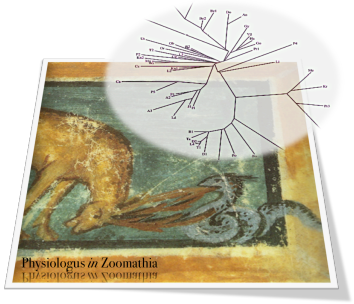
Contrat postdoctoral/Postdoctoral position

Location:	Nice. CEPAM (UMR 7264)
Salary:	2.200€ (net)
Hours:	Full Time
Contract Type:	Contract /Temporary
Placed on:	20th June 2015
Closes:	15th September 2015



INTITULE du PROJET

Application des méthodes de détection de plagiat dans l'analyse de la transmission manuscrite du *Physiologus* latin et de ses avatars (IX-XV^{e})

Description du projet

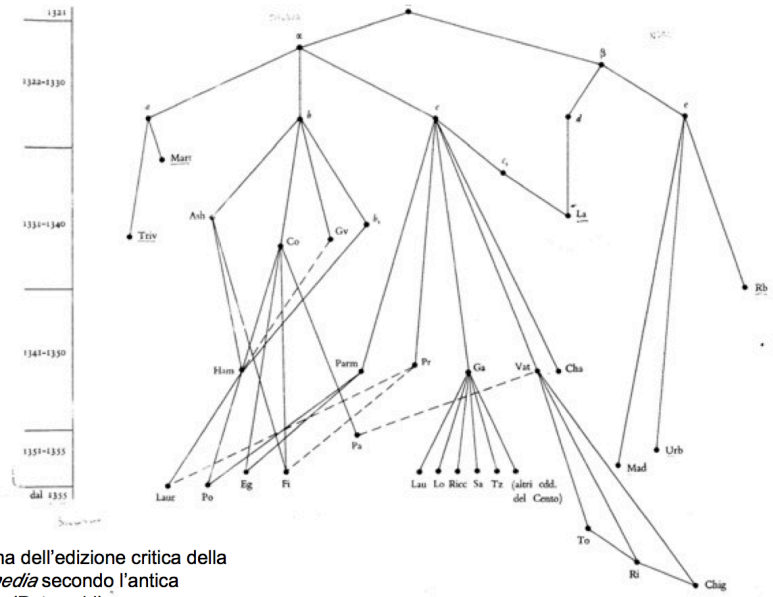
Contexte général du projet : *Digital Humanities,* histoire de la zoologie

GDRI Zoomathia

Le CEPAM est porteur d'un Groupement de Recherche International (ZOOMATHIA), consacré à l'étude de la transmission des savoirs dans l'Antiquité et au Moyen Âge [1]. Il fédère des laboratoires de trois instituts du CNRS (INEE, INSHS, INS2I) et réunit des philologues, des archéologues, des historiens, des iconologues, et des spécialistes en zoologie ancienne, en ingénierie des connaissances et en web sémantique. Un des enjeux de ce réseau est l'étude de la transmission du savoir zoologique, et le projet de recherche post-doctorale présenté ici s'inscrit dans ce cadre et vise la transmission latine d'un texte-clé composé au IV^{e} siècle et diffusé, reproduit et modifié jusqu'au XV^{e} siècle [2].

Présentation et problématiques : Détection des plagiats vertueux

Le projet de recherche post-doctorale présenté pour 12 mois consiste dans l'application de l'ingénierie et des méthodes de détection de plagiat à l'analyse de la transmission manuscrite du *Physiologus* latin et de ses avatars (IX-XV^{e}). **Ce texte**, qui fut sans doute après la Bible l'ouvrage le plus populaire du Moyen-Âge, a été l'objet d'un nombre considérable de manuscrits, que les philologues ont répartis dans des "arbres généalogiques", qui tiennent compte de l'époque de production du manuscript (siècle) et des ressemblances, exprimées en termes de filiation ou parenté plus ou moins lointaine. Le stemma de la tradition du ***Physiologus*** n'a pas été unifié, ni représenté, tant il est complexe, mais voici un exemple illustrant (pour la *Divine Comédie* de Dante) la visualisation stemmatique traditionnelle [3] :



