Transformation et visualisation de données RDF à partir d'un corpus annoté de textes médiévaux latins

Molka Tounsi Dhouib, Catherine Faron Zucker, Arnaud Zucker, Olivier Corby, Catherine Jacquemard, Isabelle Draelants, Pierre-Yves Buard

Transformation et visualisation de données RDF à partir d'un corpus annoté de textes médiévaux latins

Molka Dhouib[1], Catherine Faron Zucker[1], Arnaud Zucker[2], Olivier Corby[3], Catherine Jacquemard[4], Isabelle Draelants[5], Pierre-Yves Buard[4]

[1] Univ. Nice Sophia Antipolis, CNRS, I3S, UMR 7271, Sophia Antipolis, France
molkatounsi@gmail.com, faron@unice.fr
[2] Univ. Nice Sophia Antipolis, CNRS, CEPAM, UMR 7264, Nice, France
zucker@unice.fr
[3] Inria Sophia Antipolis Méditerranée, Sophia Antipolis, France
olivier.corby@inria.fr
[4] Université de Caen, CNRS, CRAHAM, UMR 6273, Caen, France
catherine.jacquemard@unicaen.fr, pierre-yves.buard@unicaen.fr
[5] IRHT, Paris, France
isabelle.draelants@irht.cnrs.fr

RÉSUMÉ
Cet article présente un travail préliminaire réalisé dans le cadre du GDRI Zoomathia qui vise l'étude de la transmission des savoirs zoologiques de l'Antiquité au Moyen Âge. A partir de deux textes médiévaux latins précédemment annotés en XML, nous avons construit une ontologie RDFS et une base de données RDF, puis nous avons réalisé à partir de ces annotations sémantiques RDF/S un travail d'interrogation, transformation, extraction et visualisation de connaissances pertinentes, pour aider les chercheurs épistémologues, historiens et philologues dans leur travail d'analyse de ces textes anciens.

Mots Clés
Ontologie ; Web de données ; Recherche sémantique et fouille visuelle de textes médiévaux latins

INTRODUCTION

Le Groupe de Recherche International (GDRI) Zoomathia [1], soutenu par deux instituts du CNRS, l'INEE et l'INSHS, vise l'étude de la transmission des savoirs zoologiques de l'Antiquité au Moyen Âge, à travers les ressources matérielles (bio-restes, artefacts), iconographiques et surtout textuelles. Un des objectifs de ce projet est la construction d'un thesaurus et l'annotation sémantique des ressources répertoriées, capturant différents types de connaissances : le zoonyme ; la période historique ; la spécialité zoologique, ou sous-discipline (éthologie, anatomie, physiologie, psychologie, zootechnie…) ; le genre littéraire ou iconographique. Cette inscription est déterminante dans la production et la formalisation du savoir.

Plusieurs textes en latin ont fait l'objet d'un travail antérieur de structuration et d'annotation en XML, notamment dans le cadre du projet ANR SourcEncyMe [2], portant sur l'identification des sources des encyclopédies médiévales, et du projet Ichtya [3] de la MRSH de Caen. Dans le cadre de Zoomathia, nous avons réalisé un travail exploratoire sur deux de ces textes : le *Speculum naturale* de Vincent de Beauvais (XIII^{e} s.) et l'*Hortus sanitatis* (XV^{e} s.). Plus précisément, nous avons travaillé sur les livres 16 à 22 du *Speculum naturale* qui traitent d'animaux, et sur le traité *De piscibus* de *l'Hortus sanitatis* qui comporte 106 chapitres sur les poissons. Nous avons construit une ontologie RDFS et une base de données RDF à partir de ces textes annotés, puis nous avons réalisé à partir des données RDF/S ainsi produites un travail d'interrogation, transformation, extraction et visualisation de connaissances pertinentes, pour aider les chercheurs épistémologues, historiens et philologues dans leur travail d'analyse de ces textes. Le travail collaboratif que nous avons mené entre chercheurs en sciences humaines et chercheurs en ingénierie des connaissances a permis de recueillir les besoins, les questions que se posent les premiers auxquelles il était possible aux seconds d'apporter facilement des éléments de réponse.

