Construction d'un graphe de connaissance à partir des annotations manuelles de textes de zoologie antique

Arnaud Barbe, Molka Tounsi Dhouib, Catherine Faron, Marco Corneli, Arnaud Zucker

Construction d'un graphe de connaissance à partir des annotations manuelles de textes de zoologie antique

A. Barbe[1], M. Tounsi Dhouib[2], C. Faron[2], M. Corneli[1], A. Zucker[1]

[1] Université côte d'Azur, CEPAM, CNRS, France
[2] Université côte d'Azur, INRIA, CNRS, I3S, France

arnaud.barbe@univ-cotedazur.fr, dhouib@i3s.unice.fr, faron@i3s.unice.fr,
marco.corneli@univ-cotedazur.fr, Arnaud.zucker@univ-cotedazur.fr

Résumé

Ce travail est réalisé dans le cadre de l'IRN Zoomathia qui vise l'étude de la transmission des savoirs zoologiques de l'Antiquité au Moyen Âge. Dans ce contexte, un premier travail d'annotation manuelle de l'"Histoire Naturelle" de Pline l'Ancien sur la zoologie antique (livres 8-11) en utilisant des concepts rassemblés dans le thésaurus TheZoo a été réalisé par des spécialistes de l'Antiquité. Cependant, ces annotations ont été réalisées avec des commentaires Word, ce qui rend complexe l'exploitation de ces connaissances par les chercheurs épistémologues, historiens et philologues dans leur travail d'analyse de ces textes anciens. Dans cet article, nous présentons notre approche de transformation de ces annotations manuelles en graphe de connaissance RDF permettant l'intégration et l'interrogation des connaissances extraites dans le but d'aider les chercheurs dans leur travail d'analyse de ces textes et de la transmission des connaissances à travers eux. Afin de valider la pertinence du modèle et le graphe de connaissance, nous avons recueilli auprès d'un expert du domaine un ensemble de questions de compétences que nous avons traduites en SPARQL pour y répondre en interrogeant le graphe de connaissance produit.

Mots-clés

Ontologie, Annotation sémantique de textes latins, Graphes de connaissances, Données liées et vocabulaires, Histoire de la zoologie.

Abstract

This work is carried out in the framework of the IRN Zoomathia which aims to study the transmission of zoological knowledge from Antiquity to the Middle Ages. In this context, a first work of manual annotation of Pline's Naturalis Historia (books 8-11) on ancient zoology using concepts gathered in the thesaurus TheZoo was done by classicists. However, these annotations have been stored as comments in Word documents, which complicates the exploitation of this knowledge by epistemology, history and philiology researchers in their analysis of these ancient texts. In this article, we present our approach to transform manual annotations into an RDF knowledge graph allowing the integration and the interrogation of relevant knowledge in order to support researchers in their analysis of these texts and knowledge transmission through them. In order to validate the relevance of the model as well as the knowledge graph, we elicited from a domain expert a set of competency questions that we translated in SPARQL to answer them by querying the knowledge graph.

Keywords

Semantic Annotation of Latin Texts, Knowledge Graphs, Ontologies, Linked Data and Vocabularies, History of Zoology.

1 Introduction

Les historiens et les philologues doivent faire face quotidiennement à une quantité énorme de ressources textuelles. Malgré les efforts de numérisation, les outils proposés ne répondent pas aux exigences épistémologiques en ne permettant souvent que des recherches lexicales et quantitatives des données. Les chercheurs expriment un besoin d'outils plus intelligents afin de réaliser des recherches plus élaborées qui nécessitent une annotation sémantique plus riche. Le Réseau de Recherche International (IRN) Zoomathia [1] vise l'étude de la constitution et de la transmission des connaissances zoologiques de l'Antiquité au Moyen Âge, à travers des ressources variées, et considère en particulier l'information textuelle. Dans ce contexte, un premier travail d'annotation manuelle de quatre chapitres de l'"Histoire Naturelle" de Pline a été réalisé par un chercheur en littérature latine. Ainsi, sous la forme de commentaires dans un document Word, chaque texte latin a été annoté avec les concepts du thésaurus TheZoo [2]. Malgré l'énorme effort réalisé et le temps passé à annoter ces textes, ces annotations restent inexploitables en termes de formalisation du savoir et d'intégration de ces connaissances avec d'autres sources de connaissances. L'objectif de notre travail est de transformer ces annotations manuelles en graphe de connaissance permettant ainsi l'intégration et l'interrogation des connaissances extraites dans le but de proposer des possibilités de recherche automatique plus riches et répon-

