Thezoo: un thesaurus de zoologie ancienne et médiévale pour l'annotation de sources de données hétérogènes

Irene Pajón Leyra, Arnaud Zucker, Catherine Faron Zucker

THEZOO : un thésaurus de zoologie ancienne et médiévale pour l'annotation de sources de données hétérogènes

Présentation générale du thésaurus

Nous présentons ici (THEZOO)[1] un thésaurus en cours d'élaboration conçu pour rassembler et structurer hiérarchiquement tous les termes pouvant entrer dans une recherche simple ou complexe sur la zoologie antique et médiévale à partir des données historiques croisées de nature textuelle (corpus T), iconographique (corpus I), et archéologique (corpus A). Il doit permettre de construire une base de données hétérogènes, constituée de trois corpus, tous annotés (TIA, lien interne) selon ses concepts, et ainsi interopérables. THEZOO est formalisé dans le langage SKOS[2], la recommandation du W3C pour représenter des ressources terminologiques, ce qui le rend interopérable avec les ressources du Web de données ouvertes (lien externe), sur lequel nous projetons de le publier. THEZOO est construit à partir du gestionnaire de thésaurus Opentheso[3], développé depuis 2005 et soutenu par le TGE Adonis depuis 2009, et d'une extraction sélective de données du thésaurus PACTOLS géré par FRANTIQ [4] et servant à l'interrogation d'une base bibliographique textuelle sur l'antiquité. PACTOLS a déjà été adopté pour l'indexation de diverses bases bibliographiques, mais jamais comme outil de fouille intégrale sur des données textuelles et documentaires complexes comme nous le proposons avec THEZOO. Ce thésaurus est développé dans le contexte des activités du projet de recherche *Zoomathia* (GDRI)[5], consacré à l'étude des données archéozoologiques et à la transmission des savoirs zoologiques dans l'Antiquité et au Moyen Age. Destiné à être multisources, THEZOO est pour l'instant constitué et enrichi à partir de l'annotation de textes antiques.

Objectifs et fonctions

L'objectif du thésaurus THEZOO est de fournir un instrument spécialisé pour le traitement de toutes les données historiques sur les animaux et les connaissances zoologiques de l'aire culturelle gréco-romaine du VIII^{e} siècle av. J.-C. au XV^{e} siècle ap. J.-C. Les limites géographiques de cet espace, pour les données matérielles, sont floues, et les limites linguistiques flexibles. L'ensemble de l'espace méditerranéen (y compris

[1] Le thésaurus est consultable en accès libre à l'adresse: http://134.59.79.250/opentheso/
[2] https://www.w3.org/TR/skos-reference/
[3] http://pactols.frantiq.fr/opentheso/. Un manuel d'utilisation est disponible en ligne : http://frantiq.mom.fr/sites/default/files/manuel_dutilisateur_opentheso_v4.pdf
[4] http://frantiq.mom.fr/thesaurus-pactols
[5] Le site web général de *Zoomathia*: http://www.cepam.cnrs.fr/zoomathia/?lang=fr

africain) et l'Europe romanisée ou hellénisée sont impliqués ; et le thésaurus, nativement multilingue, intégrera des données égyptiennes et sémitiques (hébreu et arabe principalement), en raison de l'interaction de la culture gréco-latine avec ces cultures, et de la tradition médiévale. L'ontologie du thésaurus comprend les concepts zoonymiques (décrits par les labels de noms d'animaux) ainsi que les concepts abordés dans les textes à caractère zoologique liant les animaux au contexte naturel (anatomie, propriétés, géographie, peuples, histoire, etc.) ou culturel (usages, représentations, etc.).

La réalisation du thésaurus a pour objectif de permettre :

1. une recherche **conceptuelle**, **sémantique** (simple ou complexe), à partir des étiquettes critiques et des métadonnées savantes à travers la littérature ancienne et médiévale, en dépassant le niveau lexical ou lemmatique (termes présents dans les textes originaux ou les traductions) grâce à l'association de plusieurs labels synonymiques à chaque concept ;
2. une recherche **multidisciplinaire**, grâce à l'intégration d'étiquettes qui correspondent aux différentes approches sur le corpus de savoirs considéré (biologie, zootechnie, spectacles, médecine, magie, littérature…) ;
3. une recherche culturelle **multiscalaire**, à différents niveaux de précision, par combinaison de descripteurs fins non limités à la sphère zoologique mais intégrant des données culturelles générales (flore, arts, anthroponymes, types de sources, etc.) ;
4. une recherche **multi-source**, par l'unification des descripteurs appliqués aux données des textes, des images et des objets archéologiques.

Le thésaurus THEZOO est formalisé dans le langage SKOS, acronyme de Simple Knowledge Organization System, qui est une recommandation du W3C. Il s'agit d'un vocabulaire RDF qui fournit un modèle commun pour partager et lier sur le web différents systèmes d'organisation des connaissances tels que les thésaurus, les taxinomies, les systèmes de classification, les systèmes d'index. Le vocabulaire SKOS permet de représenter un système d'organisation de connaissances sous la forme d'un schéma de concepts agrégeant un ensemble de concepts, d'identifier ces schémas et concepts par des URI, de sorte qu'ils puissent être publiés et liés sur le web des données, et d'associer aux concepts des labels lexicaux dans différentes langues – en distinguant pour chaque langue un label dit "préféré", i.e., de référence, et des labels dits "alternatifs", i.e. synonymes –, de leur associer des codes classificatoires, de les documenter, de les lier à d'autres concepts pour les organiser hiérarchiquement ou en réseaux d'association, de les associer à des concepts dans d'autres schémas et de les regrouper en collections.

Dans le gestionnaire de thésaurus Opentheso, nous avons donc défini un schéma de concepts *Zoomathia*, qui rassemble l'ensemble des concepts SKOS que nous avons définis. Les rubriques (natives dans le modèle d'Opentheso) renseignées pour chaque concept SKOS sont les suivantes.

- Les labels de référence du concept dans les langues prises en charge par le thésaurus (voir *infra*). Par exemple les figures 1 et 2 montrent la description du taxon poisson-scie (Concept_Id: 106762) en français (label de référence `poisson-scie`) et en latin (label de référence `serra`) ;
- *Terme(s) générique(s)* = la liste des concepts plus généraux, au niveau hiérarchique immédiatement supérieur (ou, dans le cas des taxons zoologiques, des taxons du niveau supérieur), que spécialise le concept courant. Par exemple, le taxon `poisson-scie` spécialise les archéotaxons `aquatique`, `bête`, `cetos`, etc. ;

- *Terme(s) spécifique(s)* = les concepts qui spécialisent le concept courant, au niveau hiérarchique immédiatement inférieur ;
- *Terme(s) associé(s)* = les concepts existant dans THEZOO en relation avec le concept courant, mais non connectés hiérarchiquement[6] ;
- *Terme(s) synonyme(s)* = les labels alternatifs au label préféré du concept considéré dans chaque langue traitée par THEZOO, moderne ou ancienne. Par exemple, le concept `poisson-scie` a pour labels alternatifs en latin `pistris`, `pistrix`, `pristis` et `pristix`. Le "synonyme exact" est un terme dont l'extension est égale et identique ; le "synonyme potentiel" est un terme qui peut être utilisé aussi comme équivalent d'un autre terme ou pour la désignation d'un autre animal. Les synonymes potentiels sont commentés dans les notes d'application (voir *infra*) ;
- *Notes* = Annotations savantes distribuées en quatre rubriques :
  - *Définition*, qui provient d'une œuvre encyclopédique à validité reconnue (TLFi, *Larousse*, *Encyclopaedia Britannica*, etc.),
  - *Note d'application*, servant à expliquer des particularités rencontrées dans un passage précis d'un texte du corpus,
  - *Note historique,* de nature critique, servant principalement à contextualiser les usages du label, à expliquer son évolution sémantique (en particulier pour les zoonymes et les archéotaxons) et à indiquer les différentes acceptions et les discordances de tradition,
  - *Note éditoriale*, servant si besoin à signaler des particularités de la tradition manuscrite ;
- *Alignement*, renvoyant à des concepts dans d'autres référentiels, thésaurus ou ontologies (*e.g.*, l'ontologie de DBpédia, le référentiel TAXREF, l'ontologie du projet Biblissima, etc.) ;
- *Traduction* = à chaque concept sont associés pour le moment des labels en français, anglais et espagnol. S'y ajoutent pour les zoonymes et archéotaxons les labels latins et grecs[7] ; les figures 1 et 2 montrent ainsi la description du taxon `poisson-scie`, en français et en latin, avec les labels associés.