Stemma dell'edizione critica della *Commedia* secondo l'antica vulgata (Petrocchi)

Il s'agit de proposer une approche nouvelle dans la détermination de la généalogie des manuscrits et la génétique des copies. Elle constituera un contre-point aux tableaux généalogiques traditionnels des manuscrits (*stemma codicum*), permettant de reconsidérer la méthode sur laquelle ceux-ci reposent. À partir des logiciels de comparaison de fichiers, et de détection de plagiat, il est désormais possible de mesurer quantitativement des écarts. Cette application à la **tradition textuelle** est totalement originale par rapport aux pratiques en relation avec l'extraction d'information. La problématique de la copie relève dans les deux cas d'une approche radicalement différente. Le *plagiat*, stigmatisé dès l'antiquité [4], envisage négativement (comme un vol) la reproduction de données textuelles dans la lettre ou dans l'esprit [5]. La situation est toute différente dans l'étude de la tradition manuscrite, où l'idéal de transmission est la fidélité rigoureuse des manuscrits étudiés à l'original premier (*autographe* de l'auteur) — presque toujours disparu naturellement pour les auteurs antérieurs au VIII^{e} siècle. Les chercheurs tentent en effet d'identifier des familles de manuscrits sur la base de ressemblances, à la recherche d'un ancêtre commun (appelé *archétype*). Jusqu'à présent l'évaluation des écarts et similitudes s'est faite sur la seule base de la confrontation personnelle (et sans cesse à renouveler) d'un chercheur individuel, par la collation de manuscrits, c'est-à-dire la confrontation progressive (analytique) des passages équivalents, et l'évaluation globale (synthétique) du degré de proximité et de la position du manuscrit dans l'arbre général. Mais cette évaluation est en partie intuitive et une appréhension synoptique complète des écarts est irréalisable par un simple lecteur. L'ecdotique (science de l'édition) moderne a élaboré un certain nombre de principes et de critères dans la constitution des "familles" de manuscrits (valorisation des "fautes" conjonctives caractérisant une lignée, ou inversement des fautes séparatives, etc.), mais ces règles traditionnelles reposent sur des conceptions "généalogiques" qui ne sont pas totalement dépourvues d'idéologie (le registre filial de modèle [père] à copie [fils]), et peut-être de schématisme. L'approche nouvelle et a priori "neutre" des dissimilitudes qu'offrirait l'exploitation de logiciels détecteurs de plagiat permettrait peut-être de dépasser les cadres conceptuels "hérités" de la méthode stemmatique.

Le corpus de ce projet est constitué par les manuscrits latins du *Physiologus*, nom d'un recueil de chapitres consacrés chacun à un animal traité sous un angle naturaliste et spirituel (lion, pélican, hérisson, unicorne…) ; ce recueil est considéré comme l'ancêtre de tous les bestiaires [6, 7, 8]. La transmission de ce texte implique des centaines de