1. https ://www.cepam.cnrs.fr/sites/zoomathia/
2. https ://opentheso.huma-num.fr/opentheso/ ?idt=th310

dant mieux aux besoins des chercheurs qui étudient cette littérature scientifique. Nous avons identifié trois questions de recherche : (i) Quels types de connaissances devons-nous représenter afin d'aider les chercheurs dans leur travail d'analyse et de transmission de savoir zoologique ? (ii) Quelles ontologies existantes pouvons-nous réutiliser pour représenter ces documents ? (iii) Quelle approche pouvons-nous définir pour réutiliser les annotations manuelles faites par les linguistes et les rendre exploitables ?

Notre approche de construction du graphe de connaissance repose sur (i) la proposition d'un modèle qui réutilise des ontologies et vocabulaires existants afin de structurer et représenter les annotations manuelles des textes de zoologie ancienne, (ii) l'explicitation de questions de compétences auprès d'historiens et philologues intéressés par la transmission des connaissances zoologiques. Le processus de construction du graphe de connaissances comprend cinq étapes successives : (i) la reconnaissance des entités pertinentes dans les annotations manuelles, (ii) le liage de ces entités avec les concepts du thésaurus TheZoo, (iii) l'extraction des contenus textuels des chapitres et paragraphes du texte annoté, (iv) le liage des paragraphes avec les annotations, et enfin (v) la génération du graphe RDF capturant à la fois le contenu textuel et la structure de l'Histoire Naturelle de Pline et les annotations du texte à l'aide de l'outil morph-xr2rml [5].

Cet article est organisé comme suit. Dans la section 2, nous présentons une synthèse des approches de construction de graphe de connaissance à partir de textes anciens (médiévaux) ainsi que les vocabulaires réutilisés dans ce travail. Dans la section 3, nous présentons un ensemble de questions de compétences représentatives des besoins des experts en termes d'exploitation des annotations générées. La section 4 décrit le modèle sémantique du graphe de connaissance. Dans la section 5, nous détaillons le processus que nous avons utilisé pour la génération de ce graphe de connaissance. Enfin, dans la section 6 nous présentons des requêtes SPARQL qui implémentent des questions de compétences élicitées et dont la réponse peut être recherchée dans le graphe de connaissance produit, validant ainsi celui-ci.

2 État de l'art

2.1 Construction de graphes de connaissance à partir de textes anciens

Plusieurs travaux dans la littérature ont traité la problématique d'analyse et de structuration des ressources culturelles et historiques de l'Antiquité au Moyen Âge. Des premiers travaux de recherche français s'inscrivent dans les projets SourceEncyMe4 [3] et Ichtya5 [4] portant sur la structuration d'encyclopédies médiévales en XML selon le modèle TEI et l'annotation manuelle de ces sources de données.

D'autres travaux ont fait appel aux modèles du web sémantique afin d'annoter sémantiquement des collections du patrimoine culturel et faciliter la recherche sémantique au sein de celles-ci. Le travail présenté dans [7] combine des techniques du web sémantique et du traitement automatique du langage naturel afin d'extraire automatiquement des informations à partir de textes de zoologie antique. Un modèle de publication collaboratif pour les données culturelles a été présenté dans [2]. Ce travail présente aussi des principes de conception pour la création de portails sémantiques destiné à la recherche et aux applications en Humanités Numériques. Une plate-forme orientée ontologie a été présentée dans [1] dont le but est d'aider les utilisateurs à identifier et à caractériser de nouvelles entités pour annoter les archives historiques en utilisant des techniques d'extraction automatique d'informations et les informations récupérées dans des ensembles de données externes dans le *Linked Open Data*.

