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

Graphe de connaissance outillé au service des chercheurs en histoire de la zoologie antique et médiévale

Volume 6, nº 1-2 (2025), p. 85-106.

https://doi.org/10.5802/roia.94

La Revue Ouverte d'Intelligence Artificielle est membre du Centre Mersenne pour l'édition scientifique ouverte
www.centre-mersenne.org
e-ISSN : 2967-9672

Revue Ouverte d'Intelligence Artificielle
Volume 6, nº 1-2, 2025, 85-106

Graphe de connaissance outillé au service des chercheurs en histoire de la zoologie antique et médiévale

Molka Dhouib[a], Arnaud Barbe[a], Catherine Faron[a], Arnaud Zucker[a], Marco Corneli[a]

E-mail : molka.dhouib@inria.fr, arnaud.barbe@inria.fr, catherine.faron@inria.fr, Arnaud.zucker@univ-cotedazur.fr, Marco.corneli@univ-cotedazur.fr.

Résumé. — Chaque jour, les chercheurs en sciences humaines et sociales (SHS) se trouvent confrontés à l'analyse d'une quantité importante de ressources textuelles. Les outils de recherche traditionnels axés sur des recherches lexicales et quantitatives ne répondent pas toujours aux besoins spécifiques des historiens et philologues. Dans cet article, nous présentons les méthodes, outils et services mis en place dans le cadre du projet HisINum pour répondre à cette problématique. Notre objectif est de proposer un processus générique et réutilisable pour l'analyse de textes anciens, l'indexation sémantique sous forme de graphe de connaissance des différents textes étudiés et la mise en place d'un service de recherche et de visualisation exploitant ces graphes générés.

Mots-clés. — Graphes de connaissance, ontologies, données liées et vocabulaires, annotation sémantique de textes anciens grecs et latins, extraction d'entités nommées, histoire de la zoologie.

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 livres de l'*Histoire Naturelle* de Pline a été réalisé par un chercheur en littérature latine. L'objectif était de construire un échantillon substantiel de référence en même temps que de construire un thesaurus dédié, pour envisager une annotation semi-automatique d'un corpus élargi d'une centaine de textes zoologiques. Un expert a

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

proposé pour chaque paragraphe des annotations riches sur la thématique et le contenu. Ainsi, sous la forme de commentaires dans un document Word, chaque texte latin a été annoté avec les concepts (plusieurs milliers) du thésaurus TheZoo[(2)].

Un deuxième travail a consisté à rassembler une série de textes en latin et en grec soigneusement sélectionnés par un spécialiste, puis à les convertir au format XML/-TEI en ajoutant des méta-données sur la structure. La tâche d'annotation sémantique manuelle ne s'avère pas réalisable à grande échelle sur le corpus ainsi construit : il a fallu une année entière pour annoter les quatre livres de l'*Histoire Naturelle* de Pline, qui représentent moins de 5 % du corpus. L'annotation de l'intégralité du corpus nécessiterait des centaines d'hommes-mois.

Notre objectif est de (i) transformer les annotations manuelles faites sur le texte de Pline en graphe de connaissance et (ii) annoter automatiquement l'ensemble des textes latins et grecs du corpus pour permettre l'intégration et l'interrogation des connaissances extraites afin de proposer des possibilités de recherche automatique plus riches et répondant mieux aux besoins des chercheurs qui étudient cette littérature scientifique. Nous avons identifié les questions de recherche suivantes : (i) Quels types de connaissance devons-nous représenter pour aider les chercheurs dans leur travail d'analyse et de transmission du 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 ? (iv) Quelle approche pouvons-nous utiliser pour annoter automatiquement le corpus de textes ? 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 et automatiques des textes de zoologie ancienne, (ii) l'explicitation de questions de compétence auprès d'historiens et philologues intéressés par la transmission des connaissances zoologiques. Le processus de construction du graphe de connaissance comprend cinq étapes successives : (i) la reconnaissance des entités pertinentes dans les annotations manuelles ou l'annotation automatique pour les textes non annotés, (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 [8].

Cet article est une extension de notre publication dans les actes de la conférence IC 2023 [1] dans laquelle nous nous focalisions sur l'annotation manuelle des quatre livres de l'*Histoire Naturelle* de Pline. Dans le présent article nous intégrons la présentation de notre travail d'annotation automatique de tout un corpus de textes de zoologie antique et médiévale ainsi que l'application d'exploration visuelle du corpus développée. L'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,

[(2)] https://opentheso.huma-num.fr/opentheso/?idt=th310

nous présentons un ensemble de questions de compétence 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. Dans la section 6 nous présentons des requêtes SPARQL qui implémentent les questions de compétence élicitées et dont la réponse peut être recherchée en interrogeant le graphe de connaissance produit, validant ainsi celui-ci. Enfin, dans la section 7, nous présentons une interface de visualisation des textes annotés et de recherche des paragraphes spécifiques en fonction d'un ensemble de concepts. Nous concluons en section 8.

2. État de l'art

Dans cette section, nous discutons de quelques approches existantes pour la construction de graphes de connaissance à partir de ressources culturelles. Par la suite, nous présentons les vocabulaires et les ontologies présents dans la littérature pour structurer la connaissances. Finalement, nous étudions les différentes approches pour l'extraction des entités à partir de textes.

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 sur la valorisation de textes anciens dans le domaine de l'Histoire Naturelle ont porté sur la structuration d'encyclopédies médiévales en XML selon le modèle TEI et l'annotation manuelle de ces sources de données, notamment dans le cadre des projets SourceEncyMe4[(3)] et Ichtya5[(4)]. Ces travaux représentent les premiers pas pour la structuration des textes anciens. Notre travail présenté dans [13] a été pionnier dans la combinaison de l'utilisation des modèles du web sémantique et du traitement automatique du langage naturel afin d'extraire automatiquement des informations à partir de textes de zoologie antique et les annoter sémantiquement. Un graphe de données RDF et son vocabulaire ont été construits afin de représenter les connaissances extraites à partir des règles syntaxiques. Ce travail ne réutilisait pas encore de vocabulaires existants pour la description du graphe.

