.add(vertex1)
          end if
          if vertex2 notEqual vertex AND vertex2 is candidate then
            vertexCheck.add(vertex2)
          end if
        end for
        if size of vertexCheck $< 2$ then
          vertexToDelete.add(vertex)
        end if
      end if
    end for
    graph.removeAllVertices(vertexToDelete)
    chosenURIs = calculateCentrality(measure, graph)
  end for
  return  chosenURIs, chosen candidate per mention
```

evaluate the performance of the algorithm, the linking is performed on correctly identified and classified authors.

6.1 Experiment settings

The test corpus consists of a French text of literary criticism titled "Une thése sur le symbolisme" (A thesis about Symbolism) and it is the first volume of the work named "Réflexions sur la littrature" (Reflexions on literature) published by Albert Thibaudet in 1938.

The text is drawn from a larger "Corpus critique"[11], published in TEI by the Labex OBVIL and containing a large collection of critical essays by different authors.

The chosen text in particular presents a high density of authors' mentions, so that each paragraph generally contains an average of 2-3 mentions that are treated at the same time by the algorithm. Mentions concerning authors were manually annotated by two experts in French literature; URIs assigned to mentions are those from Idref[12]. Guidelines to manual annotation were those proposed by the MUC7 conferences as well as those defined by the XML/TEI standard. The resulting test corpus contains 1021 manually annotated mentions of

[11] http://obvil.paris-sorbonne.fr/corpus/critique/

[12] www.idref.fr

person entities. We measure the precision of the proposed NEL approach in terms of the attribution of the right URI to a mention with respect to the URI manually assigned by humans. The authors lookup dictionary was automatically built in advance thanks to the BnF LOD source which is rich in SameAs predicates pointing to DBpedia and Idref URIs. The resulting lookup dictionary is composed of 4,218,798 author names including their alternative names (e.g. M. Lamartine, Monsieur Lamartine, etc.). We chose 3 centrality measures commonly used in social network analysis and the word-sens disambiguation problem, these are: *DegreeCentrality*[3], *BrandesBetweennessCentrality*[2], *FreemanClosenessCentrality*[3], as implemented in the JgraphT-SNA tool.

6.2 Results and Analysis

The test results with the three algorithm are shown in table 1,

Table 1. Results with different centrality measures on test corpus.

Centrality Measure Used	Precision	Unassigned Links
DegreeCentrality	0.73	23
BrandesBetweennessCentrality	**0.74**	23
FreemanClosenessCentrality	0.43	23

Precision is calculated comparing the number of correctly assigned links over the total of manually annotated entities of authors. The best result is obtained with BrandesBetweennessCentrality, with a precision of 0.74. DegreeCentrality has a comparable performance, FreemanCloseness centrality seems to heavily underperform with respect to the other centrality measures. The last column of table 1 shows the number of empty links over the total.

These first results are satisfying: though far from the 85% accuracy that is normally achieved by similar algorithms on the news domain, such levels of precision are nevertheless remarkable, considering that in many cases the text discusses minor authors, today unknown, that are not necessarily listed in DBpedia. Moreover, the use of BnF makes the number of candidates (and thus the possibility of error) explode, with sometimes as much as 20 or more possible candidate for a mention.

To quantify authors incompleteness in both the DBpedia and BnF data sets used in this experiment, we count the number of mentions in which the algorithm (using DegreeCentrality measure) does not find any corresponding URI in the chosen KB. In this manner, there are 160 author mentions, out of 1021 mentions identified in the corpus by the algorithm, that have no match in DBpedia, that is around 16%. Remarkably, there are only 23 mentions (i.e. 2%) that have no match in either BnF or DBpedia. Notice that all authors in this test set that are in DBpedia are also in BnF.