2.2 Vocabulaires et ontologies existantes

Pour représenter à la fois le corpus littéraire et les annotations sémantiques extraites de ce corpus, nous avons ré-utilisé un ensemble de vocabulaires et d'ontologies. Nous avons tout d'abord utilisé le vocabulaire schema.org [5] afin de représenter la structure des textes (c.-à-d. chapitres, paragraphes, auteur, éditeur...). Ce vocabulaire propose un ensemble de classes et propriétés génériques visant à décrire initialement des ressources du web. Nous avons choisi d'utiliser ce vocabulaire car il nous permettra à terme d'intégrer facilement d'autres types de ressources tels que des images et des vidéos. Nous avons aussi utilisé le vocabulaire Web Annotation Vocabulary (OA) [6] qui est une recommandation W3C pour représenter les zones textuelles des annotations manuelles. Ce vocabulaire permet de représenter de manière uniforme des annotations sur le Web dans un format interopérable [3]. Finalement, nous avons utilisé le vocabulaire de domaine TheZoo [4] afin de lier les entités extraites des annotations sémantiques aux concepts du thésaurus. Ce vocabulaire est conçu pour représenter et structurer hiérarchiquement tous les termes d'intérêt pour l'étude de l'histoire de la zoologie antique et médiévale à partir de trois types de corpus : (i) Textuel, (ii) Iconographique et (iii) Archéologique. TheZoo contient 6019 concepts structurés en 11 niveaux hiérarchiques. Ces concepts concernent différents aspects de la description d'animaux comme, par exemple, le concept d'anatomie interne (*internal anatomy*), les noms d'animaux (*tiger*) et de lieux géographiques (*Geographic space*). Une hiérarchie permet de classifier avec précision les concepts comme par exemple le concept de *tigre* dans la hiérarchie de la famille des organismes vertébrés : "*eumetazoa* > *bilateria* > *deuterostomia* > *vertebrata* > *tetrapoda* > *mammalia* > *carnivora* > *feliformidae* > *felidae* > *pantherinae* > *tigre*". Les concepts sont également regroupés en 14 collections qui font office de méta-concept qui leur offre un sens supplémentaire, comme la collection des *Anthroponymes* rassemblant les différents noms de personnes et d'animaux nommés par des humains ou la collection des *Archéotaxons* qui rassemble les taxons d'animaux antiques.

3. http ://sourcencyme.irht.cnrs.fr
4. http ://www.unicaen.fr/recherche/mrsh/document_numerique/projets/ichtya
5. https://schema.org/docs/about.html

3 Questions de compétences

Afin de déterminer la spécificité des connaissances à représenter, nous avons collecté et explicité sept questions de compétences (QC) formulées par les experts dans le but de comprendre précisément leurs attentes et les besoins des chercheurs du domaine afin d'apporter à ces derniers une réponse adéquate en terme d'exploration des liens entre les concepts du domaine et leur contexte de co-occurence dans les textes étudiés. Nous présentons ici des exemples de ces QC.

QC1. Quels sont les animaux qui construisent un habitat ? Le besoin des chercheurs est d'identifier dans la littérature les animaux capables de construire un habitat favorable et adapté à leurs besoins.

QC2. Quelles anecdotes mettent en relation un homme et un animal ? Le besoin des chercheurs est d'identifier les passages textuels qui permettent de repérer des interactions entre l'humain et l'animal, en particulier des formes de complicité ou de coopération et des formes d'hostilité ou de prédation.

QC3. Quels sont les remèdes (thérapeutiques) dont un ingrédient est une partie d'animal, e.g. la langue (ou un morceau de langue) ? Cette question permet aux chercheurs d'identifier l'ensemble des animaux qui ont été utilisés pour des raisons médicales et plus précisément une partie exploitée du corps de l'animal.

QC4. Quels sont les animaux qui communiquent entre eux ? Le besoin des chercheurs est d'identifier le texte où il est question d'un type de communication inter-individuelle dans une espèce animale.

QC5. Quels sont les animaux capables de jeûner et quelles sont les informations sur la fréquence ou le rythme des repas ? Cette question permet de discriminer des pratiques alimentaires et de mesurer la pertinence des savoirs antiques sur ce point.

QC6. Quelles sont les données transmises sur le temps de gestation des animaux ? Le besoin des chercheurs est d'identifier les passages textuelles qui permettent de récupérer des informations sur le temps de gestation des animaux.