Plus récemment, un modèle de publication collaboratif pour les données culturelles a été introduit dans [3]. Ce travail présente aussi des principes de conception pour la création de portails sémantiques destinés à la recherche et aux applications en Humanités Numériques. Une plate-forme orientée ontologie a été conçue dans [2] 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. Le projet ISSA [12] présente un pipeline pour l'analyse des documents d'une archive scientifique ouverte dans le domaine de l'agriculture. Ce pipeline utilise trois outils pour identifier et lier les entités nommées à partir des articles scientifiques :

[(3)] `http://sourcencyme.irht.cnrs.fr`

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

DBpedia Spotlight, Entity-fishing et l'outil pyclinrec qui permet d'obtenir une annotation par projection sur dictionnaire. En utilisant le résultat de l'extraction des entités nommées et les métadonnées des articles, un travail d'indexation sémantique sous forme de graphe de connaissance est par la suite mis en place. Des services de recherche et de visualisation à partir de ces graphes de connaissance sont développés pour aider à l'analyse des articles scientifiques. Ce dernier travail propose une approche facile à mettre en œuvre, tout en incluant une dimension de visualisation.

2.2. Vocabulaires et ontologies existantes

Plusieurs travaux ont proposé des vocabulaires pour la représentation des données littéraires. Parmi ces vocabulaires, nous pouvons commencer par citer des vocabulaires génériques comme le Dublin Core (DCMI – Dublin Core Metadata Initiative)[(5)] et Schema.org[(6)]. Ces deux vocabulaires proposent un vocabulaire de métadonnées simple permettant une description rapide et flexible des documents (titre, auteur, date, format, sujet, etc). Cependant, ils ne permettent pas de décrire la structure hiérarchique d'une œuvre. Il existe dans la littérature, d'autres vocabulaires plus spécifiques à la représentation des données littéraires : (i) BIBFRAME (Bibliographic Framework) [7] est une initiative développée par la bibliothèque du congrès des États-Unis qui adopte une approche de modélisation avec des types d'entités tels que œuvre, instance (pour les différentes éditions d'une œuvre), et item (pour les exemplaires physiques). (ii) FRBR (Functional Requirements for Bibliographic Records) [10] est un modèle conceptuel élaboré par l'IFLA (International Federation of Library Associations) pour organiser les données bibliographiques en fonction des relations entre œuvres, expressions, manifestations et items. Il vise à clarifier la manière dont les différentes versions d'une œuvre sont liées entre elles. (iii) RDA (Ressource Description and Access) [11] est dédié à la description des ressources et l'accès aux données bibliographiques. Il définit quatre niveaux d'entités bibliographiques : œuvre, expression, manifestation et item. La conception de ce vocabulaire permet de gérer plus facilement les différentes versions d'une œuvre telles que des traductions ou des éditions différentes. (iv) RDA-FR[(7)] est la variante francophone de RDA pour répondre aux besoins des bibliothèques francophones en introduisant des spécificités terminologiques et culturelles. Ces différents vocabulaires partagent un point commun : examiner les liens entre une œuvre, ses différentes éditions et exemplaires présents dans les bibliothèques. Cependant, malgré leur capacité à bien représenter ces liens, ces vocabulaires manquent de précision pour décrire les détails internes d'une œuvre, tels que sa décomposition en section, sous-sections ou paragraphes. Pour la représentation des annotations textuelles, nous trouvons dans la littérature le vocabulaire OA (Web Annotation Vocabulary [9] qui est une recommandation du W3C pour représenter les zones textuelles des annotations de textes. Ce vocabulaire permet de représenter de manière uniforme des annotations sur le Web dans un format interopérable [5].

[(5)] https://www.dublincore.org/specifications/dublin-core/dces/

[(6)] https://schema.org/docs/about.html

[(7)] https://rdafr.fr/

Pour représenter à la fois le corpus littéraire de Zoomathia et les annotations sémantiques extraites de ce corpus, nous avons réutilisé les vocabulaires standards OA et Schema.org et le vocabulaire de domaine TheZoo [6] pour lier les entités extraites des annotations sémantiques aux concepts du domaine de la zoologie antique et médiévale. TheZoo 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. Il nous permettra à terme d'intégrer facilement d'autres types de ressources tels que des images et des vidéos. 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 les 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-concepts et offrent un sens supplémentaire, comme la collection des *Anthroponymes* rassemblant les différents noms de personnes, ou la collection des *Archéotaxons* qui rassemble les taxons zoologiques utilisés dans les textes antiques.

2.3. Reconnaissance d'entités nommées dans les textes anciens

La reconnaissance d'entités nommées (EN) est une tâche de traitement automatique du langage naturel (TALN) qui consiste à identifier et classer les entités dans un texte telles que des personnes, des lieux, des organisations des dates, etc. La mise en place de cette tâche dans le contexte de textes anciens, qu'ils soient en latin ou grec ancien, présente des défis en raison de la faible disponibilité de corpus annotés, la nature morphologique complexe et l'utilisation de vocabulaires anciens. Les approches de reconnaissance d'entités nommées (NER) pour les textes anciens peuvent être regroupées en trois catégories. (i) Les approches basées sur des règles [13] : elles se basent sur des règles linguistiques, syntaxiques et des listes d'entités connues pour identifier et extraire les entités dans les textes. Bien que ces approches ont l'avantage de ne pas nécessiter un grand corpus annotés, elles sont limitées par la rigidité des règles face aux variations morphologiques, leur incapacité à être généralisées sur de nouveaux textes, et l'incapacité de découvrir de nouvelles entités. (ii) Les approches basées sur les lexiques : elles sont basées sur la tâche de mise en correspondance entre les termes du texte et les termes de listes d'entités nommées préexistantes. Parmi les outils, nous pouvons mentionner Entity-fishing et DBpedia Spotlight. Entity-fishing est un outil d'annotation d'EN conçu pour fonctionner avec des textes multilingues, en utilisant des bases de connaissances comme Wikidata[(8)] pour découvrir et désambiguïser les entités dans les textes. DBpedia Spotlight utilise les données de DBpedia [(9)] pour extraire les entités nommées des textes. La limite de ces approches réside dans le

[(8)] `https://www.wikidata.org/`
[(9)] `https://fr.dbpedia.org/`

fait que la reconnaissance des entités dépend de la couverture de ces bases de connaissance en terme de contexte et de langues. (iii) Les approches basées sur l'apprentissage automatique (chaînes de Markov, LSTM, transformers). Notamment, CLTK (Classical Language Toolkit) [4] fournit un outil pour l'analyse des langues anciennes en intégrant des modèles d'apprentissage supervisés pour identifier les entités nommées classiques dans les textes latins comme les noms de personnages historiques, les lieux, les divinités, etc. Toutefois, la performance de ces modèles est souvent limitée par la taille, la qualité et le type des entités présentes dans les corpus annotés disponibles.