Dans la partie suivante nous présentons le processus général d'ingénierie des connaissances mis en œuvre. Puis nous décrivons les connaissances formelles produites et la base de requêtes et transformations que nous avons réalisée pour répondre aux questions qui se posent lors de l'étude de la transmission des savoirs zoologiques dans les textes médiévaux latins. Finalement nous présentons quelques visualisations graphiques produites, qui facilitent l'analyse des textes.

PROCESSUS GENERAL D'INGENIERIE DES CONNAISSANCES

La première étape de notre travail a consisté à analyser les annotations XML du *Speculum naturale* et de l'*Hortus sanitatis*, et leur format, le schéma XML TEI (Text Encoding Initiative) [4], pour, d'une part, construire un schéma RDFS à partir des termes utilisés dans le balisage et, d'autre part, écrire un ensemble de règles de transformation XSL pour transformer les données XML en données RDF qui respectent le schéma RDFS produit.

Dans un deuxième temps, nous avons construit une base de requêtes SPARQL pour répondre aux questions que se posent les chercheurs en sciences humaines dans leur activité d'analyse des textes anciens, en exploitant les données RDF produites. Certaines requêtes permettent de répondre directement à certaines questions à partir des données RDF produites ; d'autres permettent de transformer les données RDF initiales pour produire de nouvelles données RDF permettant de répondre à d'autres questions posées ; d'autres requêtes enfin permettent de transformer des données RDF au format CSV pour offrir une visualisation graphique de certaines connaissances extraites, à l'aide de l'outil Gephi [5]. La Figure 1 montre le processus général d'ingénierie des connaissances mis en œuvre.

CONNAISSANCES PRODUITES

Vocabulaires RDFS

Le travail d'annotation précédemment réalisé sur le *Speculum naturale* et l'*Hortus sanitatis* a consisté à expliciter la structure logique des textes selon le standard TEI et à identifier les auteurs et/ou œuvres dits « sources » cités dans ces deux textes par leurs auteurs, deux encyclopédistes du Moyen Age. En outre, pour *l'Hortus sanitatis*, un travail supplémentaire a été réalisé d'identification naturaliste des poissons dont il est question.

A partir de ces données XML, nous avons construit manuellement une ontologie RDFS des termes du schéma TEI. Cette ontologie permet de décrire en RDF la structure logique des textes sur lesquels nous avons travaillé. D'autre part, nous avons construit une ontologie RDFS des classes zoologiques identifiées dans *l'Hortus sanitatis*. Cette ontologie a été construite automatiquement à l'aide d'une feuille de style XSL appliquée à l'annotation XML du texte. Nous l'avons manuellement enrichie en liant ses classes à celles de l'ontologie de DBpedia [6].

Base d'annotations RDF

Nous avons construit une base d'annotations RDF pour le *Speculum naturale* et l'*Hortus sanitatis* qui repose sur les ontologies que nous avons produites. Ces données RDF ont été automatiquement produites à l'aide d'une feuille de styles XSL que nous avons écrite et appliquée aux annotations XML des deux textes.

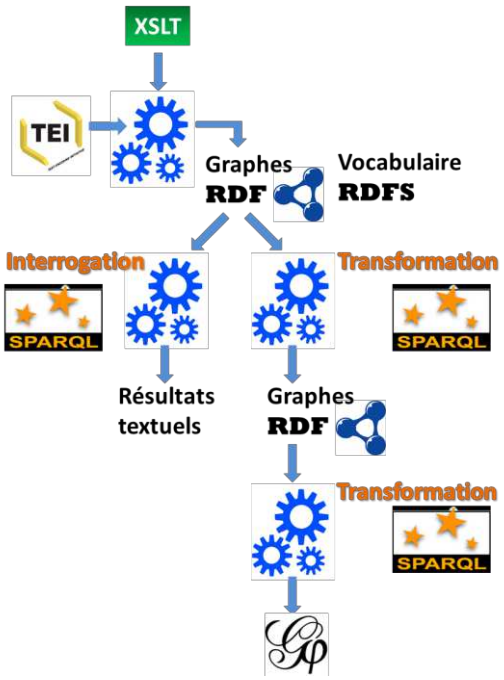

Figure 1. Processus général d'ingénierie des connaissances

Base d'annotations RDF

Nous avons construit une base d'annotations RDF pour le *Speculum naturale* et l'*Hortus sanitatis* qui repose sur les ontologies que nous avons produites. Ces données RDF ont été automatiquement produites à l'aide d'une feuille de style XSL que nous avons écrite et appliquée aux annotations XML des deux textes.