manuscrits [9] et son inventaire méthodique est très incomplet [10]. De fait, la situation est très difficile et complexe, et on ne s'étonne pas que les schémas proposés par les chercheurs pour le *stemma* soient à la fois variés et sujets à caution [11, 12]. L'apparition ou l'identification de manuscrits nouveaux augmente la complexité de la tradition et élargit la famille, rendant ardue une approche systématique du corpus [13]. Celui-ci a plusieurs particularités qui rendent l'approche proposée prometteuse, et qui tiennent à la nature de ce texte, qui est à la fois un *texte*, une *collection* ouverte (inclusion, prolongements), et un *genre* (avatars et variations). En effet, s'il est naturel qu'aucune copie manuscrite ne soit exactement identique à une autre, la situation se complique ici du fait que le *Physiologus* est (1) un recueil variable en nombre de chapitres, en organisation, en contenu (présence de paraphrases ou d'extension, ellipses), et (2) qu'il évolue progressivement vers la forme du Bestiaire. La typologie traditionnelle proposée pour ce réseau de textes est grossière et instable. Carmody [14, 15] identifia 4 principales *versions* ou "branches" du Physiologus (A, B, C, Y), mais ajouta ensuite une branche W comprenant 4 versions secondaires du *Physiologus* latin : W(A) (B) (C) (D). Ces quatre versions, transmises par une centaine de manuscrits connus, forment, avec deux groupes de bestiaires, la première "famille". Une typologie secondaire distingue, à partir de B deux sous-familles, et cinq subdivisions inférieures de Bestiaires dérivés représentant de nouveau une centaine de manuscrits selon Clark et MacMumm [16], sans compter plus de 100 manuscrits d'une version poétique du *Physiologus* connu comme *Physiologus* de Théobald.

Les dimensions relativement faibles du texte (*ca* 10.000 mots), et le fait qu'aujourd'hui un grand nombre de ces manuscrits soit disponible sous une forme digitalisée (par les grandes bibliothèques, comme Florence, Heidelberg Münich, Oxford, Paris…), voire numérisée [17] rendent possible une comparaison automatique d'au moins une bonne partie d'entre eux. Ce ne sont pas les manuscrits eux-mêmes qui seront traités, mais le *texte* de ces manuscrits. Quand ce texte n'est pas déjà disponible aux chercheurs sous forme océrisée, le post-doc aura en charge de transcrire le manuscrit, en s'appuyant sur les versions déjà répertoriées, puisque les différences manuscrites sont souvent localisées.

La méthode employée repose sur les ressources des logiciels de détection de plagiat dont le marché s'étend et les performances s'améliorent, mesurées par des tests comparatifs réguliers [18]. Le développement des outils existants (Ephorus, Turnitin, Viper…) pour la détection du plagiat (depuis Compilatio.net en 2003) ne vise pas seulement à étendre le volume de documents traités et la diversité de format, mais à pondérer les types de différences ou de ressemblance par une évaluation qualitative et représentative de leur importance. La comparaison ne concerne pas seulement les identités mais intègre des variations (modification lexicale et synonymie, déplacement de mots, similitudes structurelles, etc.). Ce que les logiciels scrutent pour déterminer une copie délictueuse ou un maquillage intentionnel peut servir à révéler des accidents révélateurs et des copies vertueuses. Des logiciels pilotes d'**alignement** de transcriptions linéarisées de textes, tel MEDITE, développé par l'OBVIL [19], pourront être sollicités de manière complémentaire. Une **visualisation** des stemmas par dendrogrammes produits à l'aide de logiciels développés pour la phylogénétique en biologie a déjà été proposée dans le cadre de l'édition de textes et permettra, en outre, de trouver un mode de formalisation plus adéquat [20].

Objectifs : Au-delà du *Lorem Ipsum*

L'objectif principal de ce travail est l'expérimentation d'une méthodologie nouvelle, s'appuyant sur des outils de traitement automatique du langage (TAL) et de détection de similitudes, pour analyser une tradition manuscrite et textuelle d'une telle richesse telle (des centaines de documents plus ou moins "similaires") qu'elle est impossible à traiter