Dans le cadre de notre travail, notre objectif n'est pas simplement d'identifier des entités nommées avec des types prédéfinis, mais de découvrir de nouveaux types d'entités qui n'ont pas été annotés dans un corpus, comme par exemple l'usage des animaux dans des domaines tels que la médecine ou la cuisine. C'est pourquoi nous avons choisi d'utiliser les outils Entity-fishing et DBpedia Spotlight, qui permettent d'explorer cet aspect de découverte et d'élargir la reconnaissance des entités au-delà des catégories classiques.

3. Questions de compétence

Afin de déterminer la spécificité des connaissances à représenter, nous avons collecté et explicité sept types de questions de compétence (QC) formulées par les chercheurs du domaine pour comprendre précisément leurs attentes et besoins et leur apporter une réponse adéquate en terme d'exploration des liens entre les concepts du domaine et leur contexte de co-occurence dans les textes étudiés.

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 leurs vertus médicales supposées et plus précisément les parties exploitées du corps de l'animal.
QC4. Quels sont les animaux qui communiquent entre eux ? Le besoin des chercheurs est d'identifier les textes où il est question d'un type de communication interindividuelle dans une espèce animale, voire interspécifique.
QC5. Quels sont les animaux capables de jeûner et quelles sont les informations sur la fréquence ou le rythme de leurs 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 textuels qui donnent des informations sur le temps de gestation des animaux et de confronter les informations sur les animaux sauvages ou domestiques avec les données contemporaines.

QC7. Quelles sont les expérimentations faites sur les animaux ? Cette question permet aux chercheurs de recenser les différentes expérimentations effectuées dans l'antiquité afin de les comparer aux connaissances modernes.

4. Modèle proposé

4.1. Représentation de la structure et du contenu des textes anciens

Nous avons examiné la structuration des textes annotés et non annotés afin d'identifier les diverses classes et relations nécessaires pour représenter et décrire ces textes. En général, nous avons constaté que les fichiers XML présentent une structuration en quatre niveaux : (i) œuvre, (ii) livre, (iii) chapitres et (iv) paragraphes. Certains fichiers peuvent également comporter un niveau additionnel destiné à représenter les sections des livres. Ces différents niveaux de structuration varient d'une ressource textuelle à l'autre. Nous n'avons pas trouvé de vocabulaire complet capable de représenter tous ces niveaux de structuration. Par exemple, le vocabulaire *Schema.org* contient des classes pour représenter les livres et les chapitres mais ne contient pas des classes pour représenter les paragraphes et les sections. Pour cela, nous avons créé notre propre vocabulaire et nous l'avons aligné avec le vocabulaire *Schema.org*. Ainsi, par exemple l'*Histoire Naturelle* de Pline est représentée par une instance de la classe `zoo:œuvre` qui est alignée avec la classe `schema:CreativeWork`. Les propriétés `zoo:author`, `zoo:title` et `zoo:editor` relient une oeuvre à son auteur, son titre et son éditeur. Un livre est une instance de la classe `zoo:Book`, il est relié à une œuvre par la propriété `zoo:isPartOf` (inverse de la propriété `zoo:hasPart`) et le numéro du chapitre est décrit par la propriété `zoo:identifier`. Enfin, le lien entre les paragraphes et leur livre est représenté par la propriété `zoo:isPartOf` et le contenu textuel du paragraphe est capturé comme valeur de la propriété `zoo:text`. La partie droite de la figure 4.1 présente un exemple de graphe RDF représentant le quatrième paragraphe du livre 11 de l'*Histoire Naturelle* de Pline.

4.2. Représentation des annotations des textes anciens

Afin de représenter les annotations manuelles ou automatiques des textes anciens, 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 du texte. 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 d'un vocabulaire de domaine, dans notre cas 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` à 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`).

En utilisant cette représentation RDF, nous avons pu modéliser d'une part les paragraphes des chapitres qui ont été annotés manuellement par les experts ou automatiquement 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 co-occurrences des annotations dans les textes mais aussi d'obtenir plus d'informations et d'inférences sur ces annotations grâce au liage du graphe avec les ontologies et vocabulaires du domaine.

La figure 4.1 présente un exemple d'annotation du paragraphe 4 du livre 11 de l'*Histoire Naturelle* de Pline portant sur le texte « *tigrium rapinas* » qui apparait en valeur de 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 de label « Pantherinae ».

Pour décrire les œuvres, nous avons développé un vocabulaire aligné avec Schema.org et qui étend celui-ci avec des classes répondant à notre besoin particulier de modélisation : `zoo:Section`, `zoo:Paragraphe`, `zoo:PageBekker` et `zoo:Fable`.

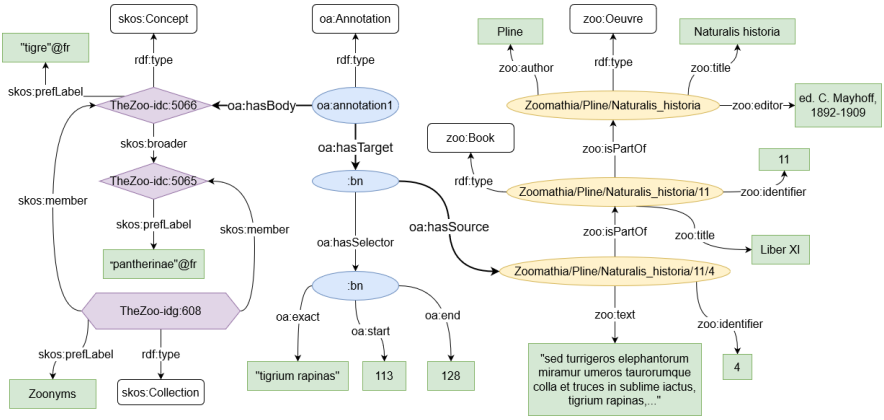

FIGURE 4.1. Graphe RDF représentant l'annotation de « *tigrium rapinas* » présente dans le paragraphe 14 du livre 11 de l'*Histoire Naturelle* de Pline.