Recueil des besoins en langue naturelle

Nous avons mené un travail d'explicitation de connaissances auprès des chercheurs en sciences humaines qui nous a permis de construire une liste des questions qui se posent dans le cadre de l'étude de la transmission des connaissances zoologiques antiques dans les textes médiévaux. Ce travail a été incrémental (et n'est pas achevé). La présentation des résultats de premières requêtes a permis de montrer aux experts interviewés quels types de connaissances il était possible de produire grâce à un travail d'ingénierie des connaissances et nous avons ainsi procédé itérativement.

Voici un extrait des connaissances explicitées comme utiles dans l'étude de la transmission des connaissances zoologiques : la présence (et l'absence) de zoonymes dans les textes du corpus, le volume textuel relatif des notices consacrées à un zoonyme/animal, les mentions (et la fréquence des occurrences) des zoonymes hors de leur chapitre dédié, les lieux pour lesquels sont mentionnés les animaux (zoogéographie), les données numériques présentes dans le texte (taille, longévité, fécondité, etc.), les allonymes, ou noms alternatifs donnés à un animal (traitement de la polyonymie), le volume textuel absolu et relatif pour chaque auteur-source, la répartition des auteurs-sources utilisés selon

les familles animales traditionnelles dans les textes (poissons, oiseaux…), la présence de séries régulières d'auteurs cités.

Base de requêtes SPARQL

Pour répondre à ces besoins exprimés en exploitant les données RDF et RDFS produites, nous avons construit une base de requêtes SPARQL permettant, pour les unes, (1) d'interroger les données RDF et de produire des listes de résultats, pour d'autres, (2) de transformer les données du graphe RDF initial pour produire des nouveaux graphes RDF dont l'analyse permette de répondre aux questions relatives à l'étude de la transmission des connaissances zoologiques, et enfin, pour d'autres encore, (3) de transformer les graphes RDF produits en données CSV qui, chargées dans l'outil Gephi, permettent la visualisation des graphes de connaissances produits et ainsi l'analyse visuelle de ces connaissances par les chercheurs en sciences humaines.

Requêtes SPARQL de la forme SELECT

Les requêtes SPARQL de la forme SELECT permettent d'extraire directement du graphe de données RDF les informations recherchées. Une des requêtes les plus simples que nous avons produites est celle permettant de répondre à la question « *Quels sont les auteurs sources cités dans le texte étudié ?* ». Cette requête peut être complexifiée pour retrouver le nombre d'occurrences de chaque auteur source dans le texte considéré, retrouver les œuvres sources en même temps que leurs auteurs, ou les œuvres sources seules, dont les auteurs ne sont pas forcément mentionnés. Cette requête peut encore être modifiée pour restreindre la recherche à un livre ou chapitre particulier du texte ou bien retourner les informations recherchées pour chaque livre ou chapitre. Elle peut encore être adaptée pour étudier les cooccurrences d'auteurs sources dans plusieurs encyclopédies. Toutes ces requêtes peuvent être utilisées aussi bien dans l'étude du *Speculum naturale* que dans celle de *l'Hortus sanitatis*.

D'autres requêtes relatives aux zoonymes dont il est question dans le texte peuvent être posées sur les données RDF représentant *l'Hortus sanitatis ;* l'extraction de cette connaissance du *Speculum naturale* reste à faire, qui n'était pas présente dans les annotations XML que nous avons exploitées. La requête SPARQL la plus simple que nous avons produite pour cela est celle permettant de répondre à la question « *Quels sont les zoonymes dont il est question dans le texte étudié ?* ». Cette requête peut être complexifiée pour retrouver le nombre d'occurrences de chaque zoonyme abordé dans le texte, pour restreindre la recherche à un livre ou un chapitre ou une citation, ou un paragraphe ou pour produire l'information pour chaque livre, chapitre, citation ou paragraphe, pour étudier la cooccurrence de zoonymes.

