Extraction et modélisation de connaissance à partir de texte – Applications à la biologie

Claire Nédellec

MÉMOIRE DÉPOSÉ EN VUE DE L'HABILITATION À DIRIGER DES RECHERCHES

Claire Nédellec

Extraction et modélisation de connaissance à partir de texte

– Applications à la biologie

Université Blaise Pascal, département Informatique – Clermont Ferrand, 2013

Table des Matières

Introduction

Ce mémoire décrit mon parcours scientifique depuis 1990, année de l'obtention de mon diplôme de DEA à l'Université Paris VI. Il couvre deux périodes d'activité, d'étudiante en thèse et de maître de conférence à l'Université d'Orsay au LRI (*Laboratoire de Recherche en Informatique*) jusqu'à la fin 2001, puis de chercheuse à l'INRA (*Institut National de Recherche Agronomique*) dans l'unité MIG (*Mathématiques, Informatique et Génome*).

Mon domaine de recherche est l'intelligence artificielle, principalement l'apprentissage automatique[1] et le traitement automatique de la langue naturelle (TALN) et leurs applications à l'acquisition de connaissances pour les systèmes à base de connaissance (SBC). Les SBC sont des logiciels capables de raisonnement automatique grâce à l'utilisation et l'apprentissage de connaissance. Ils ont en commun l'exploitation de connaissances symboliques propres à un domaine et représentées dans un langage formel, manipulable par un ordinateur. Ils sont utilisés dans une très large palette d'applications : citons la capitalisation des connaissances, la planification, la traduction, l'aide à la décision.

Depuis les années 80, la communauté de recherche en acquisition de ces connaissances, fameux "goulet d'étranglement" des SBC, s'est structurée autour de deux grandes questions, celle de l'apprentissage automatique de connaissances à partir d'exemples et celle de l'acquisition manuelle de connaissances expertes et de leur modélisation. Mon activité de recherche relève de ces deux domaines dans un objectif de conception de systèmes *assistants*, c'est-à-dire automatiques lors que cela est possible et interagissant avec l'utilisateur si c'est nécessaire ou utile. L'intérêt pour ce sujet au confluent de l'acquisition des connaissances et de l'apprentissage automatique est particulièrement vif en France et en Allemagne où il structure une large communauté scientifique. Il a émergé au début des années 90 dans une perspective de sciences cognitives. Il a évolué en se renforçant au cours du temps en deux courants très dynamiques, la modélisation formelle de connaissance en particulier pour le Web sémantique et le développement de méthodes d'acquisition et d'apprentissage de connaissances, en particulier en sciences de la vie et médecine.

J'ai sélectionné pour ce mémoire des résultats structurants de mon activité de chercheuse et qui sont illustratifs de l'évolution de la recherche en acquisition de connaissance. Ils portent sur la formalisation de l'apprentissage automatique dans un cadre unificateur, en logique et en sciences cognitive et sur l'exploitation de documents en langue naturelle comme source de connaissances.

[1] Dans la suite, le terme apprentissage désigne apprentissage automatique de connaissances.

Apprentissage automatique symbolique pour l'acquisition de connaissances

1. Apprentissage dans un treillis de concepts

Mon stage de DEA, puis ma thèse au Laboratoire de Recherche en Informatique (LRI), ont débuté en plein essor de l'apprentissage automatique en France en 1990 à l'Université Paris-Sud, dans l'équipe d'Yves Kodratoff, un des fondateurs du domaine. L'équipe *Inférence & Apprentissage*, forte d'une quinzaine de chercheurs et étudiants, travaillait principalement en apprentissage symbolique et en particulier sur l'apprentissage de connaissances relationnelles discriminantes en présence de théorie du domaine (*background knowledge*) à partir d'exemples positifs et négatifs. Mon mémoire de DEA [7] et ensuite de thèse [6] ont porté sur l'acquisition de règles de résolution de problème, tels que des opérateurs de planification pour un robot, dans le cadre de l'analyse de concepts formels (*Formal Concept Analysis, FCA*) [Wille, 1982]. J'ai proposé une méthode de généralisation et de révision de théorie du domaine qui lève l'hypothèse contraignante, mais fréquente, de cohérence des exemples. Cette hypothèse suppose que tous les exemples nécessaires à l'apprentissage sont suffisants et corrects et qu'il existe une solution dans l'espace des hypothèses. Ma méthode de généralisation dite "à petit pas", extension de celle proposée par Claude Sammut et Ranan B. Banerji en 1986, parcourt de proche en proche le treillis des hypothèses en produisant automatiquement les exemples manquants et les soumettant à un oracle [64], dans une préfiguration de l'apprentissage actif (*active learning*) [r129, 18]. En cas de surgénéralisation, la révision la plus appropriée est choisie grâce à des étapes d'inférence et d'interaction avec l'utilisateur qui est sollicité à différents niveaux de connaissances. Les résultats en révision de théorie ont donné lieu à publication [38] dont ECAI [17] et EKAW [16]. Avec Ronen Feldman de l'Université de Tel Aviv, nous avons formalisé les opérateurs de révision dans un cadre de logique d'ordre un [15], grâce à un projet bilatéral [r169], à la suite de ses travaux à Cornell University sur la révision de théorie (*revision for explanation-based learning (EBL)*) [Feldman, 1993].