5. CONSTRUCTION DU GRAPHE DE CONNAISSANCE

5.1. DESCRIPTION DU CORPUS DE TEXTES ANNOTÉS MANUELLEMENT

Nous avons construit un graphe de connaissance à partir des annotations manuelles du texte latin des livres 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 descendants directs d'un même autre
- combinaisons des motifs précédents.

La figure 5.1 montre un exemple d'annotation manuelle faite par les linguistes.

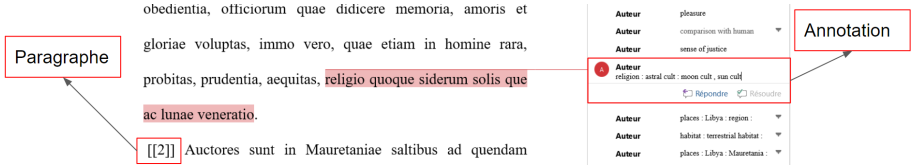

FIGURE 5.1. Texte de Pline avec annotations manuelles.

Ainsi, notre corpus de quatre livres contient 7,283 commentaires à partir desquels 13,241 références de concepts du thésaurus THEZOO ont été annotées.

5.2. DESCRIPTION DU CORPUS DE TEXTES ANNOTÉS AUTOMATIQUEMENT

Dans un second temps, en utilisant le même modèle de graphe, nous avons annoté automatiquement le corpus de 41 textes (3 textes latins et 38 textes grecs) de Zoomathia[(10)], écrits pendant la période antique et portant sur des thématiques variées comme par exemple l'agriculture, la chasse, les animaux, la géographie, etc. Le nombre de paragraphes de ces textes varie entre une quinzaine et des milliers selon le texte traité. Ces textes sont disponibles au format TEI/XML, c'est-à-dire que leur structure est explicite.

5.3. PROCESSUS DE CONSTRUCTION DU GRAPHE DE CONNAISSANCE

La figure 5.2 présente le processus général de transformation des annotations manuelles ou automatiques des textes anciens en graphe RDF.

[(10)] `https://github.com/Wimmics/zoomathia/tree/main/Named_entity_recognition/data/original_files`

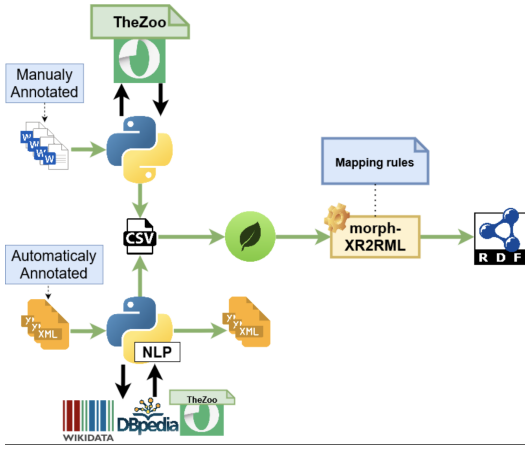

FIGURE 5.2. Schéma général du processus de construction du graphe de connaissance.

5.3.1. *Processus de transformation des annotations manuelles en graphe RDF*

La première étape de transformation consiste à extraire les annotations manuelles à partir des commentaires dans le fichier Word contenant le texte de l'*Histoire Naturelle* de Pline. 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 sélectionnée 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 renseignés.

La deuxième étape consiste à rechercher dans le thesaurus TheZoo, à l'aide d'une requête SPARQL, le concept ayant pour label le terme utilisé dans une annotation. 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 considérant la structure hiérarchique lorsqu'elle est exprimée par l'annotateur. Pour palier l'existence d'homonymie dans le thésaurus, afin d'éliminer toute ambiguïté dans la recherche du concept, nous vérifions l'appartenance du concept à une collection ou à une branche de la hiérarchie.

Toutes ces informations extraites (paragraphes de commentaire, concepts extraits, position de la mention annotée, taille de la mention annotée et référence vers le livre) sont enregistrées dans un fichier CSV qui est injecté dans un système de gestion de base de données orientée documents MONGODB.

La dernière étape consiste à utiliser l'outil MORPH-XR2RML afin de transformer les annotations produites en un graphe RDF. Pour cela, nous avons écrit un ensemble de règles de mapping génériques permettant de générer des triplets RDF à partir des données MongoDB. Ces règles sont écrites en RDF à l'aide du vocabulaire XR2RML[(11)]

[(11)] `https://hal-lirmm.ccsd.cnrs.fr/UNICE/hal-01066663v1`

basé sur R2RML[12] qui expriment des patrons de transformation 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 correspond à la collection de la base de données MongoDB contenant les données à utiliser 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.

```
SELECT ?candidate WHERE {
    { ?candidate skos:prefLabel ?label.
      ?parent skos:prefLabel ?collection;
          skos:member ?candidate.
      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
    { ?candidate skos:prefLabel ?label;
          skos:broader+ ?parent.
      ?parent 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 concepts dans le thésaurus.

Nous avons défini deux bases de règles de mapping : (i) l'une pour décrire la structuration des textes de Pline en œuvres, livres, chapitres, sections 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 les concepts de TheZoo.

```
<#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;
  ];
```

[12] https://www.w3.org/TR/r2rml/

```
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. "rr" et "xrr" sont les préfixes des ontologies "r2rml" et "xr2rml".

À 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. La Table 5.1 résume les caractéristiques du graphe de connaissance produit.

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 5.1. Caractéristiques du graphe de connaissance produit.

```
prefix skos: <http://www.w3.org/2004/02/skos/core#> .
SELECT DISTINCT ?x WHERE {
    ?x a skos:Concept; skos:prefLabel ?label.
    FILTER ( lang(?label) = "en" || lang(?label) = "fr" ).
    FILTER ( contains(str(?label), "{inputLabel}") ).
}
```

Listing 3. Requête SPARQL permettant de trouver le ou les concepts du thesaurus étant donné un label

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 (fautes de frappe, utilisation du pluriel, erreurs de langue, etc.) liées au caractère fastidieux de la tâche. Dans notre processus de transformation, nous utilisons le texte de ces annotations manuelles dans le filtre des requêtes SPARQL pour rechercher des correspondances avec des labels de concepts de TheZoo. Une des requêtes utilisées est présentée dans le Listing 3. 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 se contente de mentionner le terme « relative » en utilisant la syntaxe "`size : relative`". Malgré une phase de nettoyage des données, toutes les irrégularités d'annotation n'ont pu être corrigées étant données leur variété.

```
prefix skos: <http://www.w3.org/2004/02/skos/core#> .
SELECT DISTINCT ?x WHERE {
    ?x a skos:Concept; skos:prefLabel ?label.
    FILTER ( lang(?label) = "en" || lang(?label) = "fr" ).
    FILTER ( contains(str(?label), "{inputLabel}") ).
}
```