efficacement par les moyens traditionnels et artisanaux. Elle s'inspire de divers travaux récents appliquant les outils informatiques à l'approche stemmatique pour l'élaboration d'une nouvelle stemmatologie "digitale" [20, 21, 22, 23, 24, 25, 26]. Cet objectif est à la fois méthodologique et scientifique, car la procédure proposée sera susceptible d'être adoptée pour vérifier ou formuler des hypothèses sur d'autres œuvres ayant fait l'objet d'une transmission diversifiée et de mutations dans la tradition. La comparaison portera sur l'ensemble des similitudes et écarts (structure, littéralités, orthographe, ellipses, ajouts,…). Après avoir testé les outils disponibles et comparé leurs résultats aux conclusions des philologues, il s'agira d'envisager les possibilités de combiner ces outils avec des extensions ou des instruments nouveaux capables, grâce à l'ingénierie linguistique, de détecter des variations fines, des substitutions et des opérations de paraphrasage permettant de les évaluer qualitativement dans une perspective d'histoire de la transmission. Ainsi la plupart des logiciels traite tout texte en alphabet romain, quelle que soit la langue, mais des adaptations spécifiques pourraient permettre d'accroître la pertinence des résultats. La détermination des réseaux manuscrits ne repose pas exclusivement sur des écarts qui offrent un schéma synchronique : la datation matérielle *des manuscrits* (et non de la version textuelle) contribue aussi à l'élaborer, mais celle-ci est souvent assez vague (degré d'imprécision de plusieurs décennies). La présence fréquente d'iconographie dans le cas du *Physiologus* permettrait, ultérieurement, de croiser les résultats avec une expertise d'historien de l'art pour le schéma obtenu par la détection des écarts. Enfin, au-delà des Bestiaires recensés dans les familles Carmody-MacCulloch [11, 13, 14], existent de nombreux avatars partiels et produits dérivés de la littérature "physiologique" que cette méthode permettrait de traiter, éclairant sa difusion historique au Moyen-Âge. La méthode permettra d'identifier dans des textes apparemment déconnectés de la tradition du *Physiologus* des emprunts insoupçonnés et non repérés par les enquêtes traditionnelles.

Organisation du travail :

Le travail du post-doctorant sera encadré par A. Zucker, porteur du projet. Les tâches seront les suivantes *(avec estimation approximative du temps consacré à chacune)* :

- Familiarisation avec le texte, les familles et sous-familles et l'état de l'art sur la tradition manuscrite [1 mois]
- Océrisation du texte de quelques manuscrits digitalisés du *Physiologus*, en collaboration avec l'IE (CEPAM) spécialiste d'analyse des sources (pour compléter le corpus de départ) [2.5 mois]
- Prospection et identification des logiciels de détection de plagiat les plus adaptés ; et test de ces derniers sur le corpus de textes du réseau du *Physiologus* [1.5 mois]
- Confrontation des séries de résultats avec le *stemma codicum* des philologues et analyse des différences [2.5 mois]
- Comparaison des résultats avec ceux obtenus par des logiciels d'alignement de textes spécialisés dans la génétique textuelle [1 mois]
- Réflexion sur les algorithmes des logiciels, en collaboration avec des collègues d'I3S et de BCL, pour une adaptation de l'ingénierie utilisée au cas spécifique de la confrontation positive de variantes textuelles pour des textes latins [2.5 mois]
- Bilan sur les potentialités de la méthode, les aménagements techniques et les perspectives d'utilisation pour d'autres corpus textuels [1 mois]

Sélection du candidat

Le candidat (francophone ou anglophone) sera recruté sur dossier (CV détaillé et lettre de motivation) et après entretien, par un jury de cinq personnes constitué

du porteur de projet (A. Zucker), du directeur d'unité (M. Regert), du responsable de l'équipe MTI du CEPAM (M. Lauwers), d'un membre de l'équipe *Logométrie et corpus politiques, médiatiques et littéraires* du laboratoire BCL, et d'un membre de l'équipe WIMMICS (*Web-Instrumented Man-Machine Interactions, Communities, and Semantics*) du laboratoire I3S.