J'ai implémenté la méthode dans le système APT initialement conçu par George Tecucci (Carnegie Mellon University) et Yves Kodratoff [Tecucci et Kodratoff, 1990]. J'ai intégré APT comme algorithme d'apprentissage du *Machine Learning ToolBox* (MLT, boîte à outils) [Kodratoff et al., 1994] du projet européen éponyme [r175, r176], dont j'ai animé le principal lot, *Algorithms*[r177]. Le projet MLT a fédéré une grande partie de la communauté en apprentissage automatique en Europe, de 1989 à 1994. J'ai ensuite complété le volet fondamental de ces travaux par des expérimentations destinées à évaluer la pertinence de la méthode dans un cadre applicatif et à valider les scénarios d'interaction homme-machine. La méthode a été validée par des applications dans les domaines médical [r173] et bancaire [s84, r174] grâce au projet bilatéral franco-portugais *Disciple* avec l'équipe d'Intelligence Artificielle d'Ernesto Costa à l'université de Coimbra (Portugal). Les résultats sont publiés notamment dans le journal *Applied Artificial Intelligence* [4]. L'expérimentation de la méthode APT, s'est poursuivie avec le projet européen ESF *Learning in Human and Machine* (LHM), sur la modélisation cognitive pour l'enseignement de la physique au niveau élémentaire en collaboration avec Andrée Tiberghien (ENS Lyon), spécialiste de didactique de la physique [s85, 73, 75, 182]. Ces trois expériences très diffférentes ont montré la pertinence de l'apprentissage coopératif, mais aussi la limitation du cadre algébrique en terme d'expressivité de la représentation.

2. Apprentissage coopératif et modélisation

Selon le paradigme dominant, les connaissances déclaratives au système d'apprentissage, exemples et connaissances du domaine, étaient connues *a priori* et le processus était ensuite automatique. Cette limitation, intéressante d'un point de vue théorique, est réductrice dans un cadre applicatif quand une intervention experte au cours de l'apprentissage produit un résultat de meilleure qualité pour une charge acceptable. Pour que la participation de l'utilisateur soit efficace, la démarche de résolution de problème de la machine doit être explicable à l'utilisateur [14, 61, 62, s107]. Cela suppose des capacités d'explication de la part de la machine et la modélisation de l'activité de l'expert, des

opérateurs d'apprentissage et des connaissances [36, 44, 45, s89, s109]. Ma recherche s'est inscrite conjointement dans deux paradigmes de modélisation naissants au début des années 90 :

(1) Modélisation dans un cadre logique de l'apprentissage automatique comme une inférence inductive, avec la naissance de la programmation logique inductive (PLI),

(2) Modélisation de l'acquisition de connaissances et des SBC par exemple, par les méthodologies KADS (*Knowledge Acquisition and Design Support*) et KOD (*Knowledge Oriented Design*), développées par la communauté d'acquisition des connaissances.