Listing 4. Requête SPARQL permettant de trouver le ou les concepts du thesaurus étant donné un label

5.3.2. *Processus de transformation des annotations automatiques en graphe RDF*

Une étape préalable au processus d'annotation automatique et de transformation est la traduction automatique du contenu textuel de sa langue d'origine (latin ou grec) vers l'italien et l'anglais à l'aide de l'outil Google Translation[(13)]. Nous avons opté pour ces deux langues car il n'y a pas d'outil permettant de traiter directement le latin ou le grec ancien. L'italien représente la langue la plus proche du latin et l'anglais a été retenu en raison de la bonne performance des outils d'extraction dans cette langue. Suite à cette phase de traduction, nous avons utilisé les outils DBpedia Spotlight[(14)] et entity-fishing pour identifier, désambiguïser et lier les entités nommées (EN) à partir des traductions des textes. DBpedia Spotlight ajoute des annotations d'entités DBpedia aux textes en italien et en anglais, tandis que entity-fishing annote les textes traduits avec les entités de Wikidata. Notre approche d'annotation vise à considérer le plus grand nombre possible d'entités nommées (pertinentes), facilitant ainsi la découverte de concepts non présents dans le thésaurus TheZoo. Cela ouvre la possibilité d'enrichir celui-ci par ces concepts candidats. Pour affiner les annotations de DBpedia spotlight et entity-fishing, et après une première analyse sur les données, nous avons introduit des filtres pour exclure certains résultats d'annotation automatique. En premier lieu, nous avons fixé empiriquement un seuil de confiance, en ne retenant que les entités dont le score de correspondance est supérieur à 0,6. Concrètement, les annotations dont le score de confiance est inférieur à ce seuil sont exclues, car le risque est grand d'introduire du bruit dans les résultats. Un seuil plus élevé pourrait réduire le rappel en excluant des annotations potentiellement correctes mais avec une confiance plus faible, tandis qu'un seuil plus bas augmenterait le risque d'inclure des annotations erronées.

[(13)] https ://translate.google.com
[(14)] https ://www.dbpedia-spotlight.org/

De plus, nous avons remarqué que plusieurs entités sont annotées avec des catégories DBpedia qui ne sont pas en relation avec nos thématiques d'intérêt. Nous avons alors exclu 17 catégories DBpedia telles que : "DBpedia:MusicalWork", "DBpedia:Company", "DBpedia:MusicalArtist", "DBpedia:TelevisionShow", etc. Nous avons également filtré les entités DBpedia en fonction de la chaîne de caractères qui constitue leur URI, par exemple en excluant celles contenant les sous-chaînes « film », « music », « song ». Le résultat de ces étapes de traduction automatique des textes et d'extraction et reconnaissance automatiques d'entités nommées est enregistré dans des fichiers XML incluant l'annotation dans le paragraphe annoté à l'aide d'un nouvel élément fils `note` possédant un ensemble d'attributs tels que le texte de l'annotation, le lien vers un concept de DBpedia, le score de confiance. Un exemple d'annotation est fourni dans le Listing 5.

```
<p lang="en">
a heartburn occurs, the veins are cut open in the summer,...and a
    very strong fever, ...
    <note DBPEDIA_ENT="fever" lang="en"
          category="Wikidata:Q12136,DBpedia:Disease"
          source="DBpedia"
          link="http://dbpedia.org/resource/Fever"
          type="automatic"
          score="0.9999999999823785" start="138" end="142">
    </note>
</p>
```

Listing 5. Extrait d'une annotation automatique de paragraphe.

Nous avons ensuite enrichi ces annotations automatiques en recherchant les concepts du thesaurus TheZoo qui correspondaient aux concepts de DBpedia et Wikidata extraits. La requête SPARQL présentée dans le Listing 6 repose sur une correspondance exacte entre le label du thésaurus TheZoo et les annotations issues de DBpedia et Wikidata. Ces annotations sont intégrées à la requête sous forme de paramètre, et en tenant compte de différentes langues à savoir le latin, le grec, l'anglais ou l'italien.

```
SELECT ?x WHERE {
    ?x a skos:Concept; skos:prefLabel ?label.
    FILTER ( LANG(?label) IN ("en", "it", "la", "el") )
    FILTER ( STR(?label) = "{inputLabel}" )
}
```

Listing 6. Requête SPARQL pour l'enrichissement automatique du graphe d'annotation avec des concepts du thesaurus TheZoo

Finalement, nous avons enregistré toutes ces informations dans un fichier CSV de même format que celui résultat des annotations manuelles de l'*Histoire Naturelle* de Pline que nous avons transformé en RDF à l'aide du logiciel Morph-xR2RML en utilisant les mêmes règles de transformation que précédemment.

À la fin de ce processus, nous avons pu extraire automatiquement 29981 concepts à partir des annotations automatiques, et nous avons généré 1082452 triplets RDF. La Table 5.2 résume les caractéristiques du graphe de connaissance produit. Nous avons également généré un thésaurus contenant les entités extraites automatiquement de DBpedia et Wikidata et qui n'ont pas été liées avec les concepts de TheZoo. L'objectif est de le soumettre aux experts pour qu'ils l'analysent et valident les entités que nous n'avons pas pu lier, en raison soit de l'absence de ces concepts dans TheZoo, soit d'erreurs d'orthographe. De cette manière nous proposons une correction et un enrichissement semi-automatisé du thésaurus.

Nbre de fichiers XML	31
Nbre de paragraphes	975629
Nbre d'entités reconnues	97945
Nbre d'entités liées	29981
Nbre d'entités non liées	67964
Nbre de triplets RDF générés	1082452

TABLE 5.2. Caractéristiques du graphe de connaissance produit automatiquement.

6. Évaluation de la qualité du graphe produit

Afin de valider notre approche de construction de graphe RDF, nous avons choisi deux méthodes : (i) une validation en utilisant les métriques de précision et rappel du domaine de la recherche d'information, (ii) une validation à travers les questions de compétence.

