CorZoo: An Open TEI Corpus and Knowledge Graph for Semantic Analysis of Ancient Zoological Texts

Hajer Akid[1], Gabriel Loirat[1,4], Léa-Cassandre Tir[2,3], Kossi Zangbe[2,5], Catherine Faron[1] and Arnaud Zucker[2]

[1]*Université Côte d'Azur, Inria, CNRS, I3S (UMR 7271), Sophia Antipolis, France*
[2]*Université Côte d'Azur, CEPAM (UMR 7264), Nice, France*
[3]*École nationale des chartes – PSL, Paris, France*
[4]*Polytech Nice Sophia, Université Côte d'Azur, Sophia Antipolis, France*
[5]*Centre d'Études Supérieures de la Renaissance (CESR), Université de Tours, Tours, France*

Abstract

The IRN Zoomathia investigates the constitution and transmission of zoological knowledge from Antiquity to the Middle Ages. This paper introduces the corpus (CorZoo), resources, and semantic workflow developed within this framework to support structured corpus exploration. CorZoo comprises 89 texts encoded in TEI P5/XML under a unified schema, including 35 ancient Greek and 10 Latin source texts together with 44 modern English translations. To enrich this corpus, we propose an enhanced two-stage semantic annotation pipeline combining named entity recognition and domain thesaurus matching, producing annotations linked to TheZoo, a SKOS thesaurus of 6,019 concepts dedicated to ancient and medieval zoology. These annotations are subsequently integrated into an RDF knowledge graph serialized in Turtle format, preserving textual structure and linking source texts, translations, and semantic annotations. The resulting knowledge graph contains 3,839,346 RDF triples, describes 140,393 annotations linking textual mentions to 1,990 distinct TheZoo concepts, and is accessible via a SPARQL endpoint and a dedicated web interface, enabling fine-grained semantic navigation of the corpus. Altogether, the corpus, annotation workflow, and knowledge graph constitute open and reusable resources for the semantic enrichment and exploration of ancient scientific corpora.

Keywords

Ancient Zoological Texts, TEI Corpus, Named Entity Recognition, Semantic Annotation, Knowledge Graph, Digital Humanities, History of biology

CEUR-WS.org/Vol-4320/paper10.pdf

1. Introduction

Over the past decade, digital textual resources have grown rapidly, and with them the methods available to study historical, cultural, and scientific knowledge. Semantic technologies, knowledge graphs, and natural language processing now sit alongside keyword-based search as part of the standard Digital Humanities toolkit, opening the way to concept-based retrieval, semantic linking, and cross-textual exploration [1].

The International Research Network (IRN) Zoomathia[1] investigates how zoological knowledge evolved and circulated across ancient and medieval textual traditions, building on these advances. Aristotle's *Historia Animalium*, Pliny the Elder's *Naturalis Historia*, Aelian's *De Natura Animalium*, and Oppian's *Halieutica*, among many others, are not isolated works: zoological observations, classifications, and interpretations circulated between them, were reused, adapted, and reinterpreted across authors and centuries [2, 3]. Tracing these circulations calls for philological expertise, but also for computational methods able to integrate and analyze a large, heterogeneous textual collection [4]. Several factors make

SWODCH'26: International Workshop on Semantic Web and Ontology Design for Cultural Heritage, Workshop days of ISWC 2026, Bari, Italy
hajer.akid@etu.univ-cotedazur.fr (H. Akid); gabriel.loirat@etu.univ-cotedazur.fr (G. Loirat); lea-cassandre.tir@chartes.psl.eu (L. Tir); kossi.zangbe@etu.univ-tours.fr (K. Zangbe); faron@i3s.unice.fr (C. Faron); arnaud.zucker@univ-cotedazur.fr (A. Zucker)
0000-0002-5976-9901 (H. Akid); 0009-0007-3325-5981 (G. Loirat); 0009-0009-0415-1133 (L. Tir); 0009-0004-6725-9743 (K. Zangbe); 0000-0001-5959-5561 (C. Faron); 0000-0002-4382-5112 (A. Zucker)
© 2026 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

[1]https://www.cepam.cnrs.fr/sites/zoomathia/

CEUR Workshop Proceedings ceur-ws.org ISSN 1613-0073
published 2026-11-09

this computational work difficult. Many of the relevant works have already been digitized, through initiatives such as the Perseus Digital Library [5, 6] or the Thesaurus Linguae Graecae [7], yet they remain scattered across repositories with inconsistent encoding practices, which hampers any attempt at corpus-level integration [8]. Ancient Greek and Latin add a further layer of difficulty, since standard natural language processing tools offer only limited support for these languages [9, 10]. Orthographic variation compounds the problem, as does the gap between ancient zoological vocabulary and modern biological categories: the same species can appear under different names depending on the author or the period, a single term can shift meaning across textual traditions, and zoological descriptions are frequently woven into mythological, medical, or geographical material that calls for domain-specific interpretation [2].