[6] Parfois ces associations sont explicites dans les textes (p. ex., l'association entre le tigre et la martichore, voir infra.). Mais parfois l'association est implicite et résultat de l'interprétation (p. ex., c'est le cas des animaux et des constellations de même nom).
[7] La nature multilingue du thésaurus est traitée en détail *infra*.

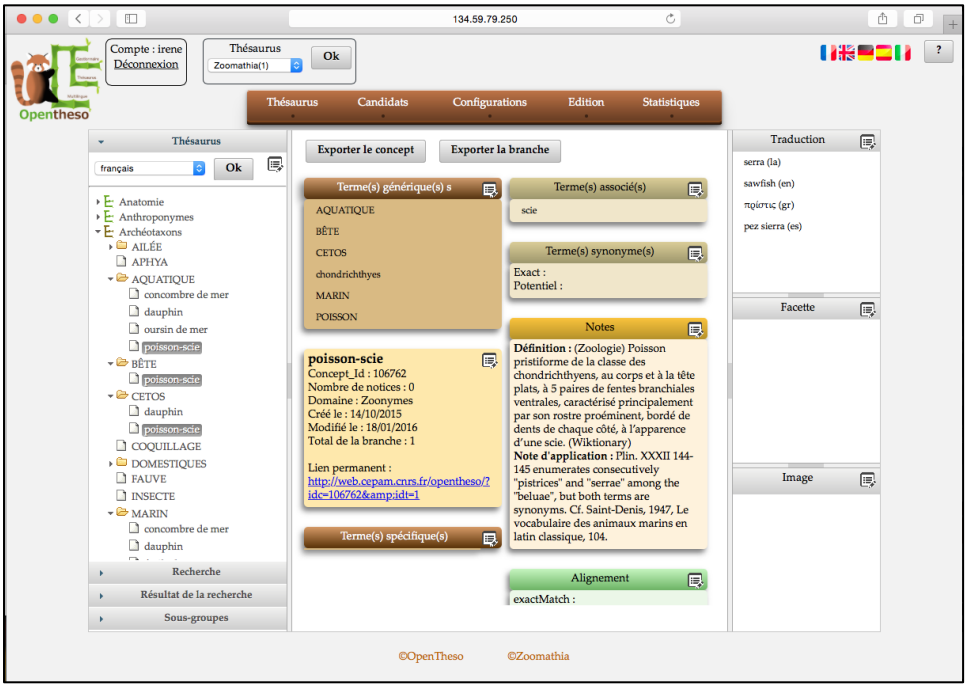


Figure 1. Description du zoonyme *poisson-scie* en français

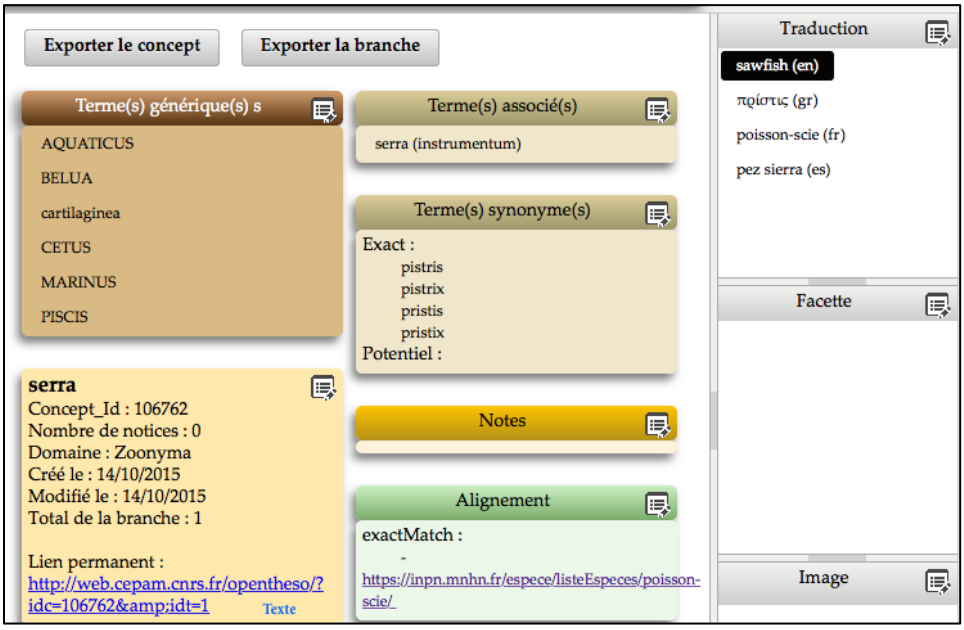


Figure 2. Description du zoonyme *poisson-scie* en latin (*serra*)

La consultation de THEZOO dans Opentheso s'opère soit en naviguant dans ses "dossiers" (concepts de niveau supérieur) qui peuvent être déroulés, soit en recherchant des termes précis, dans une langue au choix du thésaurus.

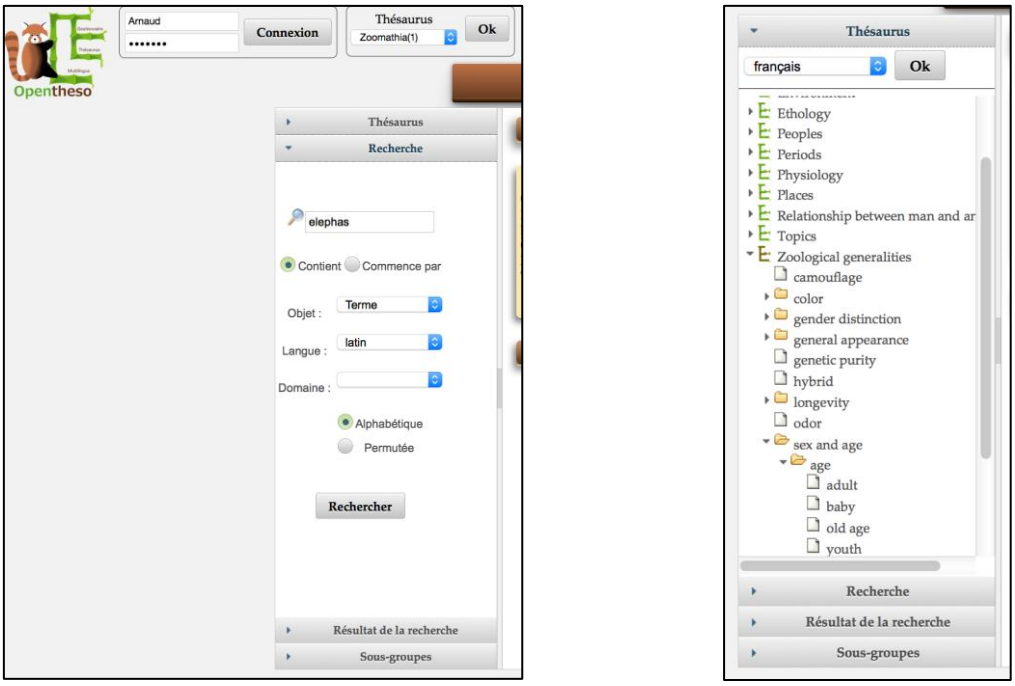


Figure 3. Les deux modes de consultation de THEZOO dans Opentheso

Le thésaurus et le travail parallèle d'annotation des textes

Le thésaurus se présente aux utilisateurs sous la forme d'une structure complexe de "dossiers" (correspondant à des concepts ayant des spécialisations), dont certains éléments ont été repris de PACTOLS quand ils étaient pertinents pour le sujet. Il est conçu pour converger avec un travail conduit parallèlement d'annotation des textes de zoologie ancienne[8]. Le texte de départ est constitué par les livres VIII-XI de l'*Histoire naturelle* de Pline l'Ancien[9], aujourd'hui annoté aux trois-quarts (VIII, IX, XI) ; les prochains textes seront les *Parties des Animaux* d'Aristote et les livres XII-XIII-XVIII de Barthélémy l'Anglais, afin de représenter les trois grands domaines du projet.