6.1. Évaluation du processus d'extraction de connaissances

Nous pouvons évaluer la qualité des graphes produits en termes de qualité du processus d'extraction des connaissances à partir des annotations textuelles, en utilisant les métriques classiques de précision et rappel. Notre approche de construction de graphe RDF à partir des annotations manuelles de texte est conçue de telle manière que la précision est maximale (P=1). Nous avons généré 11 590 liens vers les concepts de TheZoo et 2 632 annotations n'ont pu être liées (erreurs d'orthographe, typographiques, etc. dans les annotations, concepts absents ou labels manquants dans le thésaurus). Ainsi, la performance de notre processus en termes de rappel pour les annotations manuelles sur le corpus de l'*Histoire Naturelle* de Pline est de 0,814, et donc nous avons une valeure de F-mesure de 0,913. L'évaluation de notre approche d'extraction automatique d'entités nommées à partir de textes par des experts est en cours.

6.2. Implémentation et réponse aux questions de compétence recueillies

Nous avons utilisé les questions de compétence présentées en section 3 pour valider le graphe RDF produit. À travers leur traduction 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 compétence élicitées ont pu être formalisées en SPARQL[15] 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 à travers cette question est d'identifier et rassembler les passages où un auteur mentionne des animaux capables de construire leur habitat pour les étudier ensemble. Le Listing 7 présente la requête SPARQL qui implémente QC1.

```
SELECT DISTINCT ?paragraph ?name_animal ?name_construction WHERE {
    ?annotation1 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>.
} ORDER BY ?paragraph
```

Listing 7. Requête SPARQL implémentant la question de compétence QC1.

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* »).

TABLE 6.1. Tableau résultat de la QC1.

Paragraph	Animal	Habitat
Pliny/historia_naturalis/10/10	eagle	nest
Pliny/historia_naturalis/10/10	osprey	nest
Pliny/historia_naturalis/8/218	ferret	burrow
Pliny/historia_naturalis/8/218	rabbit	burrow
Pliny/historia_naturalis/11/16	bee	hive
Pliny/historia_naturalis/11/22	worker bee	hive

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

[15] https://github.com/Wimmics/zoomathia/tree/main/Pline

chapitres qui mentionnent des informations sur le temps de gestation des animaux. Le Listing 8 présente la requête SPARQL qui implémente QC6.

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

Listing 8. Requête SPARQL implémentant la question de compétence QC6.

Le résultat de cette requête contient les passages où sont mentionnés des informations concernant le temps de gestation d'un animal ainsi que son type de gestation. La table 6.2 présente un extrait de ces résultats. Par exemple, l'annotation de l'*Histoire Naturelle* de Pline indique que la durée de la grossesse du « blackbird » figure dans le paragraphe 147 du livre 10.

TABLE 6.2. Tableau résultat de la QC6.

Paragraph	Animal	type de grossesse
Pliny/historia_naturalis/11/50	bee	length of incubation
Pliny/historia_naturalis/11/85	phalangium	length of incubation
Pliny/historia_naturalis/10/147	blackbird	length of pregnancy
Pliny/historia_naturalis/10/147	swallow1	length of pregnancy

L'ensemble des requêtes implémentant des questions de compétence ainsi que leurs résultats sont disponibles sur le Github du projet[16].

7. Représentation visuelle et exploration du graphe de connaissance

Dans leur travail d'analyse textuelle, les chercheurs en SHS sont amenés à construire un corpus ciblé et dégager des paragraphes intéressants pour une ou des thématique(s) recherchée(s). Ces thématiques sont identifiés par des ensembles de concepts zoologiques spécifiques ou généralistes. Les historiens souhaitent, par ailleurs, pouvoir

[16] https://github.com/Wimmics/zoomathia/blob/main/Pline/SPARQL_queries_jupyter.ipynb

réaliser une analyse croisée des textes afin d'identifier et comparer la transmission du savoir zoologique des textes qui traitent de ces thématiques ou de concepts proches. Nous présentons ici une application Web permettant la définition, la recherche et l'exploration de corpus de textes de zoologie antique ainsi que la visualisation de questions de recherche génériques, développée à destination des chercheurs en SHS pour supporter leur analyse de corpus. Elle est accessible à cette adresse : `http://zoomathia.i3s.unice.fr`.

Cette application sur mesure a été développée en JavaScript, en adoptant une architecture client-serveur dont les détails de l'implémentation sont décrits sur le github du projet[17]. Le frontend repose sur le framework ReactJS, tandis que le backend utilise le framework Express.js facilitant la définition d'API REST. Toutes les données affichées dans l'interface, telles que les listes d'auteurs, d'œuvres, les textes et les annotations, proviennent du graphe de connaissance et sont obtenues via des requêtes SPARQL exécutées depuis le serveur. Autrement dit, le graphe de connaissance que nous avons construit est la base de données qui alimente l'interface. Un exemple de code présenté dans le listing 9 illustre une requête SPARQL paramétrée simple permettant de retrouver les relations de partonomie dans une œuvre donnée en entrée pour générer sa table des matières.

```
prefix zoo: <http://www.zoomathia.com/2024/zoo#>
SELECT DISTINCT ?parent ?current ?type (xsd:integer(?id_t) AS ?id)
    ?title ?file
WHERE {
    ?current a ?type;
        zoo:isPartOf+ <${uri}>;
        zoo:isPartOf ?parent_t;
        zoo:identifier ?id_t.
    BIND(IF(?parent_t = <${uri}>, ?current, ?parent_t) AS ?parent)
    OPTIONAL { ?current zoo:title ?title_t. }
    BIND(IF(BOUND(?title_t), ?title_t, "") AS ?title)
} ORDER BY ?id ?parent
```

Listing 9. Génération de requêtes SPARQL permettant de récupérer la hierarchie d'une oeuvre donnée

L'application offre deux interfaces d'exploration des textes du corpus : (1) au sein d'une œuvre préalablement sélectionnée par l'utilisateur et (2) sur l'ensemble du corpus annoté ou en définissant un sous-corpus personnalisé. La figure 7.1 présente l'interface pour la visualisation des annotations du livre 8 de l'*Histoire Naturelle* de Pline et la figure 7.2 présente l'interface de visualisation d'un corpus. Chacune de ces interfaces de visualisation est découpée en trois parties. Une zone de métadonnées dans le bordereau grisé en entête, une zone à gauche contenant la table des matières de l'œuvre ou des œuvres affichées, et enfin une zone centrale divisée en trois colonnes contenant,

[17] `https://github.com/Wimmics/zoomathia/tree/main/web-app`

respectivement, de gauche à droite, le numéro de paragraphe, le texte du paragraphe et les concepts annotés du paragraphe. Le composant affichant le paragraphe permet de visualiser des passages annotés en les surlignant et au survol d'un concept et au clic sur un concept de rediriger vers la fiche du concept du thésaurus TheZoo.