Keyword-based search cannot answer the kind of questions this raises. Historians and philologists working within Zoomathia ask questions such as: *Which passages across different authors describe the gestation period of the same animal? Which texts mention animals in a medical context? How is a given zoological concept transmitted from one author to another?* None of these can be resolved through lexical or quantitative retrieval alone; they require a semantic layer that links textual mentions to a shared conceptual model and supports reasoning across texts [11, 1]. TheZoo fulfills this role within Zoomathia: a SKOS thesaurus of 6,019 concepts modeling the taxonomic and thematic categories of ancient and medieval zoological knowledge [12].

This paper builds on TheZoo and on previous annotation experiments within Zoomathia [13, 14, 15] to present three resources. The first is a curated corpus of 89 TEI-encoded texts, comprising 45 ancient Greek and Latin source texts and 44 modern English translations, encoded in TEI P5/XML [16] under a unified schema, letting NLP tools work through translations instead of ancient languages. The second is an extended semantic annotation pipeline, building on our previous work [14, 15], combining named entity recognition via DBpedia Spotlight[2] and Entity-fishing[3] with complementary exact, lemma-based, and fuzzy matching strategies for alignment with TheZoo.

The third is an RDF knowledge graph serialized as Turtle files using Morph-xR2RML [17], integrating original texts, translations, and semantic annotations linked to TheZoo; it contains 3,839,346 RDF triples and 140,393 annotations linked to 1,990 distinct TheZoo concepts, queryable through a SPARQL endpoint and a dedicated web exploration interface.[4] All resources are openly available on GitHub[5] in line with FAIR data principles [18].

The remainder of this paper is organized as follows. Section 2 reviews related work. Section 3 describes the TEI corpus. Section 4 presents the semantic annotation workflow. Section 5 describes the semantic model of the knowledge graph and presents resource statistics. Section 6 illustrates competency questions and their SPARQL implementation. Section 7 presents the web exploration interface. Finally, Section 8 concludes the paper and outlines future work.

2. Related Work

2.1. Structuration and Annotation of Ancient Textual Resources

Early efforts in the digital structuration of ancient and medieval texts established foundations on which our work builds. SourceEncyMe[6] and Ichtya[7] pioneered the encoding of medieval encyclopedias in XML/TEI and their manual semantic annotation. The Perseus Digital Library [6] has digitized large collections of Greek and Latin texts in TEI format, providing open access to a wide range of classical works, partially complemented for Medieval Latin texts by Corpus Corporum.[8] Perseus texts constitute

[2] https://www.dbpedia-spotlight.org/
[3] https://github.com/kermitt2/entity-fishing
[4] http://zoomathia.i3s.unice.fr
[5] https://github.com/Wimmics/zoomathia
[6] http://sourcencyme.irht.cnrs.fr
[7] http://www.unicaen.fr/recherche/mrsh/document_numerique/projets/ichtya
[8] https://mlat.uzh.ch/

the primary source for a significant portion of our corpus. Since 1985, the Thesaurus Linguae Graecae program[9] [7] has developed a wide corpus of ancient Greek texts; more recently, the Open Greek and Latin initiative[10] has contributed to extending the availability of ancient Greek texts in TEI format, further enriching the sources from which our corpus is drawn. The TEI Guidelines themselves [16] provide the encoding standard that underlies the structural and editorial choices made throughout our corpus.

2.2. Knowledge Graphs for Cultural Heritage

The construction of knowledge graphs from cultural and historical resources has attracted growing interest in the Digital Humanities community. Hyvönen's work on the Sampo model [11, 1, 19] introduced a framework for building semantic portals dedicated to cultural heritage data and demonstrated its application across multiple domains. This work highlights the importance of combining structured vocabularies, linked data principles, and user-oriented exploration interfaces, all of which are central to our approach. Colla et al. [20] proposed an ontology-based platform to help users identify and characterize entities in historical archives using automatic information extraction techniques. The ISSA project [21] presented a pipeline for the semantic indexing of a large open scientific archive, combining named entity recognition and linking strategies similar in spirit to those we adapt to the specific challenges of ancient Greek and Latin texts.