Le travail d'annotation se développe pour le moment manuellement, sur un fichier Word, et il est validé par plusieurs chercheurs. Le texte est annoté de manière souple, par sélection de segments plus ou moins longs. Les annotations portent (1) sur un mot (zoonyme, toponyme, anthroponyme, date…) ou (2) sur une ou plusieurs "unités de savoir" ou *épistémons* (i.e. des informations minimales correspondant généralement à une phrase simple) présentant une certaine unicité et autonomie. Elles s'appliquent donc à un mot, une phrase, un ou plusieurs paragraphes. Elles ne constituent pas des éléments séquentiels, mais peuvent se chevaucher totalement ou partiellement, ou être incluses les unes dans les autres. À ces unités de savoir délimitées dans le texte, on associe un ou plusieurs termes qui dénotent des concepts dans le thésaurus, à un niveau hiérarchique le plus bas possible pour capturer l'information la plus précise possible.

L'enrichissement et la structuration du thésaurus THEZOO et l'annotation des textes avec les concepts du thésaurus sont conduits de front. Ce travail implique un dialogue constant entre les co-auteurs de cet article : l'équipe du CEPAM constituée par Arnaud Zucker et Irene Pajón Leyra et une équipe d'I3S travaillant sur l'annotation automatisée et la cohérence de l'ontologie autour de Catherine Faron Zucker. Le travail conduit à de

[8] Ce travail double a fait l'objet d'un projet post-doctoral soutenu par la fondation UNICE : http://fondation-unice.org/

[9] Le choix de Pline est dû principalement à son statut de pivôt culturel entre le monde ancien et le Moyen Age.

fréquents échanges avec le concepteur du gestionnaire Opentheso, Miled Rousset[10], qui intervient sur le programme pour l'aménager en fonction des besoins et des difficultés rencontrées.

La base de textes annotés et le thésaurus sont pour l'instant déconnectés, mais l'objectif est naturellement de parvenir rapidement à les lier pour circuler de l'un à l'autre. L'utilisateur pourra alors effectuer des recherches à partir d'un texte ou du thésaurus et construire ses itinéraires de recherche en exploitant la structure hiérarchique du thésaurus et en passant d'un texte à l'autre ou d'un texte au thésaurus (et vice-versa), à partir des relations sémantiques. Contrairement aux autres bases de données courantes de textes anciens et médiévaux (e.g., le TLG [11] ou le Corpus corporum[12]) où les termes de la recherche sont les lemmata, les coïncidences littérales de caractères ou des co-occurrences, THEZOO permettra de ne pas être limité pour la langue des textes ou pour la formulation concrète des idées, et de trouver des coïncidences au niveau sémantique en traitant les relations d'inclusion de concepts, ou de comparer des données indépendamment des langues originales utilisées ou des expressions littérales servant à les exprimer.

Perspective d'enquête

Les enquêtes que doivent permettre THEZOO, une fois complété le système dans lequel il doit s'inscrire, peuvent être conduites à quatre niveaux :

1. *Lexical* (TH) : A travers le thésaurus (TH), pour accéder à une liste complète des zoonymes, obtenir un inventaire des occurrences d'un zoonyme, ou étudier les cas de polyonymie ;
2. *Thématique textuelle, iconographique ou archéologique* (TH-C1, TH-C2, TH-C3) : pour l'étude de thématiques ou d'objets dans un corpus comme le corpus textuel (C1), par exemple sur "la durée de gestation des mammifères", "la fabrication par les oiseaux de nids non arboricoles","les substances animales commercialisées", ou bien les sources iconographiques sur la chasse ou les biorestes équins ;
3. *Thématique inter-sources* (TH-C123) : pour l'étude de thématiques liant les différents corpus textuels, iconographiques et archéologiques ;
4. *Critique naturaliste des données* (TH-C123-Web) : pour l'étude en perspective de données anciennes, accompagnée d'une documentation scientifique contemporaine.

[10] GDS-Frantiq, MOM, CNRS Université Lyon 2.
[11] http://stephanus.tlg.uci.edu/
[12] http://mlat.uzh.ch/MLS/index.php?lang=0

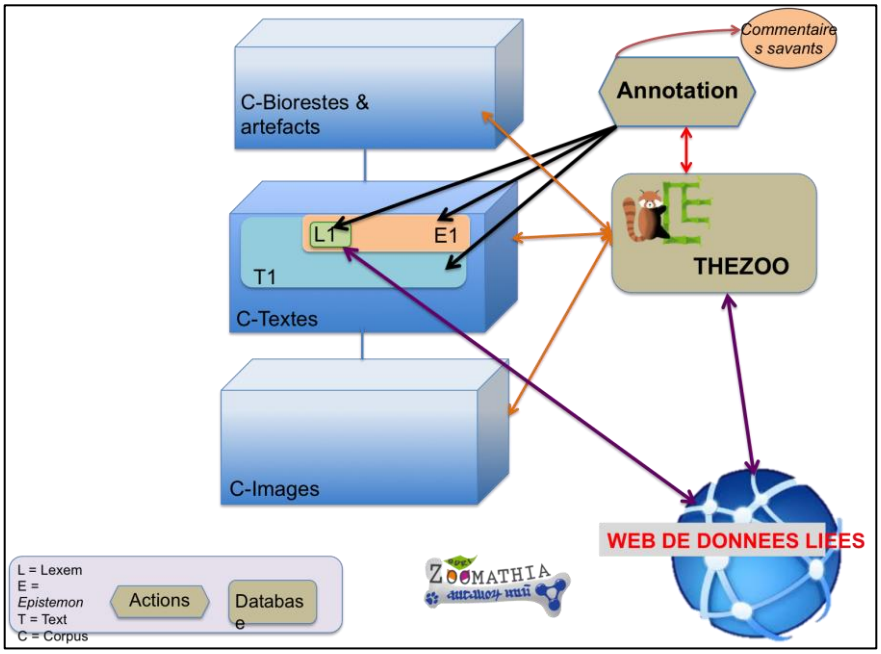


Figure 4. Système général d'annotation et de traitement sémantique.

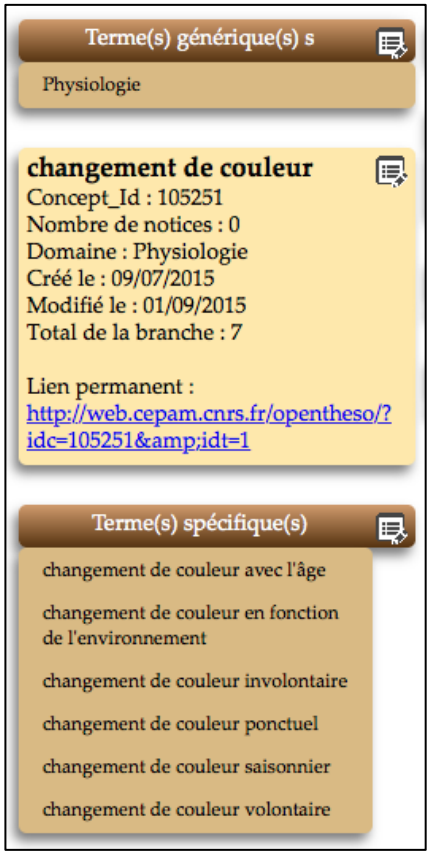
L'architecture de THEZOO

À l'intérieur du thésaurus, la construction des "dossiers" (concepts donnant lieu à des spécialisations) et des étiquettes correspondantes est le résultat de l'accord entre les chercheurs membres de l'équipe, et se fondent sur des travaux historiques et épistémologiques sur la zoologie ancienne et médiévale[13], pour l'identification des catégories pertinentes. Elle s'élabore aussi à travers la lecture critique des textes annotés