2.1 Modélisation de l'apprentissage dans un cadre logique

Ma recherche en modélisation logique a été réalisée en collaboration avec Céline Rouveirol, fondatrice de la PLI en France, qui venait alors d'obtenir des résultats novateurs sur l'utilisation de la théorie du domaine [Rouveirol et Puget, 1990]. Au début des années 90, le domaine de l'apprentissage automatique souffrait de l'absence de description formelle pour la floraison de nouveaux algorithmes d'apprentissage. Leurs langages et leurs propriétés (espace de recherche, complexité, algorithme, *etc*.) n'étaient pas comparables. Notre ambition en PLI était de concevoir un modèle dans un cadre logique qui permette non seulement de décrire, mais de définir et d'expérimenter de nouveaux algorithmes de façon similaire à David Chapman en planification (87). A la suite de la modélisation de APT pris comme exemple d'algorithme à modéliser [46, 47], nous avons défini *Haïku* [r127], un modèle unificateur et opérationnel, qui modélise les méthodes de classification symboliques à partir de la combinaison d'opérateurs unitaires de PLI et de biais déclaratifs [37, 60, 63, r128]. Cette recherche s'inscrit à la suite des travaux de Tom Mitchell en 1982 sur les espaces de recherche [Mitchell, 1982], de Wray Buntine en 1988 sur la subsomption généralisée et de Luc de Raedt (Université de Louvain) et Steve Muggleton (Université de York) (94), avec qui nous collaborions dans la tâche *Declarative Biases* du projet européen ILP (Inductive Logic Programming) [r170, 74]. L'implémentation de la méthode par Céline Rouveirol a permis de modéliser et de comparer expérimentalement différentes méthodes d'apprentissage, de définir de nouvelles méthodes comme une combinaison d'opérateurs élémentaires et de biais déclaratifs et de réfléchir à un niveau théorique sur les interactions entre les différents opérateurs du modèle. Ces travaux furent ensuite poursuivis sur le plan théorique par Céline Rouveirol et Fabien Torre, alors en thèse [Torre et Rouveirol, 1997].

2.2 Modélisation de l'apprentissage en acquisition de connaissances

Parallèlement, mes travaux en apprentissage coopératif se sont poursuivis dans une perspective d'acquisition des connaissances pour l'enrichissement d'ontologie. La proximité des notions de *théorie du domaine* en la PLI et d'*ontologie* en acquisition de connaissance permet une application de certains résultats de l'un à l'autre. Comme en PLI, ma contribution a porté sur la modélisation du processus d'acquisition [183]. J'ai défini une typologie des connaissances mobilisées au cours de l'apprentissage, connaissances qui sont connues *a priori* ou acquises auprès de l'utilisateur [61, 62]. J'ai approfondi cette typologie avec Paul-André Tourtier (INRIA Sophia) lors de mes séjours à l'Université de Berkeley dans l'équipe *Machine Learning* de Stuart Russell, et dans l'équipe de Barbara Hayes-Roth à KSL (Knowledge System Laboratory, Université de Stanford) en 1992 et 1993 publiée dans *Knowledge Engineering Review* [3]. J'ai contribué à la création et à l'animation du thème de l'apprentissage coopératif en organisant successivement avec Yves Kodratoff deux ateliers internationaux associés aux conférences ECAI en 1994 et IJCAI en 1995 sur la thématique apprentissage et acquisition de connaissance [259, 260].

Ces résultats fondamentaux ont été approfondis et validés par la thèse CIFRE de Luc Poittevin avec la SNCF (Société nationale des chemins de fer français), que j'ai encadrée de 1994 à 1998. La SNCF appliquait de nombreuses règles complexes – plus de 600 – pour accorder des réductions du prix des billets à ses employés et à leurs ayant-droits. Elles étaient décrites dans un manuel papier et étaient partiellement incohérentes. Nous avons défini une typologie claire des différentes connaissances en jeu, formalisé l'ensemble de ces règles dans un cadre logique et implémenté un SBC. Ce type de formalisation a connu depuis un grand développement sous le nom d'application de *Business Rules.* Nous avons conçu un module de raisonnement et de révision des règles basée sur la méthode de APT, c'est-à-dire détectant les incohérences, proposant des révisions et en montrant les conséquences sous forme de simulation de cas. Cette application a confirmé l'importance de la compréhension par l'expert des conséquences de ses corrections et du rôle de ses connaissance expertes dans le

raisonnement et l'intérêt d'une communication homme-machine sous forme d'exemples [267]. Le système *Edinos / Revinos*, développé en SmallTalk par Luc Poittevin, a été publié à EKAW en 1997 et KAW en 1998. L'application a été déployé et exploitée par la SNCF sous la forme d'un service Minitel consultable par tous les employés ayant-droits et révisable par un administrateur.