The most frequent mistakes considering DegreeCentrality and BrandesBetweennessCentrality measures (the most similar and precise ones) concern the following authors: Vielé-Griffin, Francis (1864-1937); Boileau, Nicolas (1636-1711); Barrès, Maurice (1862-1923); Payen, Fernand (1872-1946); Lefranc, Abel (1863-1952); Shakespeare, William (1564-1616); Spencer, Herbert (1820-1903); Goncourt, Edmond de (1822-1896) and brother Goncourt, Jules de (1830-1870); Mentré, Franois (1877-1950). The algorithm makes three types of mistakes.

MISSING CANDIDATES - In 23 cases the algorithm is unable to retrieve any candidate from the lookup dictionary, since the author is not present in any knowledge base. This is the case of author Francis Vielé-Griffin. In other cases the correct entity is present but not associated with the required pseudonym. This is the case of William Shakespeare's alleged alter ego William Stanley[13]. This alias is not listed in the dictionary for Shakespeare, therefore, it is not possible to assign both mentions to the same person (and thus the same URI).

MISSING CONTEXT - In some rare cases only one ambiguous author's mention is present in a single paragraph, thus the algorithm resorts to a fall back strategy, choosing the entity with more links in absolute. Sometimes this strategy causes errors, as in the case of "Vigny", for whom, in isolation, the wrong link to Auriane Vigny is chosen.

INCOMPLETE INFORMATION - In some cases the context of the sentence should be sufficient to produce a correct disambiguation but the NEL algorithm makes mistakes due to lack of links in the knowledge base, which prevents the centrality measure to produce the desired result. For instance, "Shakespeare", when mentioned in the context of Shakespearian critic Abel Lefranc, should produce the correct linking to William, but Nicolas is chosen instead. Clearly explicit links between Abel Lefranc and the object of his studies are missing in the knowledge bases. Ancient authors also tend to cause problems due to lack of information, e.g. the Greek author Lysias is mistaken for an homonymous French revolutionary collective.

WRONG, MISLEADING INFORMATION - Sometimes the knowledge bases contain wrong or misleading information. For instance there exist a BnF entry for the "Ronsard family", classified as foaf:Person, which is chosen instead the correct assignment, namely one of its members, Pierre de Ronsard. The opposite is also true, so some mentions refer to both Goncourt brothers as a collective noun, but the algorithm chooses one of the two. Finally, wrong or misleading pseudonyms are sometimes listed in BnF for an author, causing wrong candidates to be injected in the graph and sometimes selected. So for instance "Descartes" is listed as a pseudonym for novelist Horace Walpole and thus sometimes Walpole is wrongly chosen as the link for philosopher Descartes.

Error analysis also shows that sometimes relevant information that is present in the knowledge base is not used in the decision process because it cannot be encoded in the graph in the form of links. A typical example is temporal information which is encoded in the form of dates (data-typed literals). In other words the fact that - for a given context - two candidate referents lived in the same period of time cannot be taken into account.

[13] Stanley is believed by some to be the real author behind Shakespeare's works.

To evaluate the impact of the temporal dimension, we chose to evaluate against an index from which we removed authors born after the date of publishing of the work. The results show a slight improvement with **DegreeCentrality reaching precision 0.78** and BrandesBetweennessCentrality 0.77. A greater improvement may be obtained using a more sophisticated graph building algorithm, that transforms information about dates of birth and death in links that can connect authors in a measurable way.

7 Conclusion and future work

We presented an algorithm to perform NEL on a corpus of 19th century literary criticism, with the specific goal of disambiguating and referencing author mentions for research purposes. The NEL module is meant to be used in combination with a NER module, and will help researchers in the creation of digital literary editions enriched with information about authors. The main purpose of this work is to help scholars in history of literature to perform complex queries in order to study the literary appreciation of authors over time, and investigate the history of literary criticism in French literature. More specifically the enrichment of the aforementioned "Corpus critique" is meant to enhance ongoing research in the history of scientific ideas, and to provide a way to follow the dissemination of theories and concepts defined by Charles Darwin, Claude Bernard, Henri Bergson in non scientific texts of their time.