4 Modèle proposé

4.1 Représentation de la structure et du contenu de l'Histoire Naturelle de Pline

Pour représenter et décrire les textes annotés, nous avons utilisé le vocabulaire *Schema* pour capturer la sémantique de la décomposition de l'oeuvre de Pline en chapitres et paragraphes. Ainsi, l'Histoire Naturelle de Pline est représentée par une instance de la classe `schema:Book` dont l'auteur est décrit par la propriété `schema:author`, le titre via `schema:headline` et l'édition via `schema:editor`. Un chapitre est une instance de la classe `schema:Chapter`, il est relié à une oeuvre via la propriété `schema:isPartOf` et le numéro du chapitre est décrit via la propriété `rdf:value`.

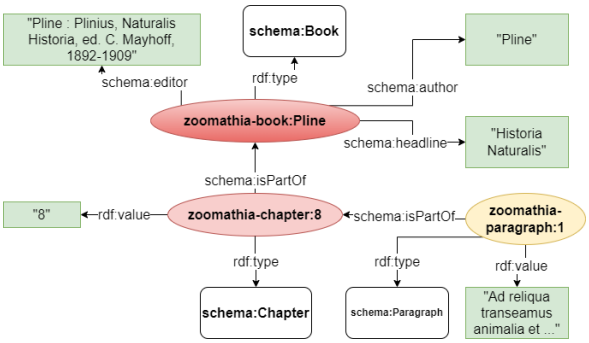

FIGURE 1 – Exemple de graphe RDF représentant un paragraphe du chapitre 8 de l'Histoire Naturelle de Pline.

Enfin, le lien entre les paragraphes et leur chapitre est représenté par la propriété `schema:isPartOf` et le contenu textuel du paragraphe est capturé comme valeur de la propriété `rdf:value`. La figure 1 présente un exemple de graphe RDF représentant le premier paragraphe du chapitre 8 de l'Histoire Naturelle de Pline.

4.2 Représentation des annotations de l'Histoire Naturelle de Pline

Afin de représenter les annotations manuelles du texte de l'Histoire Naturelle de Pline, nous avons réutilisé le vocabulaire *OA*. Une annotation a_i est une indication qu'une mention m_e d'un concept c a été identifiée dans le ou les paragraphes de l'un des quatre chapitres de l'Histoire Naturelle de Pline. Une annotation a_i est représentée comme une instance de la classe `oa:Annotation` et est décrite comme suit :

- a_i est reliée avec la propriété `oa:hasBody` à un concept c dans un vocabulaire de domaine, ici le thesaurus TheZoo.
- a_i est reliée avec la propriété `oa:hasTarget` à sa cible qui elle-même est reliée avec la propriété *oa :hasSelector* la zone de texte sélectionnée pour l'annotation et avec la propriété `oa:hasSource` au paragraphe contenant cette zone de texte. Cette zone de texte est décrite par sa valeur littérale (propriété `oa:exact`) et son début et sa fin relativement au début du paragraphe source (propriétés `oa:start` et `oa:end`).

La figure 2 présente un exemple d'annotation du paragraphe 14 du chapitre 11 de l'Histoire Naturelle de Pline. Cette annotation porte sur le texte "*tigrium rapinas*" qui est accessible via la propriété `oa:exact`. Cette annotation a été mise en correspondance avec le concept idc :5066 du thesaurus TheZoo dont un label est "Tigre" et qui est un sous concept du concept "Pantherinae".

En utilisant cette représentation RDF, nous avons pu modéliser d'une part les paragraphes des chapitres qui ont été annotés manuellement par les experts et d'autre part, la mise en correspondance de ces annotations avec les concepts des ontologies et vocabulaires du domaine (ici, le thésaurus TheZoo). Cette représentation offre la possibilité aux chercheurs d'explorer non seulement les occurrences et les

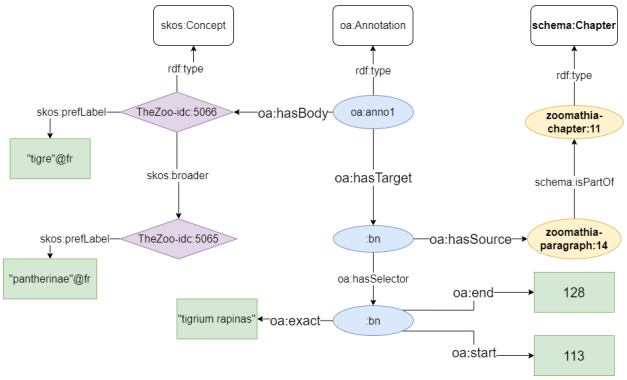

FIGURE 2 – Exemple de graphe RDF représentant l'annotation de "*tigrium rapinas*" présente dans le paragraphe 14 du chapitre 11 de l'Histoire Naturelle de Pline.

co-occurrences des annotations dans les textes mais aussi d'obtenir plus d'informations et de raisonnements sur ces annotations grâce au liage du graphe avec les ontologies et vocabulaires du domaine.