2.3. Semantic Annotation of Ancient Zoological Texts

Within the Zoomathia project itself, several works have progressively built the foundations of our current contribution. Tounsi et al. [13] presented an early approach to studying the history of premodern zoology using linked data and controlled vocabularies, establishing the role of TheZoo as the central domain thesaurus. Building on the thesaurus described by Pajón Leyra et al. [12], Barbe et al. [14] described the construction of a knowledge graph from manual annotations of texts from Pliny the Elder's *Naturalis Historia*, demonstrating the feasibility of formalizing expert annotations as RDF triples linked to TheZoo. Tounsi Dhouib et al. [15] extended this work toward automatic annotation, applying named entity recognition tools to a larger subset of the corpus; however, since ancient Greek and Latin texts cannot be processed efficiently by NLP tools, the extraction was performed on automatic translations of the source texts, which inevitably introduces noise and limits the quality of the resulting annotations. The present paper builds on and significantly extends these contributions in three concrete directions, moving from single-author manual annotation [14] and single-strategy automatic annotation on a partial corpus [15] to a full-corpus, multi-strategy pipeline: the corpus now covers 89 texts with systematic TEI P5 encoding, including 44 aligned human English translations, rather than a single work; the annotation pipeline is enriched with additional thesaurus matching strategies, including lemma-based and fuzzy matching, leveraging the Classical Language Toolkit [22] for lemmatization, rather than a single matching method; and the resulting RDF knowledge graph is published as a set of Turtle files, generated using the Morph-xR2RML mapping engine [17], constituting an open, reusable benchmark for the community.

3. The CorZoo TEI Corpus

The CorZoo corpus was assembled by domain experts who selected texts relevant to the study of zoological knowledge in Antiquity and the Middle Ages. Texts were sourced from established digital repositories, primarily the Perseus Digital Library [6], complemented by the Thesaurus Linguae Graecae

[9] http://stephanus.tlg.uci.edu
[10] https://www.opengreekandlatin.org/

(TLG)[11] [7], the Latin Library,[12] and Corpus Corporum[13] for Medieval Latin texts. The current corpus comprises **89 texts**, distributed as shown in Table 1: 35 in ancient Greek, 10 in Latin, and 44 accompanied by a modern English translation. Greek and Latin texts without an available human translation are either medieval texts not yet translated, short ancient texts, or texts whose translation is currently being acquired; for these, the annotation pipeline falls back to automatic translation, as described in Section 4.

Table 1
Corpus composition by language

Language	Number of texts
Ancient Greek	35
Latin	10
English translations (of Greek texts)	35
English translations (of Latin texts)	9
Total	**89**

This composition reflects a deliberate editorial choice rather than an accident of availability. The texts span a wide range of authors and genres relevant to ancient zoology: natural history (Aristotle, Pliny the Elder, Aelian), geography and history (Herodotus, Diodorus Siculus), agriculture (Columella, Xenophon), hunting (Oppian of Apamea, Arrian), medicine (Hippocrates), fables (Aesop, Phaedrus, Babrius), and philosophy (Plato, Philo Judaeus, Plutarch). A number of medieval texts have also been included, though a systematic inventory of relevant medieval sources remains to be carried out; many of these texts are not yet critically edited or are only available in manuscript form, which makes their integration more challenging. This diversity allows the project to trace how zoological knowledge was treated differently depending on disciplinary context, from the systematic classification found in natural history treatises to the moral allegory underlying fables and philosophical writing. The presence of English translations for only a subset of the corpus is itself a meaningful constraint, addressed directly in the annotation pipeline described in Section 4, since ancient Greek and Latin remain comparatively under-resourced for natural language processing [9].

All texts are encoded in TEI P5/XML [16] following a shared editorial schema, which ensures interoperability with the wider Digital Humanities community [8]. The choice of TEI rather than relying exclusively on direct NLP processing of the source texts is motivated by the need for a stable representation that preserves textual structure independently of any particular NLP pipeline. This is especially important for ancient textual resources, for which computational processing remains challenging because of linguistic variation, limited training data, and the heterogeneous state of digitized sources [23, 24]. TEI also provides a persistent structural layer onto which different annotation methods can subsequently be applied and compared.