Références

[1] http://www.cepam.cnrs.fr/zoomathia/
[2] Zucker, Arnaud (2004). *Physiologos. Le Bestiaire Des Bestiaires*. Grenoble.
[3] Petrocchi, Giorgio (1966-1967). *La Commedia secondo l'antica vulgata*, Milano.
[4] Ziegler, Konrat (1950). *Plagiat*, RE XX.2: 1956-1997.
[5] Selle, Hendrik (2008). "Open Content? Ancient Thinking on Copyright." *Revue Internationale Des Droits de L'antiquité* 55: 469–84.
[6] Carmody, Francis J. (1953). *Physiologus, the very ancient book of beasts, plants and stones, translated from Greek and othe languages*, San Francisco.
[7] Sbordone, Francesco (1936). *Physiologus,* Milano.
[8] Henkel, Nikolaus (1976). *Studien Zum Physiologus Im Mittelalter*, Tübingen.
[9] Sbordone, Francesco (1936). *Ricerche sulle fonti e sulla composizione del Physiologus greco*, Napoli.
[10] http://bestiary.ca/prisources/psmanu869.htm.
[11] James, Montague Rhodes (1928). *The Bestiary: Being a Reproduction in Full of the Manuscript Ii. 4. 26 in the University Library, Cambridge, with Supplementary Plates from Other Manuscripts of English Origin, and a Preliminary Study of the Latin Bestiary as Current in England. Edited ... by M.R. James.* Oxford.
[12] MacCulloch, Florence (1962). *Medieval Latin and French Bestiaries*, Chapel Hill.
[13] Muratova, Xénia (1994). *Aspects de La Transmission Textuelle et Picturale Des Manuscrits Des Bestiaires Anglais À La Fin Du XIIe et Au Début Du XIIIe Siècle*. Comprendre et Maîtriser La Nature Au Moyen Age, 579-605.
[14] Carmody, Francis J. (1939). *Physiologus latinus: éditions préliminaires, versio B*. Paris.
[15] Carmody, Francis J. (1939). *Physiologus Latinus Versio Y*, Berkeley & Los Angeles.
[16] Clark, Willene B., and McMunn Meradith T. (1989). *Beasts and Birds of the Middle Ages: The Bestiary and Its Legacy*. Philadelphia.
[17] http://www.brepols.net/Pages/BrowseBySeries.aspx?TreeSeries=LLT-O
[18] http://plagiat.htw-berlin.de/software/2013-2/
[19] http://obvil.paris-sorbonne.fr/developpements/medite
[20] Roelli, Philip and Bachmann, Dieter (2010). *Towards generating a stemma of complicated manuscript traditions: Petrus Alfonsi's Dialogus.* Revue d'histoire des textes, 5:307-321 (http://www.zora.uzh.ch/34542/).
[21] J.-B. Guillaumin : Approche informatique de l'analyse stemmatique. http://www.normalesup.org/~jguillau/calculs_philologiques.html
[22] Andrews, Tara L., and Caroline Macé (2013). "Beyond the Tree of Texts: Building an Empirical Model of Scribal Variation through Graph Analysis of Texts and Stemmata." *Literary and Linguistic Computing* 28 (4): 504–21.
[23] Andrews, Tara, and Caroline Macé (2014). *Analysis of Ancient and Medieval Texts and Manuscripts. Digital Approaches.* Turnhout: Brepols.
[24] Le Pouliquen, Marc (2010). "Filiation de Manuscrits Sanskrits et Arbres Phylogénétiques." *Mathématiques et Sciences Humaines. Mathematics and Social Sciences*, no. 192: 57–91.
[25] Le Pouliquen, Marc and Jean-Pierre Barthelemy (2009). "Construction D'arbres À Partir de Relations D'intermédiarité, Application Au Stemma Codicum." *Mathématiques et Sciences Humaines. Mathematics and Social Sciences*, no. 187: 93–105.
[26] Raynaud, Dominique (2014). "Building the Stemma Codicum from Geometric Diagrams." *Archive for History of Exact Sciences* 68 (2): 207–39.