5 Construction du graphe de connaissance

5.1 Description du corpus de textes annotés

Nous avons construit un graphe de connaissances à partir des annotations manuelles du texte latin des chapitres 8 à 11 de l'Histoire Naturelle de Pline qui traitent de zoologie, respectivement des animaux terrestres, des animaux marins, des oiseaux et des insectes. Ces livres totalisent 911 paragraphes. Ces paragraphes ont été manuellement annotés par des linguistes avec les concepts du thésaurus TheZoo.
Ces annotations manuelles ont une granularité variable (un mot, un groupe de mots, un ou plusieurs paragraphes) afin de délimiter le contexte du concept annotant le texte. Le système de commentaire de Word permet de définir ces zones d'annotation et le texte de ces commentaires fait référence au(x) concept(s) du thésaurus en fonction des motifs suivants :

- "*concept*" : référence directe à un concept
- "*concept1 : concept2 : ...*" : référence à une hiérarchie de concepts où concept1 est parent de concept2
- "*concept1 ; concept2 ; ...*" : référence à des concepts distincts annotant la même portion de texte
- "*collection : concept*" : référence à un concept faisant partie d'une collection
- "*concept1 : concept2, concept3, ...*" : référence à des concepts des descendants directs d'un autre
- combinaisons des motifs précédents.

Ainsi, notre corpus de 4 livres contient 7,283 commentaires à partir desquels 13,241 références de concepts du thésaurus TheZoo ont été annotés.

5.2 Processus de lifting

La figure 3 présente le processus de transformation des annotations manuelles du texte de l'Histoire Naturelle de

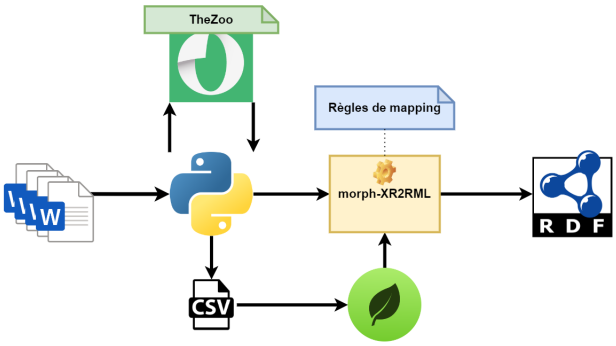

FIGURE 3 – Schéma général du processus de construction du graphe de connaissance

Pline et la construction du graphe RDF. La première étape de transformation consiste à extraire les annotations manuelles à partir des commentaires des fichiers Word. Ces informations sont stockées dans des fichiers xml internes au document word. Les informations concernant (i) le chapitre, (ii) le paragraphe, (iii) la portion du texte en latin qui a été sélectionné et qui correspond à la mention et enfin (iv) le texte du commentaire de l'expert qui correspond aux labels des concepts de TheZoo sont renseignées.
La deuxième étape consiste à utiliser l'annotation manuelle des concepts au sein de requêtes SPARQL pour rechercher dans le thésaurus l'URI du concept extrait grâce à son label. Un exemple de requête que nous avons utilisée est présenté dans le listing 1. Cette requête permet de faire la correspondance entre le label extrait et le concept du thésaurus en conservant la hiérarchie de concept ou la collection. A cause de la présence de l'homonymie dans le thésaurus, et afin d'éliminer tout cas d'ambiguïté dans l'étape de la recherche du concept, nous avons choisi de vérifier l'appartenance du concept à une collection ou à une branche de la hiérarchie.

```
SELECT ?x WHERE {
{?x a ?type; skos:prefLabel ?label.
?y skos:prefLabel ?collection; skos:member ?x.
FILTER(lang(?label) = "en").
FILTER("{label}" in (ucase(str(?label)),
  lcase(str(?label)), str(?label))).
FILTER( "{parent}" in (ucase(str(?collection)),
  lcase(str(?collection)), str(?collection))).}
UNION { ?x a ?type; skos:prefLabel ?label;
       skos:broader+ ?y.
       ?y skos:prefLabel ?concept;
FILTER(lang(?label) = "en").
FILTER("{label}" in (ucase(str(?label)),
 lcase(str(?label)), str(?label))).
FILTER( "{parent}" in (ucase(str(?concept)),
  lcase(str(?concept)), str(?concept))).}}
```