[13] Sur la terminologie zoologique, voir en particulier : J. ANDRÉ, *Les noms d'oiseaux en latin*, Paris, Klincksieck, 1967 ; I. A. BEAVIS, *Insects and other invertebrates*, Oxford, Univ. Of Exeter, 1988 ; M. DAVIES & I. A. KATHIRITHAMBY, *Greek Insects*, Londres, Duckworth, 1986 ; E; de Saint-Denis, *Le vocabulaire des animaux marins en latin classique*, Paris, Klincksieck, 1947 ; D'Arcy W. THOMPSON, *A Glossary of Greek Birds*, Londres, Oxford University Press, 1936 ; D'Arcy W. THOMPSON, *A Glossary of Greek Fishes*, Londres, Oxford University Press, 1947. Sur l'appréhension des catégories anciennes, voir Jacques BARRAU, « A propos du concept d'ethnoscience », in : *Les savoirs naturalistes populaires*, Actes du séminaire des Sommières (déc. 1983), Paris, Maison des sciences de l'homme , Cahier 2, 1984, pp. 5-12 ; Brent BERLIN, Dennis E. BREEDLOVE and Peter H. RAVEN, « Covert categories and folk taxonomies », *American Anthropologist*, 70 (2), 1968, pp. 290-299 ; Liliane BODSON, *L'interprétation des noms grecs et latins d'animaux illustrée par le cas du zoonyme* sèps-seps, Bruxelles, Académie royale de Belgique, 2009 (Mémoires de la Classe des Lettres, Collection in-8°, 3e série, XLIX, n.°2062) ; id. « Les connaissances zoologiques de l'Antiquité grecque et romaine : aperçu de leurs spécificités fondamentales et de leur actualité », *Bulletin de l'Association Guillaume Budé*, 2010 (1), pp. 53-82 ; id. « Zoological Knowledge in Ancient Greece and Rome », Gordon L. CAMPBELL (éd.), *The Oxford Handbook of Animals in Ancient Thought and Life*, Oxford : Oxford University Press, 2014, pp. 556-578 ; Barbara CASSIN, Jean-Louis LABARRIERE, Gilbert RHOMEYER DHERBEY (éd.), *L'animal dans l'antiquité* [rencontre internationale, 18-22 octobre 1994, Paris, Sorbonne et Museum National d'Histoire naturelle], Paris : Vrin, 1997 ; Jacques VOISENET, *Bêtes et hommes dans le monde médiéval: le bestiaire des clercs du Ve au XIIe siècle*, Turnhout : Brepols, 2000 ; Arnaud ZUCKER, *Les classes zoologiques en Grèce ancienne* Aix-en-Provence, 2005 ; id., *Aristote et les classifications animales*, Louvain : éditions Peeters, 2005.

(Pline), et la réflexion sur des entreprises d'annotation réalisées sur des corpus voisins, en particulier par des membres du GDRI (projet SOURCENCYME [14] de corpus des encyclopédies médiévales latines, projet ICHTYA [15] de corpus de traités d'ichtyologie, projet I2AF [16] sur les inventaires archéozoologiques et archéobotaniques de France), ainsi que la consultation de chercheurs d'autres domaines scientifiques connexes (projet TAXREF [17] sur l'inventaire du patrimoine naturel). La structuration des concepts du thésaurus est encore en chantier, et la hiérarchie est périodiquement révisée au cours du processus d'annotation, au fur et à mesure que se présentent des nouveaux cas concrets à annoter.

Au premier niveau hiérarchique (N1), i.e. sous la racine de l'arbre THEZOO, figurent des concepts correspondant à des types de données ou des perspectives hétérogènes (`Anatomie` ; `Anthroponyme` ; `Archéotaxon` ; `Description générale` ; `Environnement` ; `Information zoologique` ; `Lieu` ; `Peuple` ; `Physiologie` ; `Période` ; `Relation homme-animal` ; `Zoonyme` ; `Ethologie`). Cette disparité provient d'un désir de flexibilité qui pose néanmoins encore un problème de cohérence structurelle. Certains de ces concepts recouvrent des connaissances contextuelles (`Anthroponyme` ; `Environnement`, `Lieu`, `Période`), tandis que d'autres ciblent les espèces animales (`Zoonyme`) et les registres de savoir (`Archéotaxon` ; `Anatomie` ; `Ethologie`…). Nous avons cherché à ne pas privilégier une approche particulière et à permettre ainsi un large spectre de recherches thématiques par les utilisateurs susceptibles de se focaliser sur des paramètres secondaires liés au savoir zoologique.



[14] http://sourcencyme.irht.cnrs.fr/
[15] https://www.unicaen.fr/puc/sources/depiscibus/accueil
[16] http://bbees.mnhn.fr/I2AF
[17] https://inpn.mnhn.fr/programme/referentiel-taxonomique-taxref

Figure 5. Spécialisation du concept de changement de couleur

Au-delà des enquêtes portant sur des aspects de science naturelle le thésaurus se prête à une visée littéraire, artistique, historique ou géographique, ou à des recherches sur les acteurs de la construction du savoir (information zoologique), et il permet de combiner ces angles. Notre intention a été aussi de ne pas réduire l'information aux cadres modernes et à préserver le système intellectuel et les caractères des *épistémès* antiques ou médiévales et des configurations culturelles variables dans le temps et selon les traditions disciplinaires. Cela nous a conduits en particulier à répertorier à part les termes classificatoires indigènes (`Archéotaxon`) et les catégorisations anciennes. Mais cette attention ne porte pas seulement sur les modes de classification et doit s'étendre aux cadres perceptuels et interprétatifs. Ainsi, sous le concept `Physiologie` (N1), un concept `Changement de couleur` (N2) comporte initialement quatre spécialisations (N3) correspondant à la perception antique, même si la typologie n'est pas explicite dans les textes ([a] `changement de couleur avec l'âge` ; [b] `changement de couleur en fonction de l'environnement` ; [c] `changement de couleur saisonnier` ; [d] `changement de couleur ponctuel`). Cet exemple illustre une double difficulté méthodologique : (1) ces distinctions ne sont pas régulières dans les textes et un seul texte peut justifier leur adoption sans qu'elles reflètent l'ensemble du corpus couvert par le thésaurus ; (2) le changement de couleur peut être [e] volontaire ou [f] involontaire, et cette distinction (manifeste dans l'opuscule théophrastéen sur *Les animaux qui changent de couleur*) est tangente par rapport à la précédente ([a-d] renvoyant à l'occasion du changement) et ne peut être hiérarchisée. La solution adoptée a consisté à l'ajouter aux précédentes, au niveau N3. Il en va de même pour les espèces qui sont diversement catégorisées, selon la perspective : la classification anatomique des animaux ne correspond pas à la classification écologique ou aux diverses classifications fonctionnelles (sacrifice, chasse, gastronomie, etc.).

Cette imbrication, que résout théoriquement l'héritage multiple dans le modèle SKOS, s'avère conceptuellement problématique, en particulier en raison des décalages entre la classification moderne et *les* classifications anciennes des espèces. Les difficultés que soulève cette coexistence de points de vue anciens et modernes sont épistémologiquement fondamentales, et exigent une réflexion méthodologique approfondie. De fait, la disparité générique des pièces du corpus et son extension temporelle et culturelle considérable invitent à une grande prudence dans les choix de terminologie et de catégorie, car tout texte, pourrait-on dire en forçant à peine la réalité, diverge plus ou moins d'avec le reste de la documentation, non seulement dans les informations naturalistes, mais dans l'organisation et la caractérisation des espèces et des phénomènes. Notre volonté d'offrir avec THEZOO un portail à vocation multidisciplinaire a ajouté une exigence supplémentaire, qui explique le caractère provisoire de certaines décisions prises dans la structuration des concepts.