Extraction d'information et modélisation de connaissances

1. Introduction

Le manque d'exemples pertinents représentés à un niveau de description approprié était un des principaux verrous de la construction de modèle de connaissance par des approches symboliques comme la PLI ou l'IC. Les exemples, ou observations, doivent être décrits à un niveau d'abstraction plus élevé qu'il n'est nécessaire à approches statistiques plus robustes, capables de traiter de gros volumes de données bruitées comme les données brutes de capteurs. Par exemple, un pré-traitement manuel ou par des algorithmes de reconnaissance de forme est nécessaire aux méthodes symboliques pour exploiter les images ou la parole. Pour ces approches, la littérature scientifique et technique est une source riche d'applications potentielles de modélisation de connaissance. Elle constitue un gisement de connaissances de grande valeur, mais qui reste largement inexploité par ses utilisateurs parce qu'uniquement sous forme textuelle. La croissance très rapide du volume de publications à un niveau mondial rend impossible une veille systématique. Il est nécessaire de doter les utilisateurs d'outils semi-automatiques pour sélectionner, extraire et formaliser ces connaissances, qui seront ensuite confrontées avec des connaissances de sources et domaines étendus.

Au début des années 90, les méthodes automatiques de construction de bases de connaissance à partir de texte en langage naturel par apprentissage automatique commençaient à fleurir, mais les approches d'extraction d'information par des règles définies manuellement étaient dominantes. Elles s'appuyaient sur des indices superficiels du texte, essentiellement l'ordre des mots. L'apprentissage statistique [Manning et Schutze, 1999] a révolutionné le TALN à travers tout d'abord la segmentation, la racinisation (*stemming*), l'analyse morpho-syntaxique, l'apprentissage de grammaire, et notablement les applications de reconnaissance de la parole et de traduction. Le champ sémantique a été concerné plus tardivement et a d 'abord porté sur la classification de documents et la désambiguisation du sens.

L'extraction d'information (EI) et son évaluation se sont considérablement développés sous l'impulsion de la DARPA avec les conférences MUC (*Message Understanding Conference*) [Grishman et Sundheim, 1996]. La disponibilité d'outils de traitement automatique de la langue naturelle (TALN) et plus particulièrement, la disponibilité et la qualité grandissante d'analyseurs syntaxiques ont ouvert la porte à une nouvelle génération de méthodes d'EI basées sur une analyse profonde du texte dont les progrès ont été immédiats. Selon les théories linguistiques qui articulent syntaxe et sémantique du langage, les relations syntaxiques reflètent en effet les rôles sémantiques des entités et les relations entre ces entités. Par exemple en biologie dans la phrase, *GerE activates cotD*, la protéine *GerE* sujet syntaxique du verbe *activate* dont l'objet syntaxique est le gène *cotD*, est probablement l'agent d'une action d'activation dont le gène *cotD* est le patient, ou la cible.

En réduisant l'investissement dans le développement de ressources linguistiques, la publication de lexiques de qualité dont notablement le thésaurus WordNet [Fellbaum, 2005] a aussi été un facteur important de l'élargissement de la communauté d'EI aux spécialistes de l'apprentissage automatique. Selon la théorie de Zeilig Harris (1951), le contenu et la structure des langues de spécialités forment un sous-language dont les relations entre les éléments peuvent être analysés à partir de leur distribution statistique. C'est dans ce paradigme que s'est fédérée une communauté internationale de recherche pluridisciplinaire développant des méthodes automatiques d'acquisition de connaissances à partir de textes spécialisés pour la modélisation de connaissance et la construction d'ontologie. Elle est particulièrement dynamique dans les domaines biomédical et biologique. En Europe, les communautés françaises et allemandes se distinguaient à l'époque par leur dynamisme au confluent du TALN, de l'IC (l'ingénierie des Connaissances) et de ce que l'on commençait à appeler le Web Sémantique, qui a apporté à l'IC une dimension formelle.