Enfin, d'autres requêtes permettent de croiser l'analyse des sources citées et des zoonymes traités dans le texte étudié. Ainsi nous avons produit des requêtes SPARQL permettant de répondre aux questions « *Quelles sources traitent de tel zoonyme ?* », « *De quels zoonymes traite telle source* ? » et autres variantes pour produire simultanément l'information pour chaque auteur source ou chaque zoonyme, et filtrer éventuellement les résultats par livre, chapitre, etc.

Requêtes SPARQL de la forme CONSTRUCT

Comme cela apparaît dans le recueil des besoins, certaines analyses demandent de s'abstraire de la structure logique du texte décrite dans les annotations pour se concentrer sur les relations entre l'auteur du texte, ses sources et les zoonymes décrits. Pour cela nous avons écrit plusieurs requêtes SPARQL de la forme CONSTRUCT qui sélectionnent dans le graphe RDF initial les données nécessaires pour construire de tels graphes. Les graphes ainsi produits servent d'input pour l'analyse du texte.

Nous avons notamment transformé en requêtes SPARQL de la forme CONSTRUCT certaines des requêtes de la forme SELECT décrites ci-dessus, celles destinées à fournir une vision d'ensemble sur le texte, difficile à appréhender par des listes de résultats. Par exemple, pour offrir une vue synthétique sur l'importance relative des auteurs sources dans un texte, nous avons écrit une requête permettant de construire un graphe dont les nœuds sont les URI identifiant le texte étudié et les auteurs sources et dont les arcs sont des propriétés entre l'URI du texte et les URI des sources, indiquant le nombre de fois où elles sont citées. Une requête similaire permet de construire le graphe des œuvres sources citées dans le texte étudié, dont les arcs sont étiquetés par le nombre de citations de chaque source. Une autre requête permet de construire un graphe associant les auteurs sources cités dans le texte étudié et les livres ou chapitres du texte dans lesquels les citations sont faites. Notons que les auteurs nommés comme sources par les encyclopédiste ne sont pas nécessairement ceux qu'il a réellement utilisés. La plateforme SourcEncyMe vise à identifier peu à peu les sources documentaires *réelles* utilisées par l'encyclopédiste et à distinguer "sources alléguées" et "sources directes exploitées". Dans cette mesure, les résultats produits par les requêtes qui viennent d'être expliquées donneront une idée de plus en plus proche de la véritable documentation de l'encyclopédiste, par rapport aux sources qu'il avoue.

Des requêtes similaires permettent, pour *l'Hortus sanitatis* uniquement, de représenter sous forme de graphe quels zoonymes sont mentionnés dans le texte et quel est le nombre d'occurrences de chacun ; quels zoonymes sont mentionnés dans quels chapitres ; quels zoonymes sont mentionnés dans quelles citations et dans quels livres apparaissent ces citations ; etc.

Requêtes SPARQL de la forme TEMPLATE

Une visualisation adaptée des graphes RDF ainsi produits est évidemment déterminante pour que les

chercheurs en sciences humaines (1) se les approprient comme des données exploitables dans leur étude et (2) puissent faire une analyse *visuelle* de ces données. Pour cela, nous reposons sur le langage SPARQL Template, qui permet de définir des transformations de données RDF dans un format donné quelconque, sous la forme de règles de transformation [7]. SPARQL Template est une extension syntaxique de SPARQL qui peut être compilée en SPARQL standard. Nous en avons réalisé une implémentation dans le moteur sémantique Corese/KGRAM. Pour ce projet, nous avons écrit et utilisé une transformation RDF2CSV qui nous a permis de générer avec Corese/KGRAM des données au format CSV à partir des graphes RDF produits, pour présenter et manipuler ces derniers avec l'outil Gephi de visualisation et d'analyse de graphes.

RESULTATS

Dans cette partie nous montrons l'aboutissement du travail d'ingénierie des connaissances décrit précédemment, à travers quelques exemples de graphes qui ont pu être présentés aux chercheurs en sciences humaines.