L'héritage multiple (ou poly-hiérarchie) permet de lier un concept à plusieurs concepts plus généraux (nommés "terme générique"), c'est-à-dire de le situer à plusieurs emplacements de la structure hiérarchique. Ainsi le concept `chauve-souris` spécialise quatre concepts, de niveaux différents : `volant`, `ailé`, `oiseau` et `chiroptera`.

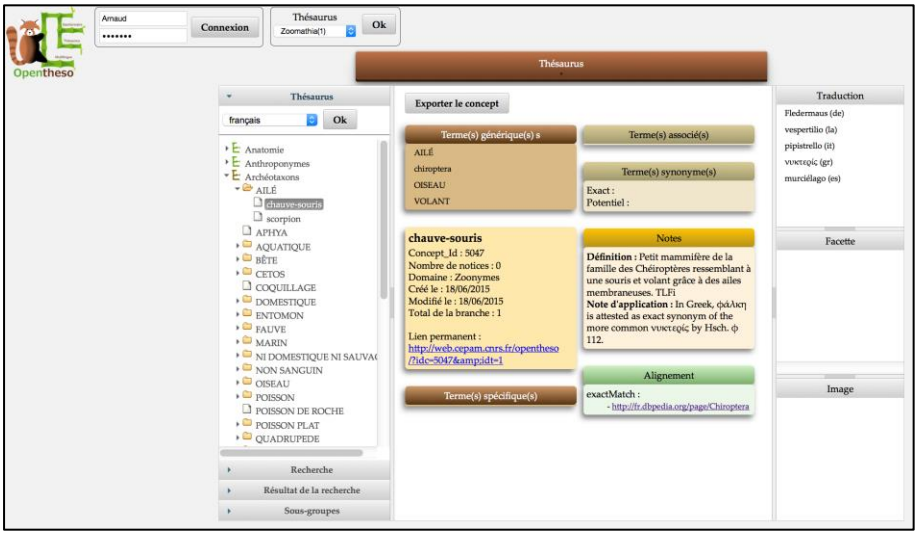


Figure 6. Occurrence du concept `chauve-souris` comme spécialisation de quatre concepts : `volant`, `ailé`, `oiseau` et `chiroptera`

LES PRINCIPES METHODOLOGIQUES

Classes modernes vs classes anciennes : `Zoonyme` vs `Archéotaxon`

Depuis le XVIII[e] siècle la culture occidentale dispose d'un cadre de référence général pour la classification systématique des espèces vivantes, animales ou végétales : la taxinomie de Linné, amorcée par Peter Artedi. Cependant la systématique est un secteur en perpétuelle évolution et la nomenclature n'est pas unifiée, même au niveau des taxa superspécifiques[18]. A cette instabilité actuellement entretenue par la communauté des systématiciens, et aux problèmes que celle-ci rencontre[19], se superpose la disparité des méthodes et des brins de classements anciens. La classification moderne des espèces animales ne correspond pas aux classements rencontrés de manière générale en ethnobiologie[20], ni aux classes anciennes qui s'en tiennent en général aux formes de vie. En outre, dans la littérature scientifique ancienne il n'y a pas de cadre référentiel général, mais plusieurs systèmes coexistent, qui sont dérivés de différents points de vue : l'habitat (terrestre, aquatique), le rapport avec l'homme (sauvage ou domestique), les caractéristiques physiologiques ou anatomiques (comme le fait d'avoir du sang ou non, le mode de locomotion, la possession de certaines phanères ou attributs anatomiques), etc.[21].

THEZOO propose deux concepts (`Zoonyme`, `Archéotaxon`) sous lesquels répertorier les *noms d'animaux*, puisque c'est bien par les *zoonymes* que sont appréhendés les animaux. Cette précision est importante, car les thésaurus zoologiques contemporains ont un fonctionnement différent. Bien qu'elle soit aussi en apparence

[18] Voir Loïc MATILE, Pascal TASSY et Daniel GOUJET, « Introduction à la systématique zoologique. Concepts, principes, méthodes », *Biosystema*, 1, 2013 [1987].

[19] Voir par exemple *Journal of Zoological Systematics and Evolutionary Research* (http://onlinelibrary.wiley.com/journal/10.1111/(ISSN)1439-0469), ou pour une problématique ciblée : http://horizon.documentation.ird.fr/exl-doc/pleins_textes/ed-06-08/010037362.pdf

[20] Scott ATRAN, *Fondements de l'histoire naturelle: pour une anthropologie de la science*, Paris : Editions complexes, 1986 ; Douglas L. MEDIN, et Scott ATRAN, *Folkbiology*, Cambridge : MIT Press, 1999.

[21] Arnaud ZUCKER, *Les classes zoologiques en Grèce ancienne*, Aix-en-Provence, 2005.

"lexicale", l'entrée de référence de ces thésaurus est constituée par un lexème quasi formel (binom latin de la nomenclature dérivée de Linné), situé à un niveau conceptuel particulier, entre dénomination populaire et idéal spécifique. Ces entrées se présentent pour ainsi dire comme simultanément méta-linguistiques et méta-réalistes. L'accompagnement iconographique accroît encore la transparence attribuée au vocable savant dans sa désignation du réel, que l'idiome employé (le latin), en raison de sa neutralité et étrangeté linguistiques, contribue à détacher de la perception immédiate. Tous les zoonymes prémodernes (concernés par THEZOO) sont des dénominations courantes et plus ou moins populaires[22], dont l'usage n'est pas contrôlé par une norme naturaliste et une définition complète et systématique (ni *a fortiori* phylogénétique), et qui se répandent textuellement avec des variations fréquentes d'extension, des équivoques et des malentendus auxquels une nomenclature standard est, elle, parfaitement imperméable. Un même terme (γαλῆ - *galè*) pouvant désigner selon les textes une belette, une martre, une fouine, un putois ou un chat (sans parler du poisson homonyme), il est nécessaire de dissocier le plan lexical et le plan naturaliste… et tout autant de les lier ! C'est pourquoi un même zoonyme est susceptible de se trouver, abstraction faite des cas de polyonymie, sous différents concepts spécialisant `Zoonyme` ou `Archéotaxon`. Il faut noter, dans le contexte du projet *Zoomathia*, que le type d'entrées lexical est adapté à la documentation littéraire, principale composante du savoir zoologique ancien, mais qu'elle est moins adéquate pour traiter la documentation iconographique (dont le répertoire devrait reposer sur une typologie des formes) ou les bio-restes (associables à un genre/espèce biologique). Le concept `Zoonyme` recouvre tous les noms d'animaux rencontrés dans les textes, en langue originale et en diverses traductions. La traduction, qui est le plus souvent claire ou explicite entre les *langues* anciennes (grec/latin), exige pour les langues modernes une identification naturaliste. De manière à articuler le savoir ancien avec les données naturalistes contemporaines, et en permettre une évaluation fine, nous avons eu recours à un système taxinomique contemporain comme cadre de répartition des concepts zoonymiques. Nous avons adopté la norme ITIS, qui est utilisée par Wikipedia et constitue un standard taxinomique complet et très largement adopté[23] ; nous envisageons également la possibilité d'utiliser TAXREF comme cadre référentiel.

Le zoonyme correspond généralement à l'ultime niveau de la hiérarchie (*equus*, *muraena*, etc.), celui du générique-spécième[24]. Parfois il est à un niveau supérieur quand il a une valeur supra-générique (*ostrea* = bivalve ; *canis marinus* = requin), en lui conservant son maximum d'amplitude, y compris pour des termes apparemment clairs (*aquila*) qui font l'objet de subdivisions explicites dans les textes anciens. Lorsque l'assignation d'un zoonyme ancien à un niveau taxinomique est problématique, car incertaine ou discordante dans la littérature, nous l'avons introduit au plus bas niveau raisonnable d'intégration. Ainsi la *lampetra* (généralement traduit par "lamproie", qui est un genre dans la taxinomie moderne)[25] est placée non pas au niveau de la sous-famille des Petromyzontinae mais au niveau des classes, comme spécialisation de Agnatha (qui est une superclasse). Lorsqu'un même terme (ὄστρεον, *ostrea*)

[22] On ne méconnaît pas l'existence en grec ou en latin de lexèmes zoonymiques savants (des γλῶσσαι), mais ceux-ci ne constituent pas les labels de référence retenus et apparaissent comme termes synonymes.
[23] http://www.itis.gov/index.html
[24] Scott ATRAN, *Fondements de l'histoire naturelle: pour une anthropologie de la science*, Paris : Editions complexes, 1986.
[25] http://dbpedia.org/page/Lampetra

correspond à deux niveaux taxinomiques différents, en raison d'usages métonymiques (huître/coquillage bivalve), un concept est créé au plus haut niveau hiérarchique (classe Bivalvia) ; comme il est impossible de déterminer pour tous les usages du terme dans les textes son extension naturaliste, le zoonyme est identifié comme un concept unique (`ostrea`), et une note critique est chargée de préciser les acceptions graduées du lexème. Les cas de polyonymie[26], tel *asinus* (= âne, cloporte, merlu) donnent lieu, quant à eux, à trois concepts différents (`asinus1`, `asinus2`, `asinus3`).