The reported experiment shows how combining different sources can be useful to perform linking on a domain specific corpus with satisfying results. While the precision is not yet state of the art, it is nevertheless remarkable, considering that it is the first time graph that centrality algorithms have been used for NEL combining DBpedia with a domain specific source. Tests showed significant differences between one implementation of centrality and the other two. Error analysis suggests possible improvements of the algorithm, including the ad hoc transformation of temporal information - present in the knowledge base in the form of literals - into links of the context graph. Another possible evolution of the algorithm would be to assign different weights to the edges so that for instance sharing the same literary circle becomes a more important relation than being born in the same town. Weights would be learned from manually annotated data. Further experiments will be carried out with different corpora and on different categories of entities, notably places.

Experimenting with the size of the context will also be necessary, in order to find the best trade-off between efficiency and informativeness. A more ample context (ideally a whole chapter) may produce a better graph of candidates, such that all mentions can disambiguate each other correctly. But at the same time this may introduce noise, and also generate a graph so big that its construction and the calculation of centrality may require too much time.

Another possible evolution of the algorithm could be to improve the graph fusion procedure. So far, our strategy does not handle the proper fusion of individuals which are described heterogeneously by the different sources (e.g. Victor

Hugo as described by BnF, as described by DBpedia, and so on). In this study we chose to study the problem from a quantitative point-of-view and thus to consider existent knowledge as it is without a pre-processing step. In the future, we foresee to make use of strategies commonly applied in Conceptual Graphs for information fusion [6]. In this way, the resulting graph would better concentrate domain knowledge (i.e. avoid redundancy and conflicts) and thus calculate a more accurate centrality measure.

Acknowledgements

This work was supported by French state funds managed by the ANR within the Investissements d'Avenir programme under reference ANR-11-IDEX-0004-02 and by an IFER Fernand Braudel Scholarship awarded by FMSH.

References

1. Bizer, C., Heath, T., Berners-Lee, T.: Linked data - the story so far. Int. J. Semantic Web Inf. Syst. 5(3), 122 (2009)
2. Brandes, U.: A faster algorithm for betweenness centrality. Journal of Mathematical Sociology 25(2), 163–177 (2001)
3. Freeman, L.C.: A set of measures of centrality based on betweenness. Sociometry pp. 35–41 (1977)
4. Hachey, B., Radford, W., Curran, J.R.: Graph-based named entity linking with wikipedia. In: Web Information System Engineering–WISE 2011, pp. 213–226. Springer (2011)
5. Hachey, B., Radford, W., Nothman, J., Honnibal, M., Curran, J.R.: Evaluating entity linking with wikipedia. Artificial intelligence 194, 130–150 (2013)
6. Laudy, C., Ganascia, J.G.: Information fusion using conceptual graphs: a tv programs case study. In: ICCS. pp. 158–165 (2008)
7. Mendes, P.N., Jakob, M., García-Silva, A., Bizer, C.: Dbpedia spotlight: shedding light on the web of documents. In: Proceedings of the 7th International Conference on Semantic Systems. pp. 1–8. ACM (2011)
8. Nadeau, D., Sekine, S.: A survey of named entity recognition and classification. Lingvisticae Investigationes 30(1), 3–26 (2007)
9. Rao, D., McNamee, P., Dredze, M.: Entity linking: Finding extracted entities in a knowledge base. In: Multi-source, Multilingual Information Extraction and Summarization, pp. 93–115. Springer (2013)
10. Rochat, Y.: Character Networks and Centrality. Ph.D. thesis, University of Lausanne (2014)
11. Sinha, R.S., Mihalcea, R.: Unsupervised graph-basedword sense disambiguation using measures of word semantic similarity. In: ICSC. vol. 7, pp. 363–369 (2007)
12. Van Hooland, S., De Wilde, M., Verborgh, R., Steiner, T., Van de Walle, R.: Exploring entity recognition and disambiguation for cultural heritage collections. Literary and linguistic computing (2013)