Le graphe de la Figure 2 permet de visualiser les auteurs cités dans le *Speculum naturale* et le nombre d'occurrences de chacun. Les nœuds représentent les auteurs cités et les arcs sont étiquetés par le nombre d'occurrences de chaque auteur cité. Par exemple *Plinus maior* (Pline) est cité 944 fois.

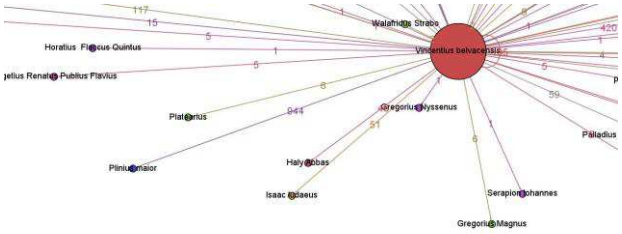

Figure 2. Visualisation des auteurs sources du *Speculum naturale* et de leur importance relative

Le graphe de la Figure 3 est une visualisation de la même information que celle présentée dans le graphe de la Figure 2 ; l'importance relative des auteurs est plus facilement appréhendable, le nombre exact d'occurrences n'apparaît pas.

Le graphe de la Figure 4 montre le nombre d'occurrences des œuvres citées dans le *Speculum naturale*. Dans ce graphe, le nombre d'occurrences des œuvres apparaît non seulement sur les étiquettes des arcs mais aussi dans la coloration des nœuds. Par exemple, les nœuds de couleur violette représentent les œuvres citées 5 fois.

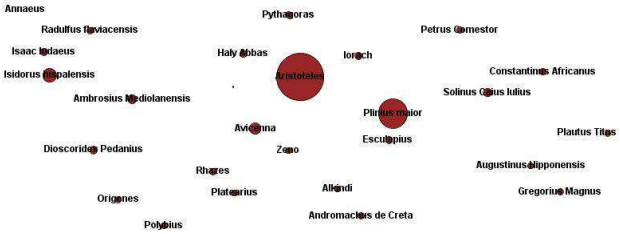

Figure 3. Visualisation des auteurs sources du *Speculum naturale* et de leur importance relative

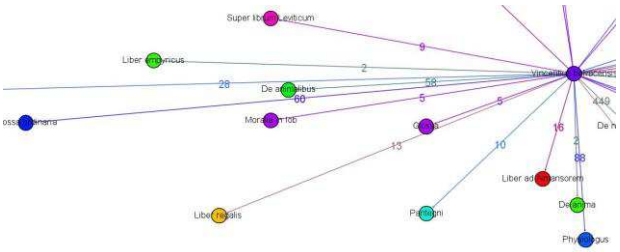

Figure 4. Visualisation des œuvres sources du *Speculum naturale* et de leur importance relative

Le graphe de la Figure 5 permet de visualiser les différents auteurs cités pour chaque livre du *Speculum naturale*. L'extrait présenté est relatif au livre 16.

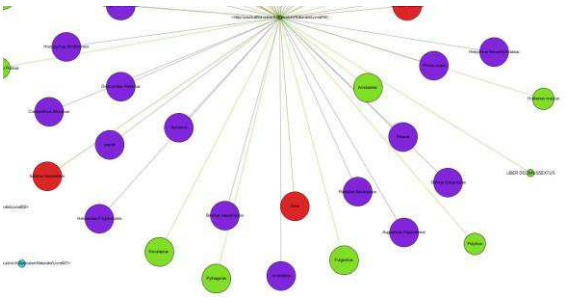

Figure 5. Visualisation de la distribution des auteurs sources du *Speculum naturale* par livre

Le graphe de la Figure 6 présente les auteurs et œuvres sources cités dans le *Speculum naturale*. Les arcs du graphe relient les auteurs à leurs œuvres. Certains nœuds sont isolés ; ils représentent un auteur ou une œuvre cité sans que Vincent de Beauvais n'indique son œuvre ou son auteur. La taille relative des nœuds rend visuellement compte de l'importance relative d'une œuvre ou d'un auteur source.