J'ai contribué à porter ce sujet dans le domaine de la biologie moléculaire depuis la fin des années 90 au LRI avec Philippe Bessières chercheur bioinformaticien de l'unité MIG (*Mathématiques, Informatique et Génome*) à l'INRA, puis avec l'équipe Bibliome que je dirige dans cette unité depuis 2002 [s87]. Au contraire du domaine médical, le domaine de la biologie était encore assez peu traité, sinon à travers la reconnaissance d'entités nommées, gènes et protéines [Fukuda, 1998] ou la mesure de la cooccurrence de leurs noms dans les textes comme dénotant une relation [Blaschke et al. 1999]. Avec Philippe Bessières, nous avons identifié la question de la modélisation de réseaux de régulations géniques comme un enjeu majeur en biologie [Babu, 2013] et dont la complexité nécessite le recours à une analyse profonde du texte. L'exemple précédant, *GerE activates cotD* est un exemple d'interaction génique entre *GerE* et *cotD*, dont les réseaux de régulation sont composé. L'abondance d'articles et de données sur *Bacillus subtilis* dont Philippe Bessières est expert reconnu a justifié notre choix de cette bactérie modèle comme modèle d'étude. La question de la modélisation de réseaux de régulations géniques diffère notablement des tâches MUC pour lesquelles des patrons lexicaux suffisaient. Sur un tel sujet, une analyse superficielle à base de patrons produisait des résultats médiocres en terme de rappel [Ono et al., 2001]. Par contre les premiers résultats d'apprentissage appliqué à l'analyse syntaxique pour l'interaction protéine-protéine étaient très encourageants [Craven et Kumlien, 1999] Notre hypothèse était que l'apprentissage de ressources linguistiques spécialisées telles que les structures prédicatives apporteraient un gain considérable en permettant d'exprimer les règles d'extraction avec des relations sémantiques entre entités à un niveau de généralisation élevé. Ce projet est encore à l'ordre du jour. Nos applications en biologie moléculaire se sont ensuite diversifiées pour répondre aux besoins croissants de structuration de l'information textuelle en biologie moléculaire. La partie *Activités Transversales* de ce mémoire en décrit les principales. Cette partie *Extraction d'information et modélisation de connaissances* est consacrée aux aspects méthodologiques.

La présentation des activités de recherche est sous-tendue par un projet unificateur de développement de méthodes d'analyse de contenu de texte qui s'intègre dans une démarche plus générale de modélisation de connaissance en sciences de la vie par des ontologies. Nos méthodes visent à formaliser et intégrer dans un modèle logique ces connaissances éparses qu'il faut interpréter et extraire du contenu scientifique de documents en langue naturelle. Dans les domaines expérimentaux qui m'intéressent, le texte scientifique relie des observations à leur interprétation dans le cadre des hypothèses de recherche et il explicite son insertion dans le paradigme qui est supposé unique et non discutable. Ce sont ces différents niveaux de connaissance, observation, interprétation, résultat, paradigme, que mes méthodes visent à analyser et intégrer dans un modèle unificateur, spécifique à chaque problématique biologique. Elles analysent des niveaux linguistiques profonds, sémantiques et pragmatiques à l'aide de *termino-ontologies* qui guident l'interprétation du texte et qu'elles contribuent à enrichir. Dans ce sens, même si certaines des méthodes mises en jeu sont communes avec la fouille de texte, ou l'analyse textuelle, l'objectif que je poursuis dans leur mise en œuvre et leur intégration diffère.

Ces méthodes s'appliquent principalement à des corpus de documents scientifiques homogènes dont le paradigme est explicite. Dans l'objectif de l'extraction d'une information potentiellement rare, chaque document est traité comme porteur d'un contenu cognitif original et complémentaire. Ces hypothèses répondent parfaitement aux besoins des champs thématiques et scientifiques de la biologie. Chaque question applicative est spécifique, mais les textes partagent un grand nombre de traits linguistiques et conceptuels qui réduit le coût d'adaptation des méthodes. Les traits communs sont principalement les sources documentaires (articles, rapports ou brevets), le genre (scientifique ou technique, l'abondance de données expérimentales, la langue) et le vocabulaire. Une fois les méthodes d'extraction de connaissance adaptées à une application donnée, le coût du développement d'une application similaire est celui de l'annotation par des spécialistes d'exemples de la connaissance à extraire du texte et de la validation du modèle de connaissance. L'automatisation de son acquisition par des méthodes d'apprentissage automatique est au centre de mon activité de recherche.