The structural hierarchy of each work, from book and chapter down to paragraph level, is preserved through nested `<div>` elements, while metadata such as author, title, date, and source repository are recorded in the TEI header. This uniform encoding is what makes the corpus usable as a single, queryable object rather than a loose collection of digitized texts: it allows paragraphs from very different authors, genres, and centuries to be addressed, compared, and annotated through the same structural vocabulary. Listing 1 shows a simplified excerpt of this encoding for a passage of Aristotle's *Historia Animalium*.

A further methodological choice underlies the corpus design: wherever a modern English translation exists, it is encoded as a separate TEI file sharing the exact same structural schema as the source text, with paragraphs aligned through matching identifiers. This alignment serves a dual purpose. For human scholars, it allows direct navigation between an ancient passage and its translation. For the annotation pipeline, it provides the means to process texts written in ancient Greek and Latin, languages

[11] http://stephanus.tlg.uci.edu
[12] https://www.thelatinlibrary.com/
[13] https://mlat.uzh.ch/

Listing 1: Abridged TEI encoding of D'Arcy Wentworth Thompson's translation of Aristotle's *Historia Animalium* (corpus file `zoo7/13e`, Book 1, Chapter 1)

```
<TEI xmlns="http://www.tei-c.org/ns/1.0">
 <teiHeader>
  <fileDesc>
   <titleStmt>
    <title>Historia Animalium</title>
    <author ref="tlg0086">Aristotle</author>
    <respStmt>
     <resp>Translator</resp>
     <persName>D'Arcy Wentworth Thompson</persName>
    </respStmt>
   </titleStmt>
  </fileDesc>
 </teiHeader>
 <text>
  <body>
   <div type="book" subtype="book" n="1">
    <div type="chapter" subtype="chapter" n="1">
     <p>
      [...] Differences are manifested in modes of subsistence, in habits, in actions performed. For
           instance, some animals live in water and others on land. And of those that live in water
           some do so in one way, and some in another: that is to say, some live and feed in the water
           , take in and emit water, and cannot live if deprived of water, as is the case with the
           great majority of fishes; others get their food and spend their days in the water, but do
           not take in water but air, nor do they bring forth in the water. [...]
     </p>
    </div>
   </div>
  </body>
 </text>
</TEI>
```

that current NLP tools cannot handle efficiently, by using the aligned English translation as input for semantic analysis. The corpus is therefore not only a textual resource but the structural foundation on which the semantic annotation described in the following sections is built.

4. Annotation Pipeline

The construction of the Zoomathia knowledge graph relies on two complementary annotation strategies, extending the methodology initially explored in [14, 15]. Figure 1 provides an overview of this pipeline.

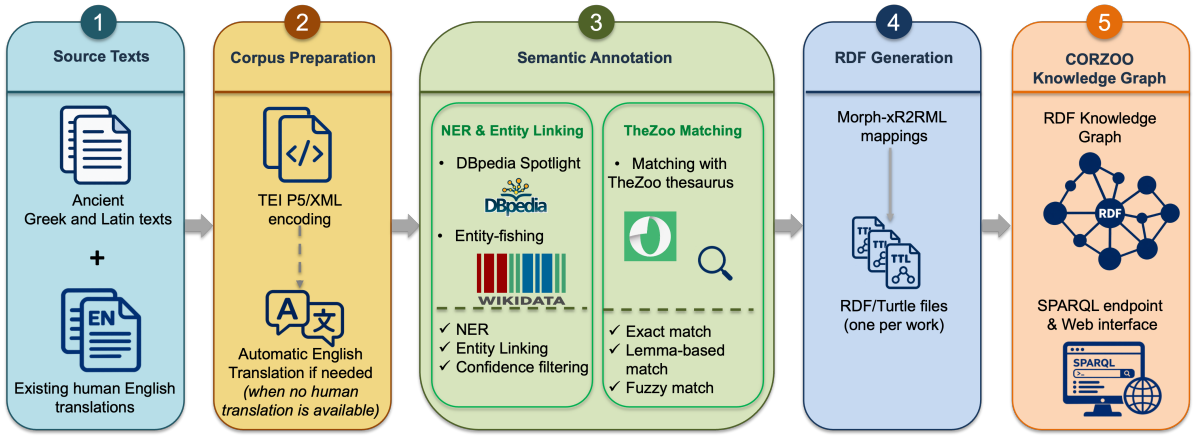


Figure 1: Overview of the CorZoo annotation pipeline.