La structure hiérarchique issue du concept `Zoonyme` reflète donc approximativement la classification taxinomique de Linné, ou au moins une version simplifiée, adaptée à la faune connue dans l'Antiquité. Il est conçu pour aider l'utilisateur à trouver les espèces traitées dans les textes, à travers la classification moderne généralement acceptée et connue. Les zoonymes peuvent également figurer dans la structure hiérarchique issue du concept `Archéotaxon` consacrée aux classements historiques, lorsque les textes rangent explicitement le noms parmi les exemples d'une classe (par exemple le lion parmi les animaux sauvages)[27].

Néanmoins, le concept `Zoonyme` n'a pas pour fonction d'amalgamer les catégories anciennes et les classes modernes ou les espèces que nous reconnaissons aujourd'hui, mais de refléter le savoir ancien tel qu'il est mis en forme. Il faut donc traiter les décalages entre données historiques et données naturalistes sans sacrifier les catégories anciennes ni leur correspondance moderne, même si celle-ci est approximative et doit être commentée dans les annotations associées aux concepts. L'adoption d'un référentiel moderne comme cadre structurel principal pour l'indexation des zoonymes peut sembler discutable, car il ne correspond pas aux modes de classement pré-modernes. Elle s'impose néanmoins pour trois raisons principales qui ont conduit à écarter le classement alphabétique ou les groupements historiques : (1) il n'existe pas de taxinomie unique et commune dans les périodes considérées et pour les corpus traités permettant un groupement par classes "historiques" ; (2) un groupement alphabétique aurait, d'un autre côté, annulé les affinités communes et globalement représentées dans des taxons transhistoriques (mammifères, ongulés, oiseaux, coquillages…) ; (3) l'objectif du thésaurus est l'articulation du savoir avec le savoir naturaliste moderne et les données scientifiques publiées sur le Web de données pour lequel la taxinomie moderne est conventionnelle. Ce principe de double identification pose un grand nombre de difficultés pratiques pour des lexèmes obscurs ou vagues (surtout parmi les arthropodes) sans référent zoologique identifié, voire identifiable. Un exemple est fourni par le manticore, ou martichore. Cet animal est ainsi décrit dans les textes selon les informations données par Pline et d'autres auteurs[28] : il a trois rangées de dents, est capable d'imiter la voix humaine et a parfois un visage humain, et il a un aiguillon venimeux dans la queue. Cette description ne correspond à aucun animal connu dans la faune actuelle. Nénamoins, dans la zoologie ancienne c'est un *animal* —au moins littéraire— courant et autonome. Le concept zoologique est considéré par Pline, Ctésias ou Élien sur le même plan que des animaux comme la girafe ou l'éléphant.

[26] Arnaud ZUCKER, « Sur l'extension de certains noms d'animaux en grec : les zoonymes pluriels », *Métis*, NS 4, 2006, pp. 97-122.

[27] Les archéotaxons apparaissent en majuscules dans le thésaurus.

[28] L'information originale procède de Ctésias de Cnide: *FGH* 688, fr. 45 (15), ap. Phot. *Bibl*. 72, 45b 30 sq. Voir aussi Aristote, *HA* 501a 24-b 1; Elien, *NA* IV.21. Voir Pietro LI CAUSI, *Sulle tracce del manticora. La zoologia dei confini del mondo in Grecia e a Roma*, Palerme : Palumbo, 2003.

L'identification du martichore comme le résultat d'une description exagérée du tigre[29], et surtout les éléments de description anatomique fournis par les textes ont conduit à placer ce taxon sous celui des Félidés (*pantherinae*), mais il conserve un statut particulier : un martichore n'est pas un tigre. La rubrique "terme associé" permet de signaler les assimilations proposées.

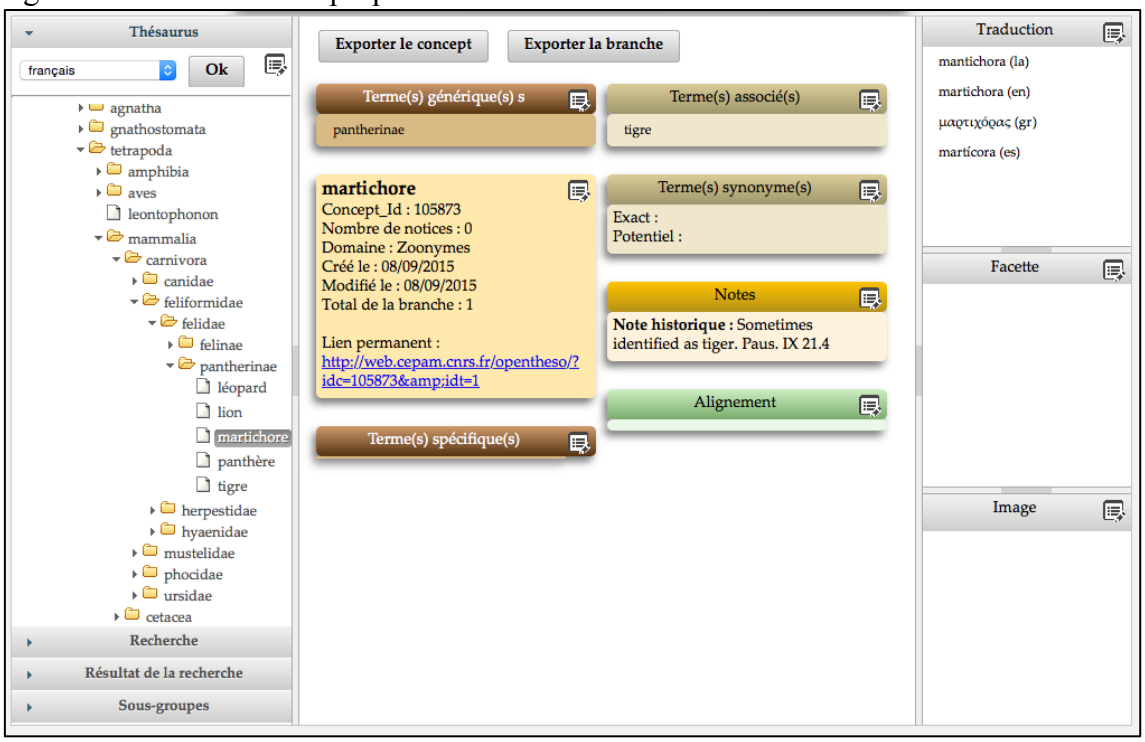


Figure 7. Description du zoonyme `martichore`, un animal non identifié

Les exemples de ce type faunique ambigu d'êtres exagérés ou déformés, et souvent qualifiés négativement de semi-légendaires abondent : outre le martichore, les textes zoologiques ont fait exister, sur le papier, des animaux comme l'unicorne, le griffon, le phénix, le basilic, etc. ; souvent aussi des erreurs d'interprétation sur les sources ou des malentendus ont généré des êtres qui ont eu une certaine vigueur (et vie !) dans la tradition scientifique postérieure. THEZOO intègre ces zoonymes, et les situe, comme les termes à spectre variable, au niveau le plus bas possible dans l'arborescence, sans aller au-delà des éléments de description fournis.