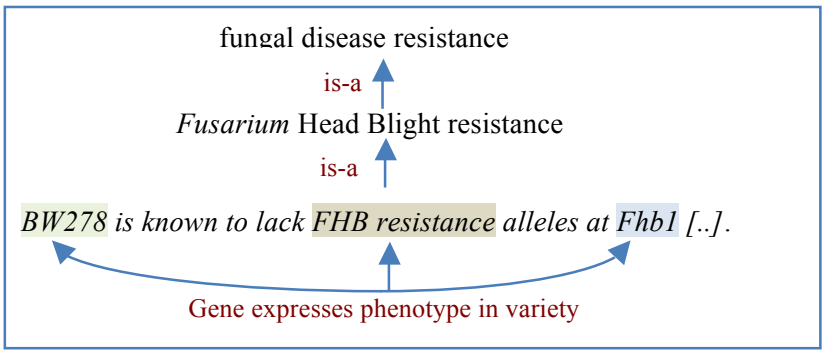
2. Analyse sémantique

Concrètement, les fragments de texte qui représentent des unités sémantiques sont finement étiquetés par les concepts et relations définis dans une ontologie, sous la forme d'annotations sémantiques formelles. Ces métadonnées relient des segments de texte brut de tailles variables, du mot au

document, à des concepts formels structurés de telle sorte que des opérations logiques de recherche, d'inférence, de validation et d'intégration avec d'autres sources de données structurées deviennent possibles [Staab et Studer, 2004 ; W3C OWL Working Group, 2004, 2012]. Prenons un exemple dans l'application sur la sélection du blé par marqueurs génétiques décrite ci-dessous :

BW278 is known to lack FHB resistance alleles at Fhb1 [..].

Notre objectif est d'identifier automatiquement les entités, ici *BW278*, *FHB resistance* et *Fhb1*, respectivement de types variété, phénotype et gène. Plus précisément, nous voulons les associer à des concepts préalablement définis dans l'ontologie et organisés en hiérarchies. Par exemple de la figure, l'entité *FHB resistance* doit être rattachée au concept *Fusarium Head Blight resistance*, feuille de la hiérarchie, mais mieux encore à ses ancêtres *fungal disease resistance*, *disease resistance*, *biotic stress response*, *stress resistance*, *response to environmental factor* par ordre croissant de généralité. La prédiction de la relation ternaire *gene_expresses_phenotype_in_variety* doit relier les trois entités du texte, gène, phénotype et variété. La relation est préalablement définie dans l'ontologie comme acceptant des arguments de ces trois types.



Une catégorisation des éléments du texte structurée comme celle-ci permet ensuite des recherches de corrélation entre des données décrites à différents niveaux d'abstraction, par exemple des données expérimentales.

Pour prédire de telles annotations sémantiques, les méthodes s'appuient sur des annotations intermédiaires appartenant à différents niveaux linguistiques interdépendants. Ma proposition est la suivante : les étapes d'analyse linguistique à chaque niveau sont modulaires et elles sont adaptées aux spécificités de la langue du domaine spécialisé et à l'objectif applicatif grâce à des connaissances spécifiques au domaine, lexiques, terminologies et ontologies. Ces ressources sont rarement disponibles et toujours incomplètes, coûteuses à acquérir manuellement. Elles sont apprises automatiquement ou semi-automatiquement par des méthodes dédiés. L'intégration des modules d'analyse linguistique et d'apprentissage dans une même chaîne de traitement d'analyse sémantique permet de configurer la chaîne de traitement pour une grande variété d'applications : de l'extraction d'information (EI) ou la recherche d'information (RI) jusqu'à la modélisation d'ontologie. Elle soulève également des questions complexes d'interaction entre modules.

3. La chaîne Alvis d'analyse linguistique et d'acquisition de connaissance

En 2001, nous avons défini avec Adeline Nazarenko spécialiste du traitement automatique de la langue naturelle (TALN) dans l'équipe RCLN (*Représentation des Connaissances et Langage Naturel*) du LIPN (*Laboratoire d'Informatique de Paris-Nord*), un projet de recherche pluridisciplinaire qui détaillait la typologie des connaissances en jeu et les méthodes de TALN et d'apprentissage et les questions de recherche ouvertes.