Since ancient Greek and Latin texts cannot be processed directly by most NLP tools, the pipeline first relies on English translations, either the human translations available in the corpus or, when none

exists, an automatic translation of the source text. Two named entity recognition tools are then applied to this English text: Entity-fishing, which links mentions to Wikidata, and DBpedia Spotlight, which links them to DBpedia. A confidence threshold of 0.6 is applied to discard low-confidence annotations, and a filter list excludes DBpedia categories irrelevant to the zoological domain, such as musical works, companies, or television shows.

In parallel with the annotations produced by the general-purpose entity recognition tools, the English text is directly matched against TheZoo to identify domain-specific zoological concepts that may not be detected by Entity-fishing or DBpedia Spotlight. This thesaurus-based annotation relies on three complementary strategies. Exact matching searches for concepts whose preferred or alternative label occurs in the text, in any of the languages covered by the thesaurus. Lemma-based matching uses the Classical Language Toolkit [22] to account for inflectional variation, while fuzzy matching, with a similarity threshold of 0.85, captures minor spelling variations. When the same concept is identified more than once for a given mention, duplicate annotations are removed.

The enriched annotations are finally converted into RDF using the Morph-xR2RML mapping engine [17]. The pipeline generates, for each work, an individual Turtle file containing its textual structure, content, and semantic annotations, which can be inspected, validated, or reprocessed independently. These per-work files are collectively available on the project repository and can be loaded together to constitute the unified knowledge graph queried through the SPARQL endpoint and the web interface described in the following sections.

The annotation pipeline is applied to the English version of each text. When a human English translation is available, it is used directly; otherwise, the source text is automatically translated into English using Google Translate. Automatic translation therefore constitutes a potential source of uncertainty, since mistranslations or context-dependent lexical choices may affect the subsequent annotation process. A qualitative inspection of selected cases illustrates this issue. For example, the term corresponding to "ox" was correctly translated when considered in isolation, but was translated as "mouth" in its sentence context, preventing its identification as a zoological concept. This example shows how translation errors can directly affect downstream semantic annotation. A systematic comparison of annotations obtained from human and automatically translated versions of the same passages, together with expert assessment of annotation correctness and completeness on a representative sample, is needed to better quantify the impact of translation quality and is left for future work.

5. Semantic Model of the Knowledge Graph

The CorZoo knowledge graph builds on standard Semantic Web vocabularies to represent both the structure and the semantic content of the corpus. A domain vocabulary `zoo:`, aligned with Schema.org[14], captures the hierarchical structure of ancient works from the work level down through book, chapter, section, and paragraph, with each level linked to its subdivisions through `isPartOf` and `hasPart` properties. Paragraph text is stored in the original language and, where available, in modern English translation.

Annotations are expressed through the W3C Web Annotation Vocabulary[15]: each `oa:Annotation` links a body, a TheZoo concept, to a target, the annotated text span within a paragraph, recording both the exact wording and its character position. The annotation layer remains independent of the underlying text, allowing paragraphs to be corrected or re-annotated without interference, and enabling traversal in either direction, from a paragraph to the concepts it mentions or from a concept to every paragraph where it occurs across the corpus. Figure 2 illustrates this representation through an example subgraph, including the connection of an annotated mention to the SKOS hierarchy of TheZoo.

TheZoo [12] provides the domain-specific grounding for these annotations. Built specifically for ancient and medieval zoological knowledge, it comprises 6,019 SKOS concepts organized into 11 hierarchical levels and 14 thematic collections, covering animal names, anatomical terms, geographic