Le graphe de la Figure 7 montre quels zoonymes sont abordés dans quels chapitres de *l'Hortus sanitatis*. La taille des nœuds montre l'importance relative des zoonymes et la couleur des nœuds montre qu'il existe trois « communautés » de chapitres.

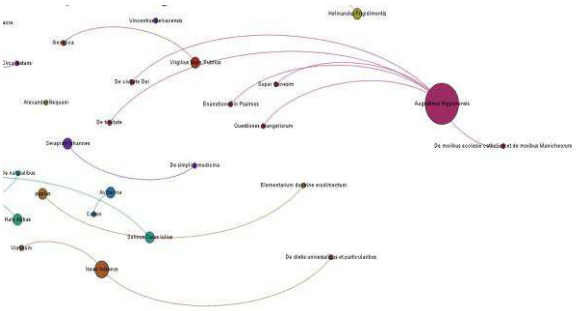

Figure 6. Visualisation des auteurs et œuvres sources du *Speculum naturale*, et de leur importance relative

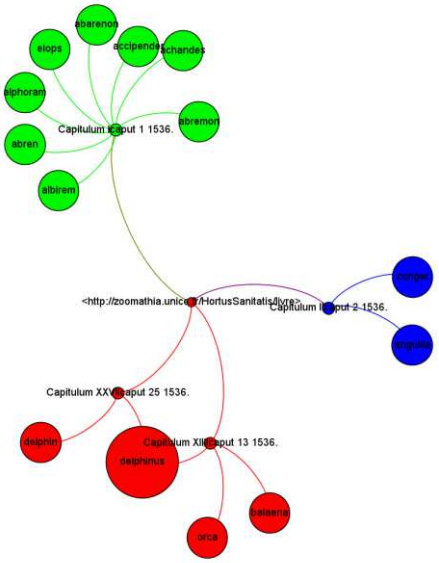

Figure 7. Visualisation des zoonymes traités dans *l'Hortus sanitatis* et de leur distribution par chapitre

Le graphe de la Figure 8 montre plus précisément où les zoonymes sont abordés : chaque zoonyme est relié aux citations dans lesquelles il apparaît, chaque citation est reliée à son paragraphe et chaque paragraphe à son chapitre.

Le graphe de la Figure 9 met en relation zoonymes, auteurs et œuvres sources. Il montre visuellement quel auteur cité traite de quel poisson et dans quelle œuvre.

CONCLUSION

Nous avons présenté dans cet article un travail préliminaire d'ingénierie des connaissances mené sur deux textes zoologiques médiévaux, travail qui a conduit à la production de connaissances destinées à supporter l'analyse de la transmission des connaissances zoologiques antiques et médiévales dans ces textes.

Les connaissances produites sont deux graphes RDF décrivant les deux textes médiévaux, un schéma RDFS capturant le vocabulaire utilisé, et une base de requêtes SPARQL permettant d'interroger ces données RDF et de sélectionner ou produire des sous-graphes RDF.

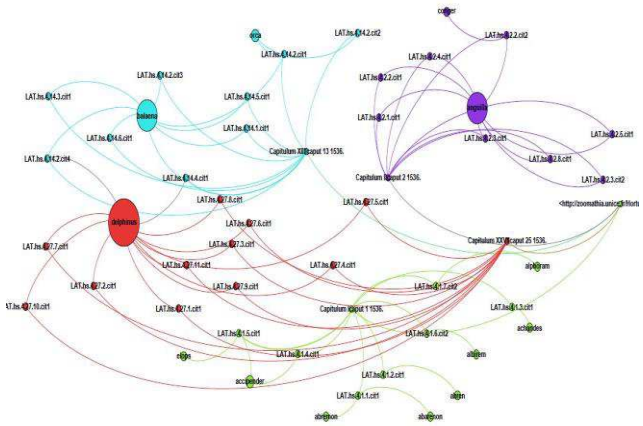

Figure 8. Visualisation des zoonymes traités dans *l'Hortus sanitatis*, de leur importance relative et de leur distribution par citation, paragraphe et chapitre