L'éventail thématique

Un thésaurus du savoir zoologique ne peut se réduire à un lexique zoonymique. En effet, le discours sur les animaux s'inscrit dans une configuration culturelle et idéologique particulière, et compose un véritable système de représentations. C'est pourquoi le thésaurus comprend d'autres concepts de niveau supérieur (N1), qui manifestent ce souci d'une prise en compte des données biologiques *et* culturelles, naturalistes *et* contextuelles, descriptives *et* interprétatives. Ils permettent d'accéder sélectivement à la gamme des registres de savoir développés dans les textes, par une approche transversale par rapport au monde animal. L'ordre retenu dans l'exposé des auteurs anciens est en effet variable, et il peut suivre une logique spécifique (traitement par animal), ou thématique voire problématique (traitement par phénomène typique ou sujet). Pline, notre auteur témoin, structure son œuvre selon les grandes formes de vie dans quatre livres (animaux terrestres [VIII], animaux marins [IX], oiseaux [X], insectes [XI]) ; et il

[29] C'est l'opinion de Pausanias (*Périégèse* IV.21.4-5).

présente successivement par brèves monographies les dossiers des différents animaux/zoonymes. Cet ordre facilite la consultation ciblée de l'œuvre, et il est reconduit dans un grand nombre d'ouvrages postérieurs de type encyclopédique. Mais il adopte aussi une perspective thématique, lorsqu'il fait, plus tard dans son œuvre (livres XXVIII-XXXV), l'inventaire des produits animaux utilisés comme substances *thérapeutiques*.

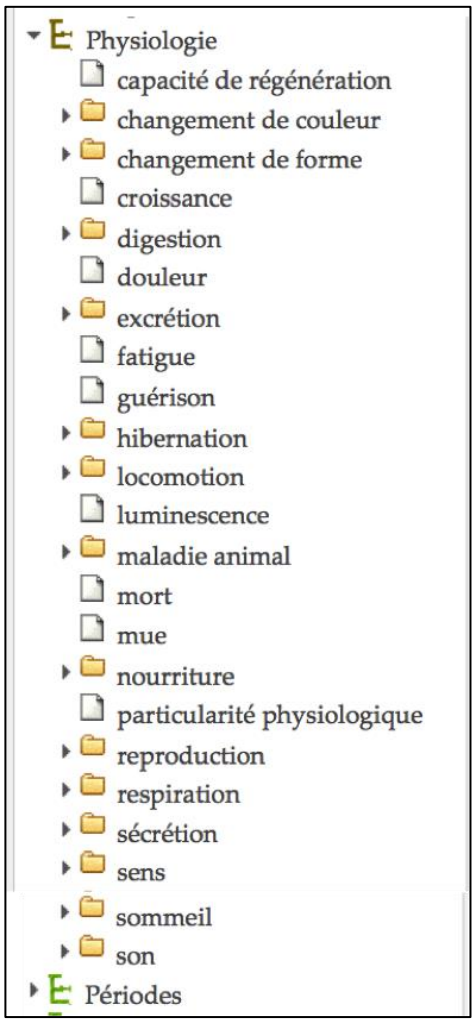


Figure 8. Liste des *spécialisations* du concept `Physiologie`[30]

Aristote, le fondateur de la biologie comme domaine scientifique, accorde pour sa part une importance cruciale à l'étude théorique et comparative des options anatomiques ou physiologiques illustrées par les espèces animales. Il est donc essentiel de pouvoir aborder directement cette dimension systématique du discours zoologique. L'expression des différences anatomiques et physiologiques a reçu une attention spéciale à travers les concepts `Anatomie` (N1) et `Physiologie` (N1), que complète un concept `Description générale` (N1), où sont rassemblées toutes les données phénoménales qui ne relèvent pas d'une description des parties (couleur, taille, âge, expressivité…). Les descriptions physiques données par les textes sont segmentées en unités et réparties dans les concepts de second niveau (N2) `anatomie externe` et `anatomie interne`, qui spécialisent le concept `Anatomie`. Les différentes

[30] Dans l'interface de visualisation d'Opentheso les concepts correspondant à une icône de fichier (et non de dossier) sont ceux qui n'ont (actuellement) pas de spécialisations.

expressions des fonctions vitales constituent les spécialisations du concept `Physiologie`.

La description des particularités anatomiques est décorrélée des zoonymes, et l'utilisateur peut donc envisager un détail anatomique, soit comme élément de la description d'un animal précis, soit comme caractère morphologique générique (indépendamment d'un animal visé), soit comme trait lié à la réalisation d'une certaine fonction. Dans la structure de THEZOO les espèces animales, les fonctions biologiques, les caractéristiques anatomiques, etc. sont donc au même niveau hiérarchique.

Les aspects relatifs au comportement et au caractère des animaux (ἦθος) sont traités dans la hiérarchie issue du concept `Ethologie` (au niveau N1), un champ qui dans les sources anciennes et médiévales apparaît comme exceptionnellement riche, et qui va au-delà des limites de la science biologique pour impliquer des aspects d'ordre symbolique, littéraire ou poétique ; les spécialisations de ce concept mettent en valeur des associations culturelles que la tradition occidentale a établies avec les diverses espèces, (comme la valeur du lion comme paradigme du courage, ou de l'éléphant, comme modèle de mémoire) ; un concept `comparaison avec l'homme` (au niveau N2) qui inventorie les rapprochements explicites (entre un animal et l'homme) focalise sur une forme de discours qui déborde la littérature spécialisée.

Enfin, les facteurs externes qui déterminent la vie des animaux et le rythme du développement de leurs fonctions (l'habitat, le climat et son influence, les périodes concrètes de l'année quand ils hivernent ou se reproduisent, etc.) sont exprimés sous le concept `Environnement` (de niveau N1).

L'approche historique et épistémologique : faits et acteurs du savoir

La description des caractéristiques physiques, physiologiques, ou éthologiques des animaux ne donne qu'un aperçu factuel et donc une image incomplète de l'état des connaissances exposées dans les textes : il est nécessaire de prêter attention non seulement aux faits, mais aussi aux acteurs de ces savoirs et aux sources d'information, qui conditionnent le discours et déterminent son statut. Ce contexte de formation de l'énoncé naturaliste concerne la source utilisée par un auteur (une autorité intellectuelle, un expert technique, une tradition orale regardée comme populaire, une image ou représentation artistique,...), les conditions sociales de production de l'observation ou du savoir (au cours d'une expédition militaire, pendant des activités techniques ou commerciales impliquant des animaux, à travers l'expérience personnelle de l'auteur, etc.) et le cadre spatio-temporel (les lieux géographiques concernées par l'information et les figures historiques impliquées). Ces aspects sont couverts principalement à travers les concepts (de niveau N1) `Lieu`, `Période`, `Peuple`, `Anthroponyme`, `Information zoologique` et `Relation homme-animal`. Les trois premiers ont une déclinaison classique. Dans `Anthroponyme`, on trouve à la fois les noms des autorités intellectuelles citées comme sources des informations, et les figures historiques que les textes mentionnent en relation avec les animaux (le premier homme qui a découvert l'utilisation d'un certain produit animal, le protagoniste d'une anecdote impliquant des animaux, l'homme politique qui a célébré des jeux avec des espèces exotiques, l'inventeur d'une manière de les préparer dans la cuisine, etc.). Mais le concept `Information zoologique` comporte aussi une spécialisation de niveau N2 `période` (date d'apparition de la donnée) et une autre `source` (origine humaine de la donnée), dans laquelle des figures historiques mentionnées dans les textes et recensées dans `Anthroponyme` apparaissent en qualité cette fois de source scientifique. Le contexte anthropozoologique de production de l'information est également exprimé sous le concept `Relation homme-animal`, où les interactions

sont énumérées dans leurs manifestations négatives (dans le sous-concept `dommage à l'homme`) et positives (dans le sous-concept `utilisation de l'animal`): l'information relative à la pratique d'activités techniques connexes avec le monde animal, comme le soin des animaux domestiques, la chasse, la pêche ou l'obtention de produits d'origine animale et leur commerce trouve ici sa place.