Ce projet a été publié à la conférence CIDE [55] (voir aussi [s78, s79]). Deux projets IMPG (*Action Informatique Mathématique Physique pour la Génomique*) *Caderige I* et *II* entre les partenaires LRI – Hélix (INRIA) – MIG – LIPN ont jeté les bases de ce programme pour une application en biologie moléculaire. La proposition et une première implémentation ont été publiés au workshop NLPBA (*Natural Language Processing in Biomedicine and its Applications*) de la conférence Coling en 2004 [31] et présentés dans des séminaires [s65, s76, s103, s105]. NLPBA est un des premiers workshops

dédiés à l'extraction d'information en biologie. La figure 1 décrit la proposition d'architecture en trois niveaux, (1) prétraitement du corpus, (2) analyse lexicale et (3) analyse sémantique.
Dans la chaîne d'analyse linguistique (bandeau bleu), le *prétraitement du corpus* comprend la sélection du corpus et la segmentation avec une première étape de reconnaissance d'entités nommée pour identifier les frontières de mot. L'*analyse lexicale* encadrée en rouge inclut l'étiquetage morpho-syntaxique et la lemmatisation, la reconnaissance des entités nommées et l'analyse terminologique. L'*analyse sémantique* encadrée en vert comprend l'analyse des dépendances syntaxiques, l'étiquetage des catégories sémantiques avec la résolution d'anaphore et la reconnaissance des relations sémantiques. Les paragraphes suivants de ce mémoire détaillent ces méthodes linguistiques et les méthodes d'acquisition de connaissances (niveau vert de la figure 1) qui produisent les ressources spécialisées nécessaires à ces traitements (niveau blanc de la figure 1).

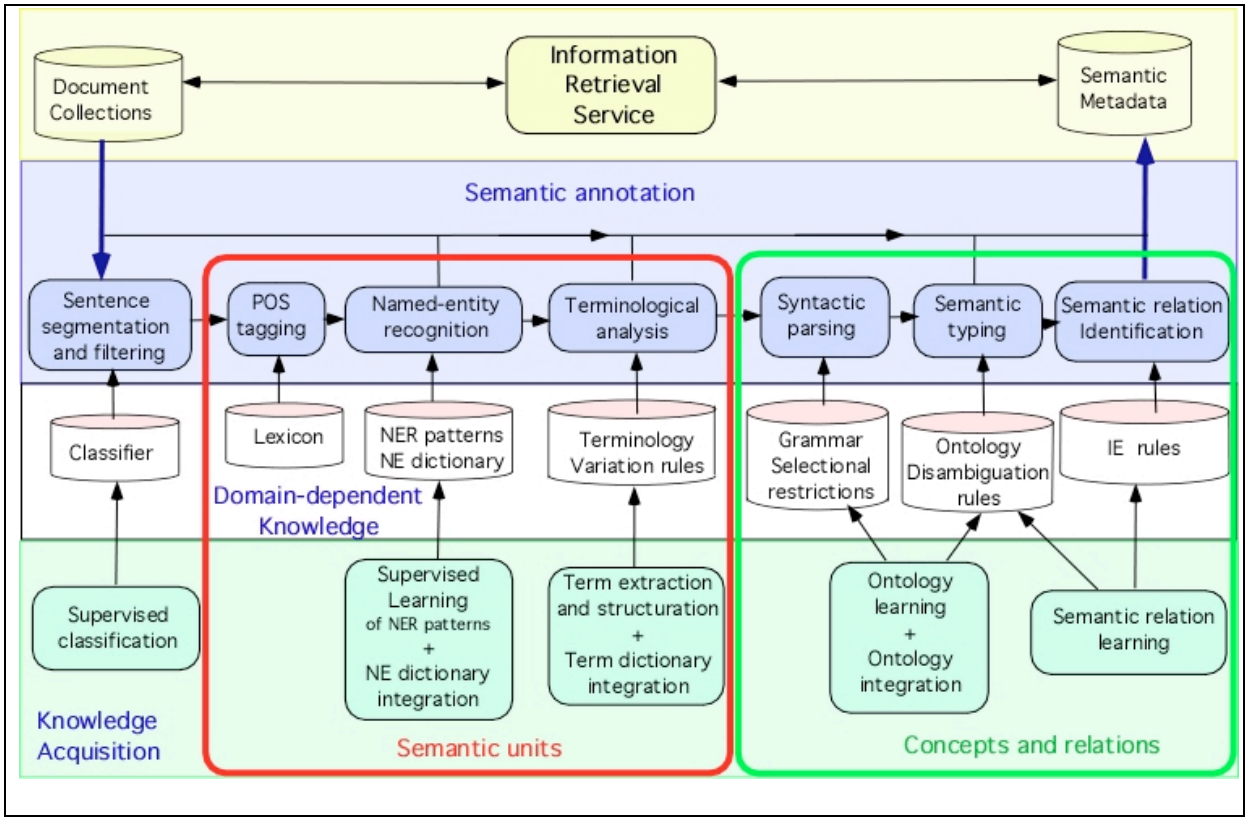


Figure 1. Architecture initiale de AlvisNLP/ML
(*Natural Language Processing and Machine Learning Alvis pipeline*).