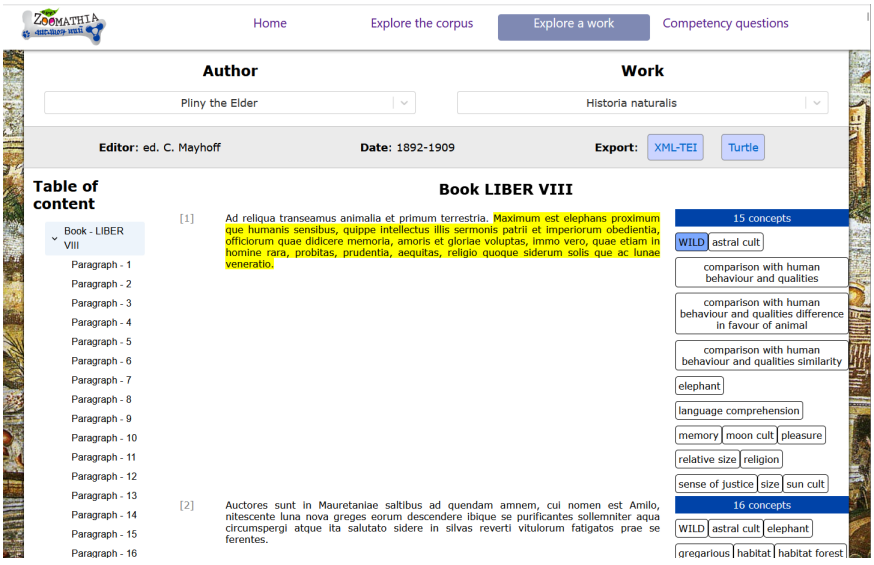

FIGURE 7.1. Visualisation des annotations du livre 8 de l'*Histoire Naturelle* de Pline.

Sur l'interface de visualisation d'un corpus (figure 7.2), une fonction de recherche multilingue permet à l'utilisateur de définir son sous-corpus d'intérêt en indiquant les œuvres, l'auteur et/ou les concepts de ce sous-corpus. Il peut sélectionner un ou

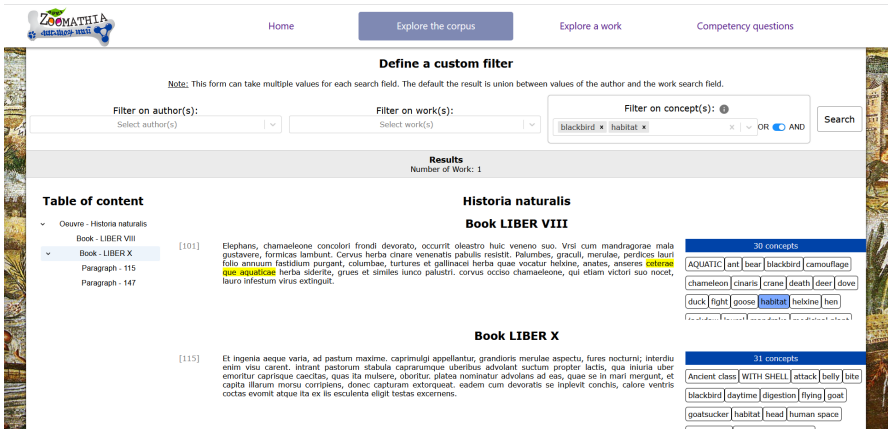

FIGURE 7.2. Visualisation des annotations d'un corpus selon le filtre utilisateur.

plusieurs concepts et indiquer si le corpus doit contenir les paragraphes où au moins un de ces concepts est présent, ou bien les paragraphes où tous les concepts sont présents. Les concepts renseignés peuvent être des concepts SKOS (instances de `skos:Concept`) ou des thématiques (instances de `skos:Collection`). Ces champs de recherche alimentent une requête SPARQL paramétrée. Cette interface offre ainsi une abstraction de requête SPARQL de recherche permettant aux non spécialistes d'exploiter le graphe de connaissance. La figure 7.2 illustre l'utilisation de la construction de sous-corpus avec un label de concept en anglais (blackbird) et un autre en français (habitat).

Un troisième onglet permet d'accéder à une interface listant les questions de compétence exprimées par les experts interviewés et présentées dans la section 3. Ces questions ont été formalisées en SPARQL et l'utilisateur peut en sélectionner une pour l'exécuter et visualiser le résultat. Deux types de visualisation des résultats sont possibles : (1) une table contenant le résultat de la requête SPARQL et (2) une présentation graphique permettant une exploration interactive du graphe à l'aide de l'outil MGExplorer. La figure 7.3 présente un exemple de rendu pour l'exploration de la première question de compétence. Ces visualisations visent à mettre en évidence des liens entre concepts ou œuvres difficiles à déceler dans une analyse linéaire d'un texte.

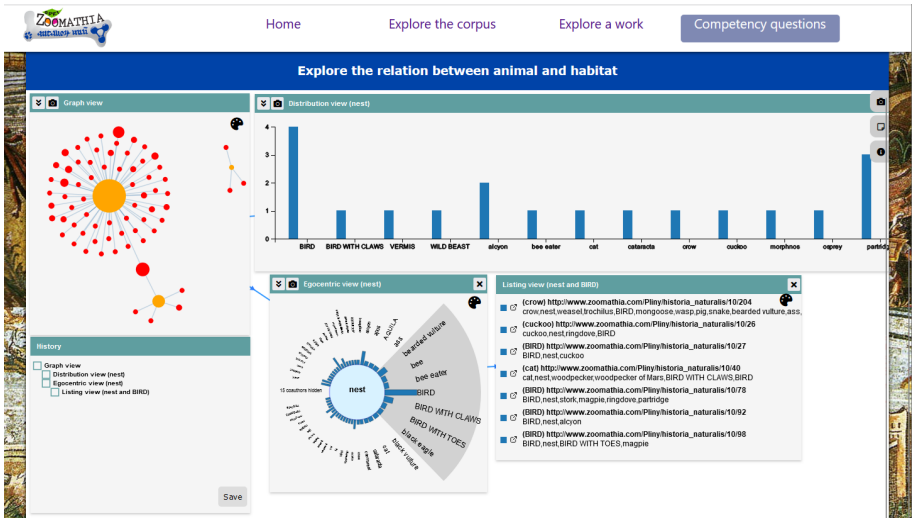

FIGURE 7.3. Visualisation des annotations du paragraphe selon le filtre utilisateur.