[14] https://schema.org/
[15] https://www.w3.org/TR/annotation-vocab/

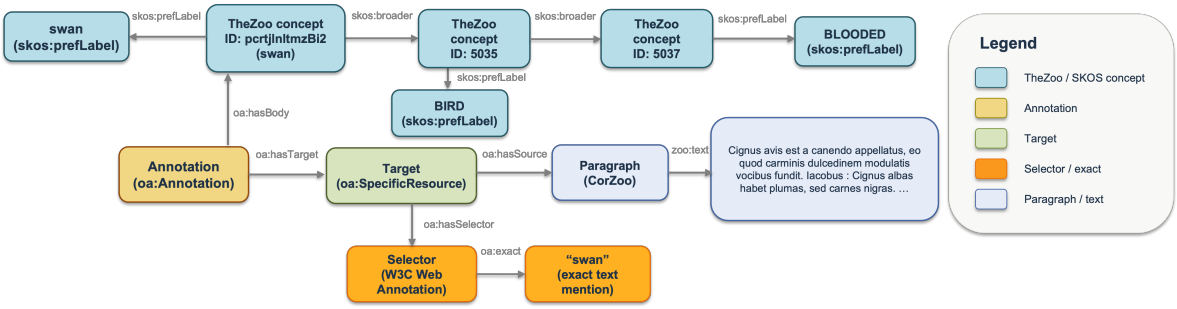


Figure 2: Example CorZoo subgraph showing the annotation of the textual mention "swan" and its connection to the TheZoo SKOS hierarchy.

spaces, and archaic taxonomic categories in Greek, Latin, English, French, and Italian. It is this thesaurus that makes the historically grounded semantic queries presented later in this paper possible.

Table 2 summarizes the resulting thesaurus and knowledge graph statistics, building on the 89-text corpus presented in Section 3.

Table 2
Statistics of the TheZoo thesaurus and the resulting knowledge graph

Resource	Count
TheZoo Thesaurus	
SKOS concepts	6,019
Hierarchical levels	11
Thematic collections	14
Knowledge Graph	
RDF triples	3,839,346
Annotations	140,393
Distinct TheZoo concepts used	1,990
Top annotated works (original language)	
Pliny, *Naturalis Historia* (Latin)	65,039
Athenaeus, *Deipnosophists* (Greek)	26,878
Herodotus, *Histories* (Greek)	20,676
Diodorus Siculus, *Library* (Greek)	16,670
Aelian, *On the Characteristics of Animals* (Greek)	13,608

The distribution of annotations reflects both the size and the zoological density of each text. Pliny the Elder's *Naturalis Historia* yields the highest number of annotations (65,039), consistent with its encyclopedic scope; the English translation available in the corpus covers only a subset of the books, which explains the lower annotation count for the English version. Aelian's *On the Characteristics of Animals* also ranks among the most richly annotated works (13,608), as expected from its explicit zoological focus.

6. Competency Questions and SPARQL Validation

The knowledge graph was designed to answer research questions formulated by historians and philologists. We present below two representative competency questions (CQs) and simplified versions of their SPARQL implementations. The numbering used here is specific to this paper; the corresponding questions can be identified in the web interface by their full titles.

CQ1. Which paragraphs mention an animal and its habitat in the same passage?

```
PREFIX oa: <http://www.w3.org/ns/oa#>
PREFIX skos: <http://www.w3.org/2004/02/skos/core#>

SELECT DISTINCT ?paragraph ?animal ?habitat WHERE {
  ?ann1 oa:hasBody ?animalConcept ;
    oa:hasTarget [ oa:hasSource ?paragraph ;
      oa:hasSelector [ oa:exact ?animal ] ] .

  ?ann2 oa:hasBody ?habitatConcept ;
    oa:hasTarget [ oa:hasSource ?paragraph ;
      oa:hasSelector [ oa:exact ?habitat ] ] .

  <https://opentheso.huma-num.fr/?idg=MT_10&idt=th310>
    skos:member ?animalConcept .

  ?habitatConcept skos:broader+
    <https://opentheso.huma-num.fr/?idc=105621&idt=th310> .
} ORDER BY ?paragraph
```

This query retrieves paragraphs in which an animal concept and a concept related to habitat co-occur, illustrating how TheZoo's semantic organization can be used to retrieve passages beyond simple keyword matching.

CQ2. Which data is available on the gestation periods of animals?

```
PREFIX oa: <http://www.w3.org/ns/oa#>
PREFIX skos: <http://www.w3.org/2004/02/skos/core#>
PREFIX zoo: <http://ns.inria.fr/zoomathia/zoo#>

SELECT DISTINCT ?paragraph ?text ?animal WHERE {
  ?ann1 oa:hasBody ?animalConcept ;
    oa:hasTarget [ oa:hasSource ?paragraph ;
      oa:hasSelector [ oa:exact ?animal ] ] .

  ?ann2 oa:hasBody
    <https://opentheso.huma-num.fr/?idc=105364&idt=th310> ;
    oa:hasTarget [ oa:hasSource ?paragraph ] .

  ?paragraph zoo:text ?text .

  <https://opentheso.huma-num.fr/?idg=MT_10&idt=th310>
    skos:member ?animalConcept .
} ORDER BY ?paragraph
```

This query retrieves passages in which an animal concept co-occurs with the TheZoo concept corresponding to gestation period, allowing users to identify and compare information about animal gestation across the corpus.