Listing 1 – Requête SPARQL de recherche de concept dans le thésaurus

Toutes ces informations (les paragraphes de commentaire, les concepts extraits, etc.) sont finalement enregistrées dans un fichier CSV qui va être injecté dans le système de gestion de base de données orienté documents MongoDB.
Puis, nous avons utilisé l'outil morph-xR2rml afin de transformer les annotations produites en un graphe RDF. Pour

Nbre de paragraphes	911
Nbre de commentaires	7283
Nbre d'entités reconnues	13241
Nbre d'entités liées	11590
Nbre d'entités non liées	2632
Nbre de triplet RDF générés	88184

TABLE 1 – Caractéristiques du graphe de connaissances produit

cela, nous avons écrit un ensemble de règles de mapping génériques permettant de générer nos triplets RDF. Les règles de mapping sont écrites en RDF à l'aide du vocabulaire xr2rml [6] basé sur r2rml [7] qui expriment des patrons de transformation de données provenant d'une base de données. Une règle de transformation définit une ressource de type *rr :TripleMap* qui est décrite par un unique sujet *rr :subjectMap*, une source logique *rr :logicalSource* qui représente la base de données MongoDB [8] et un ensemble de propriétés *rr :predicateObjectMap*. Le listing 2 présente un exemple de règle de mapping qui permet de générer une partie d'une annotation avec "rr" et "xrr" les préfixes des ontologies "r2rml" et "extended r2rml".

```
<#Anno>
a rr:TripleMap;
xrr:logicalSource [
    xrr:query """db.Annotation.find()""";];
rr:subjectMap [
        rr:template
"http://www.zoomathia.com/annotation/
sha1({$.id}_{$.chapter}_{$.paragraph})";
        rr:class oa:Annotation;];
rr:predicateObjectMap [
        rr:predicate oa:hasBody;
        rr:objectMap [
        rr:template
"https://opentheso.huma-num.fr/
?idc={$.concept}&idt=th310";];];
rr:predicateObjectMap [
    rr:predicate oa:hasTarget;
        rr:objectMap [
                rr:template "TargetBN{$._id}";
                rr:termType rr:BlankNode;];].
```

Listing 2 – Extrait d'une règle de mapping xR2RML

Nous avons défini deux bases de règles de mapping : (i) l'une pour décrire la structuration des textes de Pline en livres, chapitres et paragraphes, (ii) l'autre pour décrire les annotations de ces textes en les liant avec les paragraphes annotés, le texte de l'annotation et le lien vers les concepts de TheZoo.

A la fin de ce processus, nous avons pu extraire automatiquement 11590 concepts à partir des annotations manuelles des experts, et nous avons généré 88184 triplets RDF. Le tableau 1 résume les caractéristiques du graphe de connaissance produit.

En utilisant cette approche, nous avons échoué à lier 2632 entités annotées manuellement par les experts à des concepts de TheZoo. Cela s'explique par des irrégularités dans certaines annotations manuelles. En effet, comme la tâche d'annotation manuelle est une tâche fastidieuse pour l'annotateur, nous avons été confrontés à des problèmes d'irrégularités des règles d'annotation manuelle (faute de frappe, utilisation du pluriels, ...). Dans notre processus de transformation, nous utilisons le texte de ces annotations manuelles dans le filtre des requêtes SPARQL pour rechercher des correspondances avec les labels de concepts de TheZoo. Des annotations non uniformes engendrent des problèmes de correspondance. Par exemple, pour annoter les informations concernant la taille des animaux, l'annotateur utilise en général la syntaxe "`size : relative size`" qui fait référence au concept "relative size" dans le thésaurus. Dans certains cas, l'annotateur peut se contenter de mentionner le terme "relative" en utilisant cette syntaxe "`size : relative`".