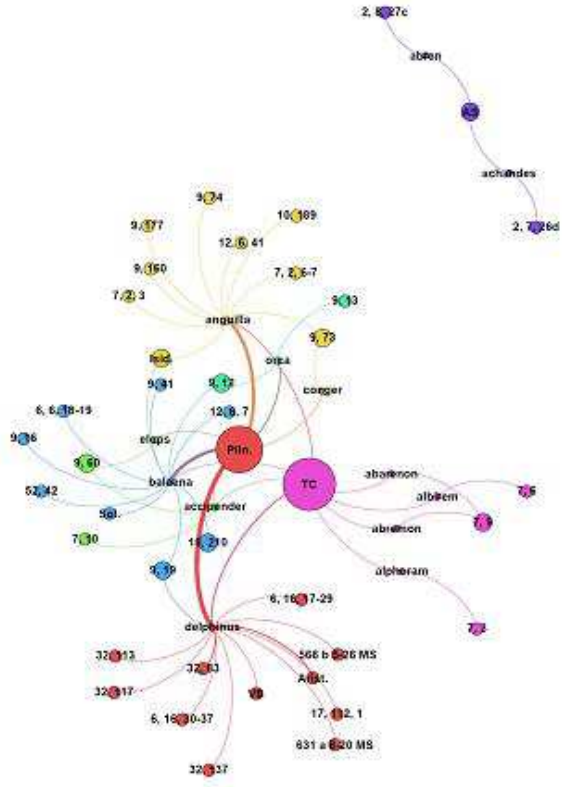

Figure 9. Visualisation des auteurs sources et des zoonymes traités dans *l'Hortus sanitatis* et des relations entre les zoonymes et les auteurs qui en traitent

Ces sous-graphes RDF capturent des connaissances dont il s'agit de permettre une visualisation supportant une analyse visuelle par les chercheurs en sciences humaines. Soulignons que cette base de requêtes est générique et pourra être utilisée pour étudier d'autres textes annotés avec le même schéma RDFS.

Plusieurs itérations entre chercheurs en ingénierie des connaissances et chercheurs en sciences humaines ont permis de recueillir les besoins et construire cette base de

requêtes de façon incrémentale. Ce travail est préliminaire ; il a permis d'amorcer une collaboration entre chercheurs de ces deux disciplines et il sera poursuivi dans le cadre du GDRI Zoomathia.

Nous projetons, d'une part, d'enrichir les connaissances produites en poursuivant le travail d'alignement du vocabulaire produit avec des vocabulaires existants et en mettant en œuvre des techniques de traitement de la langue pour extraire du texte des connaissances qui n'ont pas été annotées manuellement, à commencer par les zoonymes dans le *Speculum naturale*.

Nous projetons, d'autre part, d'enrichir la base de requêtes SPARQL actuelle qui ne répond pas encore à tous les besoins déjà exprimés et qui ne croisent pas encore les données de plusieurs textes pour une étude comparative de ces derniers.

Enfin nous projetons également de traiter à un niveau plus fin une problématique spécifique sur un corpus de textes enrichi, comme l'élaboration d'une chronologie d'apparition et d'occurrences d'espèces zoologiques étrangères à l'aire européenne, ou l'émergence et la transmission de notices sur les parasites humains dans la littérature zoologique après Aristote.

BIBLIOGRAPHIE

1. Zoomathia: http://www.cepam.cnrs.fr/spip.php?rubrique229
2. Sourcencyme: http://atelier-vincent-de-beauvais.irht.cnrs.fr/encyclopedisme-medieval/programme-sourcencyme-corpus-et-sources-des-encyclopedies-medievales
   plateforme collaborative : http://sourcencyme.irht.cnrs.fr/
3. Ichtya: http://www.unicaen.fr/recherche/mrsh/document_numerique/projets/ichtya
4. Text Encoding Initiative (TEI): http://www.tei-c.org/index.xml
5. Gephi: http://gephi.github.io/
6. DBPedia : http://fr.dbpedia.org/
7. Corby, O., Faron-Zucker C., SPARQL Template : un langage de Pretty Printing pour RDF. In Actes des *25èmes Journées francophones d'Ingénierie des Connaissances*, IC 2014, 213–224.