The full set of competency questions and their SPARQL implementations is available on the project GitHub repository.[16]

7. Web Exploration Interface

The CorZoo resources are made accessible through a dedicated web application at http://zoomathia.i3s.unice.fr, developed in JavaScript with a React.js frontend and an Express.js backend. All data is retrieved from the knowledge graph via parameterized SPARQL queries.

The interface offers three complementary exploration modes:

[16] https://github.com/Wimmics/zoomathia

1. **Explore a work**: The user selects an author and work. The interface displays the table of contents, the paragraph text with annotated mentions highlighted, and the list of annotated concepts. Clicking on a concept redirects to its entry in the TheZoo thesaurus.
2. **Explore the corpus**: The user defines a custom sub-corpus by filtering on author, work, and/or concepts, in any language supported by the thesaurus. The interface retrieves all matching paragraphs across the corpus, supporting both OR and AND search modes.
3. **Competency questions**: A dedicated tab provides access to competency questions formulated by domain experts and implemented as SPARQL queries. The two questions presented in Section 6 are examples from this collection and can be identified in the interface by their full titles. Results can be visualized as tables or as interactive graphs using the MGExplorer tool, enabling the exploration of conceptual co-occurrences across texts.

This design follows the principle, central to the Sampo model [19], of hiding the complexity of SPARQL behind task-oriented exploration interfaces accessible to non-expert users.

8. Conclusion and Future Work

This paper has presented the CorZoo open resources: a TEI P5 corpus comprising 45 ancient Greek and Latin source texts and 44 modern English translations, a two-stage semantic annotation pipeline, and an RDF knowledge graph of 3,839,346 triples describing 140,393 annotations linked to 1,990 distinct TheZoo concepts. Together, these resources constitute an original contribution to Digital Humanities and to the Semantic Web for Cultural Heritage community, and offer a reusable benchmark for semantic annotation of ancient scientific texts.

Several directions are pursued for future work. We are currently expanding the corpus with additional texts. The current pipeline relies on DBpedia Spotlight and Entity-fishing, two general-purpose named entity recognition tools trained on contemporary text, which consequently miss entities absent from their underlying knowledge bases or expressed through paraphrase rather than a named mention. We are therefore exploring Large Language Model (LLM)-based annotation as a complementary strategy, prompting an LLM to identify zoological mentions directly against the TheZoo concept inventory; this is expected to improve recall on entities currently unmatched by the NER pipeline, though it will require careful evaluation against expert annotations before integration into the knowledge graph. The web interface will also be extended to display the original text and its English translation side by side, easing navigation between the source language and the version processed by the pipeline. Finally, word-embedding-based similarity will be explored to identify candidate concepts for thesaurus enrichment, linking Wikidata entities to TheZoo via semantic proximity.

Acknowledgments

This work was carried out in the context of the IRN Zoomathia, supported by the CNRS. The internships of Gabriel Loirat and Léa-Cassandre Tir were funded by the HisINum project (Académie 5 Idex, UniCA). The authors thank Molka Tounsi Dhouib and Arnaud Barbe for their earlier contributions to the Zoomathia knowledge graph, and the members of the GDRI Zoomathia for their domain expertise and for the competency questions that guided the design of the resources.

Declaration on Generative AI

Generative AI tools were used only for limited language and drafting assistance. All content was subsequently reviewed, revised, and validated by the authors.

References

[1] E. Hyvönen, Using the semantic web in digital humanities: Shift from data publishing to data-analysis and serendipitous knowledge discovery, Semantic Web 11 (2020) 187–193.

[2] A. Zucker, Les classes zoologiques en Grèce ancienne, Publications de l'Université de Provence, 2005.

[3] S. T. Newmyer, Animals in Greek and Roman Thought: A Sourcebook, Routledge, 2011.

[4] R. Siemens, S. Schreibman, A companion to digital literary studies, John Wiley & Sons, 2013.

[5] Perseus Digital Library, Perseus digital library, 2026. URL: http://www.perseus.tufts.edu, accessed: July 1, 2026.

[6] G. Crane, Building a digital library: The perseus project as a case study in the humanities, in: Proceedings of the first ACM international conference on Digital libraries, 1996, pp. 3–10.