6 Evaluation de la qualité du graphe produit

6.1 Evaluation du processus d'extraction de connaissances

Nous pouvons évaluer la qualité du graphe produit en terme de la qualité du processus d'extraction des connaissances à partir des annotations de texte en utilisant les métriques classiques de précision et rappel. Notre approche est conçue de telle manière que la précision est maximale (P=1). Notre processus de lifting des annotations en RDF a généré 11590 liens vers les concepts de TheZoo et 2632 annotations n'ont pu être liées (erreurs d'orthographe, typographiques, etc. dans les annotations, concepts absents ou labels manquants dans le thesaurus). Ainsi, la performance de notre processus en terme de rappel est de 0.814. L'analyse des annotations qui n'ont pu être liées au thesaurus avec les experts du domaine va nous permettre d'améliorer la qualité des annotations et du thesaurus.

6.2 Implémentation et réponse aux questions de compétence recueillies

Nous avons utilisé les questions de compétences présentées en section 3 pour valider le graphe RDF produit. A travers les questions de compétences traduites en SPARQL, nous avons vérifié que le graphe produit permet de répondre aux besoins des experts en terme d'exploration des connaissances zoologiques. Toutes les questions de compétences élicitées ont été formalisées en SPARQL [9] et validées par les experts du domaine. Nous ne présentons ici que deux de ces requêtes avec leurs formalisations et les différents résultats avec le retour de l'expert.

QC1. Quels sont les animaux qui construisent un habitat ? L'intention du chercheur derrière cette question est d'identifier les paragraphes des chapitres où l'auteur mentionne des animaux capables de construire leur habitat pour les étudier ensemble. Le listing 3 présente la requête SPARQL qui implémente QC1.

6. https ://www.i3s.unice.fr/~fmichel/xr2rml_specification_v5.html
7. https ://www.w3.org/TR/r2rml/
8. https ://www.mongodb.com

9. https://github.com/Wimmics/zoomathia/tree/main/Pline

```
SELECT DISTINCT ?paragraph ?name_animal
?name_construction
WHERE {
?annotation1 a oa:Annotation;
               oa:hasBody ?animal;
               oa:hasTarget [oa:hasSource
               ?paragraph; oa:hasSelector
               [oa:exact?mention_animal]].
?annotation2 oa:hasBody ?construction;
             oa:hasTarget [oa:hasSource
             ?paragraph; oa:hasSelector
             [oa:exact ?mention_construction]].
?animal  a skos:Concept;
         skos:prefLabel ?name_animal.
<https://opentheso.huma-num.fr/idg=MT_10
&idt=th310> skos:member ?animal.
?construction skos:prefLabel ?name_construction;
              skos:broader+ <https://opentheso.
              huma-num.fr/?idc=105466&idt=th310>.
FILTER (lang(?name_animal) = "en").
FILTER (lang(?name_construction) = "en")
}ORDER BY ?paragraph
```

Listing 3 – Requête SPARQL de la CQ1

Le résultat de cette requête indique, par exemple, que le paragraphe *104* du livre *10* mentionne que les Méropidae ("*bee eater*") construisent des nids ("*nest*").
QC6. Quelles sont les données transmises sur le temps de gestation des animaux ? L'intention du chercheur derrière cette question est d'identifier les paragraphes des chapitres qui mentionnent des informations du temps de gestation des animaux. Le listing 4 présente la requête SPARQL qui implémente QC6.

```
SELECT DISTINCT ?paragraph ?name_animal
?mention_animal ?mention_pregnancy
WHERE {?annotation1 a oa:Annotation;
               oa:hasBody ?animal;
               oa:hasTarget [
               oa:hasSource ?paragraph;
                oa:hasSelector [oa:exact
            ?mention_animal]].
?animal a skos:Concept;
        skos:prefLabel ?name_animal.
<https://opentheso.huma-num.fr/?idg=MT_10&idt
=th310> skos:member ?animal.
?annotation2 oa:hasBody <https://opentheso.huma-
num.fr/?idc=105364&idt=th310>;
             oa:hasTarget [
             oa:hasSource ?paragraph;
             oa:hasSelector [oa:exact
        ?mention_pregnancy]].
FILTER (lang(?name_animal) = "en").
}ORDER BY ?paragraph
```