8. CONCLUSION ET TRAVAUX FUTURS

La capitalisation des connaissances et le développement de techniques de recherche d'informations est devenue une tâche cruciale dans le domaine des humanités numériques pour les chercheurs soucieux de valoriser le patrimoine culturel et scientifique. Dans cet article, nous avons présenté notre approche de construction de graphes de connaissance à partir des annotations manuelles et automatiques d'un corpus de textes

de zoologie antique en utilisant le thésaurus TheZoo. Dans les graphes RDF produits, nous avons pu : (i) capturer le contexte d'apparition des différentes annotations, (ii) décrire les annotations d'une manière structurée grâce aux vocabulaires standards du web sémantique, (iii) lier ces annotations manuelles avec le vocabulaire de domaine TheZoo et (iv) construire une application de visualisation et exploration de ces textes annotés. Le graphe produit permet une interrogation uniforme et avancée à l'aide de requêtes SPARQL, qui exploitent 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 à résoudre par les experts du domaine. 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. Les entités non liées à TheZoo ont fait apparaître le besoin de corriger certaines annotations et/ou réviser ou enrichir le thésaurus TheZoo. Ce travail est prévu dans le cadre du projet Zoomathia.

Une perspective connexe est l'évaluation de la tâche d'extraction et reconnaissance automatiques d'entités nommées, en tenant compte de l'impact de la qualité de la traduction automatique des textes sur cette étape et donc sur la qualité du graphe produit. Ce travail est également prévu dans le cadre du projet Zoomathia. Par ailleurs, nous souhaitons expérimenter une approche d'annotation par classification automatique de textes, en regroupant les paragraphes dans des catégories relativement larges, avec un niveau de précision élevé. Cette approche donnant une première vision globale du contenu d'un corpus de grande taille pourrait être complémentaire de celle plus fine et de ce fait plus difficilement précise d'extraction et reconnaissance d'entités nommées, qui se concentre sur des unités linguistiques plus petites. Le graphe d'annotations que nous avons produit constitue des données de très bonne qualité pour l'entraînement d'algorithmes d'apprentissage.

Une autre perspective est l'enrichissement de l'application de visualisation et exploration du corpus annoté avec une interface d'annotation semi-automatique des textes, l'extraction et reconnaissance d'entités nommées devenant une première étape du processus d'annotation, la seconde étant la validation ou édition par les experts du domaine. En plus de réduire drastiquement les irrégularités d'annotations à l'aide du système de proposition de concept directement connecté au thesaurus TheZoo, cette interface permettrait aux experts du domaine, les philologues et historiens qui ne sont pas spécialistes des modèles du web sémantique, de mettre à jour et enrichir les textes de manière simple tout en conservant un modèle de données robuste et homogène.

Bibliographie

[1] A. Barbe, M. Tounsi Dhouib, C. Faron, M. Corneli & A. Zucker, « Construction d'un graphe de connaissance à partir des annotations manuelles de textes de zoologie antique », in *IC 2023 - 34e Journées francophones d'Ingénierie des Connaissances Plate-Forme Intelligence Artificielle (PFIA 2023)* (Starsbourg, France), IC2023 : 34es Journées francophones d'Ingénierie des Connaissances, 2023.

[2] 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), nº 3, p. 1-24.

[3] E. Hyvönen, « Digital humanities on the Semantic Web: Sampo model and portal series », *Semantic Web* (2022), p. 1-16.

[4] K. P. Johnson, P. J. Burns, J. Stewart, T. Cook, C. Besnier & W. J. B. 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*, Association for Computational Linguistics, 2021, p. 20-29.

[5] J.-D. Kim, K. Verspoor, M. Dumontier & K. B. Cohen, « Semantic representation of annotation involving texts and linked data resources », *Semantic Web journal*, 2015.

[6] 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), p. 321-342.

[7] S. McCallum, « BIBFRAME development », *JLIS. it* **8** (2017), nº 3, p. 71-85.

[8] F. Michel, L. Djimenou, C. Faron Zucker & J. Montagnat, « Translation of Relational and Non-Relational Databases into RDF with xR2RML », in *11th International Confenrence on Web Information Systems and Technologies (WEBIST'15)* (Lisbon, Portugal), Proceedings of the WebIST'15 Conference, 2015, p. 443-454.

[9] R. Sanderson, P. Ciccarese & B. Young, « Web Annotation Ontology », 2017, `https://www.w3.org/TR/annotation-vocab/`.

[10] B. Tillett, « What is FRBR? A conceptual model for the bibliographic universe », *The Australian Library Journal* **54** (2005), nº 1, p. 24-30.

[11] Y. Tosaka & J.-r. Park, « RDA: Resource description & access — a survey of the current state of the art », *Journal of the American Society for Information Science and Technology* **64** (2013), nº 4, p. 651-662.

[12] A. Toulet, F. Michel, A. Bobasheva, A. Menin, S. Dupré, M.-C. Deboin, M. Winckler & A. Tchechmedjiev, « ISSA : un graphe de connaissances au service de la recherche bibliographique », *Revue des Nouvelles Technologies de l'Information* **39** (2023), p. 515-522, EGC 2023 - conférence Extraction et Gestion des Connaissances.

[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 *The First International Workshop Semantic Web for Scientific Heritage at the 12th ESWC 2015 Conference*, 2015.

Abstract. — Every day, researchers in the humanities and social sciences (SHS) are tasked with analyzing vast quantities of textual resources. Traditional lexical and quantitative research tools do not always meet the specific needs of historians and philologists. This article presents the methods, tools and services developed as part of the HisINum project to address this issue. We aim to provide a generic and reusable process for analysing ancient texts, the semantic indexing in the form of knowledge graphs of the various texts studied, and the implementation of a search and visualization service exploiting these generated graphs.

Keywords. — knowledge Graphs, Ontologies, Linked Data and Vocabularies, Semantic Annotation of Greek and Latin Texts, History of Zoology.