Le multilinguisme

Un aspect fondamental de THEZOO est son caractère multilingue qui se manifeste à la fois au niveau des textes que la base de donnés documentaire inclura et au niveau des langues dans lesquelles les labels des concepts sont exprimés. Le programme *Zoomathia* se veut translinguistique : il dépasse les frontières des langues dans lesquelles le savoir zoologique est transmis et prévoit l'établissement d'un corpus de textes multilingue sur lequel opérera THEZOO. L'annotation, qui consiste en l'utilisation d'étiquettes indexées qui ne correspondent pas nécessairement à des unités lexicales ou sémantiques explicites dans le texte permet d'établir des connexions entre idées similaires, indépendamment de leur formulation linguistique. Cette caractéristique ouvre l'utilisation de la base à un grand éventail d'approches multidisciplinaires, en proposant un outil qui donne accès aux textes à des chercheurs de diverses formations philologiques (hellénistes, latinistes, hébraïstes, arabistes) ainsi qu'à des chercheurs sans aucune formation philologique dans les langues des textes anciens et médiévaux.

Le travail d'annotation initial sur le texte de Pline est conduit en anglais, mais l'indexation dans le thésaurus se fait systématiquement en trois langues modernes (français, anglais et espagnol), et devrait se développer au moins en italien, en allemand, et en grec moderne[31]. Cette caractéristique implique que la création des concepts et de leurs labels est réalisée en tenant compte des systèmes de lexicalisation et de classification propres aux trois langues. Ce croisement offre l'occasion, au cours du processus d'extension de la palette linguistique, de modifier et d'enrichir la structure même des concepts. La relation entre les concepts de la zoologie ancienne et les concepts présents dans les langues modernes n'est pas univoque, et n'est pas non plus identique avec toutes les langues. Par exemple, les auteurs anciens parlent pour certains animaux d'un type de comportement particulier et difficile à exprimer en français qu'ils appellent φθόνος ou *invidia* : des animaux s'arrangeraient pour escamoter volontairement certaines substances naturelles qu'ils possèdent, pour éviter que l'homme, auquel elles seraient utiles, puisse en bénéficier (les cerfs, leurs bois tombés ; les lézards, leur peau morte ; etc.). Seul l'anglais dispose d'un mot capable plus ou moins de traduire le concept ancien (*begrudge*). Dans d'autres cas c'est le français ou l'espagnol qui présente l'équivalent le plus adéquat pour rendre les idées anciennes, de sorte que les concepts anciens sont approchés et restitués plus fidèlement par la *combinaison* des langues.

Conclusion

THEZOO, élaboré dans le cadre du projet *Zoomathia*, propose une ontologie de la zoologie prémoderne (antique et médiévale), qui vise à permettre l'intégration de données historiques hétérogènes –textuelles, iconographiques, archéologiques– relatives au domaine. L'annotation de ces données hétérogènes avec un référentiel commun permettra une recherche sémantique sur ces données, par des savants d'horizons différents (philologues, historiens, biologistes, archéologues…), et un trait

[31] Il va de soi que les labels devront être finement vérifiés par des natifs.

d'union entre les savoirs anciens et les connaissances contemporaines en zoologie. Il se présente aujourd'hui aussi comme un symptôme : l'effet et le signe d'une réflexion méthodologique et épistémologique sur l'évaluation anthropologique des savoirs anciens, impliquant des questions de sémantique historique, de transmission culturelle et de systèmes de représentation. Le processus d'ingénierie des connaissances mis en œuvre pour l'élaboration de THEZOO, où l'enrichissement du thésaurus va de pair avec l'annotation manuelle d'un corpus de textes représentatifs, offre l'opportunité d'une réflexion méthodologique appliquée, sur l'étude de la tradition zoologique ancienne et médiévales et la possibilité de décrire dans une longue durée, sans les uniformiser, un état de connaissance et des cadres de savoir.

Dans le même temps, un travail collaboratif est en cours qui vise à exploiter THEZOO pour annoter automatiquement l'ensemble des textes du corpus de *Zoomathia*. Les annotations manuelles servent de base de données d'apprentissage pour l'algorithme de classification automatique développé. D'autre part, l'évaluation des différentes versions de l'algorithme suppose de confronter les annotations produites automatiquement avec celles produites manuellement par des experts sur un sous-ensemble du corpus choisi. L'annotation se fait sur un groupe de traductions modernes, ce qui permet d'envisager de considérer des textes sources dans différentes langues. Pour l'instant seuls les textes latins et grecs sont concernés, mais pour le Moyen Age le corpus inclura progressivement la littérature en arabe et en hébreu impliquant un savoir sur les animaux. Ce travail systématique d'évaluation des annotations produites automatiquement sera également l'occasion de revenir sur les résultats de l'annotation manuelle en cas de discordances, et de réviser les choix adoptés dans l'annotation manuelle et éventuellement les connaissances représentées dans le thésaurus.

THEZOO est un thésaurus encore incomplet et en cours de construction. De nombreux aspects sont ouverts à la discussion et susceptibles d'amélioration, notamment la structuration et l'organisation hiérarchique des niveaux supérieurs, le développement de l'alignement avec d'autres référentiels de la taxonomie moderne (notamment avec TAXREF), et d'autres thésaurus de nature thématique ou lexicographique. A terme, notre objectif est de publier THEZOO sur le web de données afin qu'il puisse être réutilisé dans d'autres projets d'annotation sémantique de données ayant des objectifs voisins de ceux de *Zoomathia*.

Irene PAJON LEYRA
Université Nice Sophia Antipolis
CEPAM, UMR 6472
irene.pajon.leyra@gmail.com

Arnaud ZUCKER
Université Nice Sophia Antipolis
CEPAM, UMR 6472
zucker@unice.fr

Catherine FARON ZUCKER
Université Nice Sophia Antipolis
I3S, UMR 7271
faron@i3s.unice.fr

Abstract: This paper presents a thesaurus of ancient and medieval zoological knowledge, called THEZOO, constructed in the framework of the International Research Group (GDRI) *Zoomathia*. It aims at integrating heterogeneous data sources on zoology in Antiquity and Middle Ages: mainly texts, but also images, archaeological objects and archaeozoological material. The development process of THEZOO combines 1) the manual annotation of books VIII-XI of Pliny the Elder's *Natural History*, chosen as a reference dataset to elicit the concepts to be integrated in the thesaurus, and 2) the definition and hierarchical organization of the elicited concepts in the thesaurus. THEZOO is formalized in SKOS, the W3C standard to represent knowledge organization systems on the Web of data, and it is created with the Opentheso editor. Our final aim is to publish the thesaurus THEZOO as well as the corpus of annotated textual, iconographical and archeological resources, to support a semantic search in the corpus in different languages.

Résumé : Cet article présente un thesaurus pour des connaissances zoologiques antiques et médiévales, THEZOO, élaboré dans le cadre du Groupe De Recherche International Zoomathia. THEZOO vise à intégrer des données hétérogènes sur la zoologie antique et médiévale : principalement des textes, mais également des ressources iconographiques et archéologiques et du matériel archéozoologique. Son processus de développement consiste en la réalisation de deux tâches conduites parallèlement : (1) l'annotation de la section zoologique (livres VIII-IX) de l'Histoire Naturelle de Pline, choisie comme corpus témoin pour éliciter les concepts devant être intégrés au thésaurus ; et (2) la définition et l'organisation hiérarchique de ces concepts dans le thésaurus. THEZOO est formalisé en SKOS, le standard du W3C pour la représentation des systèmes d'organisation du savoir sur le Web de données, et il est construit à l'aide du gestionnaire de thésaurus Opentheso. Notre objectif à terme est de publier le thésaurus THEZOO ainsi que le corpus des sources textuelles, iconographiques et archéologiques annotées, pour permettre une recherche sémantique multilingue dans ce corpus zoologique global.

Keywords : *thesaurus, zoological knowledge, semantic annotation, SKOS, linked data, ancient and medieval sources, multilingualism*