Listing 4 – Requête SPARQL de la CQ6

7 Conclusion et travaux futurs

La capitalisation des connaissances et le développement de meilleures techniques de recherche d'informations est devenue une tâche cruciale dans la communauté des humanités numériques pour les chercheurs soucieux de valoriser le patrimoine culturel. Dans cet article, nous avons présenté un graphe de connaissance que nous avons construit à partir des annotations manuelles par des experts de l'oeuvre de Pline en utilisant le thésaurus TheZoo. Dans le graphe RDF produit, nous avons pu : (i) capturer le contexte d'apparition des différentes annotations, (ii) les décrire d'une manière structurée grâce aux vocabulaires standards du web sémantique, et (iii) lier ces annotations manuelles avec le vocabulaire de domaine TheZoo. Le graphe produit permet une interrogation uniforme, avancée à l'aide de requêtes SPARQL et qui exploite les contextes d'apparition et les liens entre les concepts du vocabulaire du domaine. La génération de ce graphe de connaissance a également permis d'identifier des problèmes d'irrégularité d'annotation. Nous avons partiellement contourné ce problème avec une recherche de correspondance approximative entre les entités extraites et les concepts du thésaurus TheZoo, par exemple en recherchant des inclusions plutôt que des égalités de chaînes de caractères entre l'entité extraite et les labels des concepts du thésaurus. Cependant, cette approche a des limites, car elle génère du bruit : par exemple, "relative" est contenue dans "relative size" mais aussi dans "tail relative size".Les entités non liées à TheZoo ont ainsi fait apparaître le besoin de corriger certaines annotations et/ou réviser ou enrichir le thésaurus TheZoo. Ce travail est prévu prochainement dans le cadre du projet Zoomathia. Une perspective de ce travail, est d'automatiser la tâche d'annotation des textes, fastidieuse pour les experts et source d'erreur. Le graphe RDF produit constitue des données de très bonne qualité pour l'entraînement d'algorithmes d'apprentissage sur lesquels reposeront l'approche que nous souhaitons développer. Une autre perspective est de faciliter l'exploitation de ce graphe RDF par les experts du domaine, philologues et historiens, qui ne sont pas spécialistes des modèles du web sémantique, en développant des interfaces de visualisation plus intelligibles et intuitives.

Références

[1] Davide Colla, Annamaria Goy, Marco Leontino, Diego Magro, and Claudia Picardi. Bringing semantics into historical archives with computer-aided rich metadata generation. *Journal on Computing and Cultural Heritage (JOCCH)*, 15(3) :1–24, 2022.

[2] Eero Hyvönen. Digital humanities on the semantic web : Sampo model and portal series. *Semantic Web*, (Preprint) :1–16, 2022.

[3] Jin-Dong Kima, Karin Verspoorb, Michel Dumontierc, and K Bretonnel Cohend. Semantic representation of annotation involving texts and linked data resources. *Semantic Web journal*, 2015.

[4] Irene Pajón Leyra, Arnaud Zucker, and Catherine Faron Zucker. Thezoo : un thesaurus de zoologie ancienne et médiévale pour l'annotation de sources de données hétérogènes. *Archivum Latinitatis Medii Aevi*, 73 :321–342, 2015.

[5] Franck Michel, Loïc Djimenou, Catherine Faron Zucker, and Johan Montagnat. Translation of Relational and Non-Relational Databases into RDF with xR2RML. In *11th International Confenrence on Web Information Systems and Technologies (WEBIST'15)*, Proceedings of the WebIST'15 Conference, pages 443–454, Lisbon, Portugal, October 2015.

[6] Robert Sanderson, Paolo Ciccarese, and Benjamin Young. Web annotation ontology. https://www.w3.org/TR/annotation-vocab/, 2017.

[7] Molka Tounsi, Catherine Faron Zucker, Arnaud Zucker, Serena Villata, and Elena Cabrio. Studying the history of pre-modern zoology with linked data and vocabularies. In *The First International Workshop Semantic Web for Scientific Heritage at the 12th ESWC 2015 Conference*, 2015.