[7] Thesaurus Linguae Graecae, Thesaurus linguae graecae, 2026. URL: http://stephanus.tlg.uci.edu, accessed: July 1, 2026.

[8] L. Burnard, What is the Text Encoding Initiative?: How to add intelligent markup to digital resources, 3, OpenEdition Press, 2014.

[9] D. Bamman, G. Crane, The ancient greek and latin dependency treebanks, in: Language technology for cultural heritage: Selected papers from the LaTeCH Workshop Series, Springer, 2011, pp. 79–98.

[10] R. Sprugnoli, F. Iurescia, M. Passarotti, Overview of the evalatin 2024 evaluation campaign, in: Proceedings of the Third Workshop on Language Technologies for Historical and Ancient Languages (LT4HALA)@ LREC-COLING-2024, 2024, pp. 190–197.

[11] E. Hyvönen, Publishing and using cultural heritage linked data on the semantic web, volume 3, Morgan & Claypool Publishers, 2012.

[12] I. P. Leyra, A. Zucker, C. F. Zucker, TheZoo: un thesaurus de zoologie ancienne et médiévale pour l'annotation de sources de données hétérogènes, Archivum Latinitatis Medii Aevi 73 (2015) 321–342.

[13] M. Tounsi, C. F. Zucker, A. Zucker, S. Villata, E. Cabrio, Studying the history of pre-modern zoology with linked data and vocabularies, in: First International Workshop Semantic Web for Scientific Heritage (SW4SH) at ESWC 2015, 2015.

[14] A. Barbe, M. Tounsi Dhouib, C. Faron, M. Corneli, A. Zucker, Zoo-kg: A knowledge graph of zoological data, in: E. Carbé, G. Lo Piccolo, A. Valenti, F. Stella (Eds.), La memoria digitale: forme del testo e organizzazione della conoscenza – Atti del XII Convegno Annuale AIUCD, AIUCD, Siena, 2023. doi:10.6092/unibo/amsacta/7721.

[15] M. T. Dhouib, A. Barbe, C. Faron, A. Zucker, M. Corneli, Graphe de connaissance outillé au service des chercheurs en histoire de la zoologie antique et médiévale, Revue Ouverte d'Intelligence Artificielle 6 (2025) 85–106. doi:10.5802/roia.94.

[16] TEI Consortium, TEI P5: Guidelines for Electronic Text Encoding and Interchange, 2023. URL: https://tei-c.org/guidelines/p5/, accessed: July 1, 2026.

[17] F. Michel, L. Djimenou, C. F. Zucker, J. Montagnat, Translation of relational and non-relational databases into RDF with xR2RML, in: Proceedings of the 11th International Conference on Web Information Systems and Technologies (WEBIST 2015), 2015, pp. 443–454.

[18] M. D. Wilkinson, et al., The FAIR guiding principles for scientific data management and stewardship, Scientific Data 3 (2016) 160018. doi:10.1038/sdata.2016.18.

[19] E. Hyvönen, Digital humanities on the semantic web: Sampo model and portal series, Semantic Web 14 (2023) 729–744.

[20] D. Colla, A. Goy, M. Leontino, D. Magro, C. Picardi, Bringing semantics into historical archives with computer-aided rich metadata generation, Journal on Computing and Cultural Heritage (JOCCH) 15 (2022) 1–24.

[21] A. Bobasheva, F. Michel, A. Tchechmedjiev, A. Toulet, Issa pipeline, Inria & Université Côte d'Azur, CNRS, I3S, 2024.

[22] K. P. Johnson, P. J. Burns, J. Stewart, T. Cook, C. Besnier, W. J. Mattingly, The classical language toolkit: An nlp framework for pre-modern languages, in: Proceedings of the 59th annual meeting

of the association for computational linguistics and the 11th international joint conference on natural language processing: System demonstrations, 2021, pp. 20–29.

[23] K. Papantoniou, Y. Tzitzikas, Nlp for the greek language: A brief survey, in: 11th hellenic conference on artificial intelligence, 2020, pp. 101–109.

[24] T. Sommerschield, Y. Assael, J. Pavlopoulos, V. Stefanak, A. Senior, C. Dyer, J. Bodel, J. Prag, I. Androutsopoulos, N. De Freitas, Machine learning for ancient languages: A survey, Computational Linguistics 49 (2023) 703–747.