Adeline et moi avons spécifié l'architecture, implémentée par nos deux équipes sous le nom Alvis NLP/ML (*Natural Language Processing and Machine Learning Alvis pipeline*) [l247] grâce à plusieurs projets successifs nationaux, Astuxe (RNRT[2]), ExtraPloDocs (RNTL) [r160, r161] et internationaux, Alvis (FP7) [30] et Quaero (Oséo). Nos deux articles précurseurs publiés dans *Ontology Learning from Text: Methods, Evaluation and Applications* [180] et *Ontology Handbook* [179] donnent une vue synthétique des enjeux et méthodes [51, r158, r126]. Mes contributions portent plus spécifiquement sur différents aspects de l'apprentissage automatique [43, 65, r159] et de l'application des méthodes à la biologie, publiées entre autres [28, 41, 42, 49, 55, 70, 181] dans *Intelligent Systems* [178]. Je les ai présentées dans des séminaires nationaux [s101, s103, s105, s111, s121, s122, s187, s188, s190] et internationaux [s77, s78, s79, s81, s82, s185, s186].
Alvis NLP/ML est dédié à l'analyse sémantique de documents scientifiques et techniques en français et en anglais. Elle diffère d'autres chaînes de traitement linguistique et d'annotation sémantique comme GATE (Bontcheva *et al.*, 2004) et plus récemment NLTK [Steven et al., 2009], OpenNLP ou UIMA (*Unstructured Information Management Architecture*) [Ferrucci et Lally, 2004] pour n'en citer

[2] Se reporter à l'annexe « Signification des acronymes » pour les formes développées.

que quelques unes. Dans Alvis, les modules d'acquisition de connaissance sont intégrés étroitement avec les modules d'analyse linguistique [s102]. Ils partagent une représentation des données communes. Les entrées des modules d'acquisition sont produites par la chaîne d'analyse linguistique qui exploite à son tour leurs sorties. Comparée à la plate-forme UIMA qui a une perspective plus large de traitement des données non structurées, Alvis est spécifique au traitement de documents et très simple à mettre en œuvre et à configurer par les non développeurs que sont certains membres de l'équipe. Le système de type de UIMA est inutilement généraliste pour le domaine qui nous concerne et rend la spécification plus complexe.
La version actuelle est le résultat du développement et de l'expérimentation de différentes versions successives du langage de représentation des données et du mode de configuration de la chaîne. En 2003 nous avions fait le choix de représenter les résultats des modules d'analyse linguistique par une annotation sémantique représentée en mémoire ou exportée dans le langage standard XML pour une meilleure interopérabilité interne et avec des modules externes. La représentation des annotations était déportée du texte (*stand-off*) pour permettre l'expression d'annotations discontinues et partiellement recouvrantes. Avec le LIPN et Erick Alphonse en post-doctorat et Alain-Pierre Manine en thèse, nous avons défini une DTD de représentation des connaissances organisée en différents niveaux de linguistiques [r153, r157]. Cette représentation partagée permet d'encapsuler un nouveau module dans la chaîne en écrivant simplement son interface (*wrapper*).
En 2008, Robert Bossy, ingénieur de recherche (IR) de l'équipe a respécifié et réimplémenté en Java le langage et la configuration. Sa solution très performante a été définitivement adoptée. Plus précisément, le langage de représentation en mémoire très succinct et générique permet une définition flexible et simple des interfaces des modules. La configuration des composants et ressources pour un traitement donné est définie par un plan unique. Le langage de définition de plan permet de configurer aisément les traitements en spécifiant les modules et ressources. Les plans sont eux-mêmes modulaires, partagés, réutilisés, combinés et adaptés par sa communauté d'utilisateurs.
Au-delà de l'objectif d'extraction d'information, l'annotation sémantique fine réalisée par AlvisNLP permet de développer un large panel d'applications documentaires, qui va de la recherche documentaire à la gestion et à la classification de documents répondant ainsi aux besoins grandissants d'accès à l'information scientifique. Une nouvelle application utilisant l'annotation sémantique est développable en quelques heures avec Alvis NLP/ML, sous réserve de disponibilité des documents et des bases de connaissance.

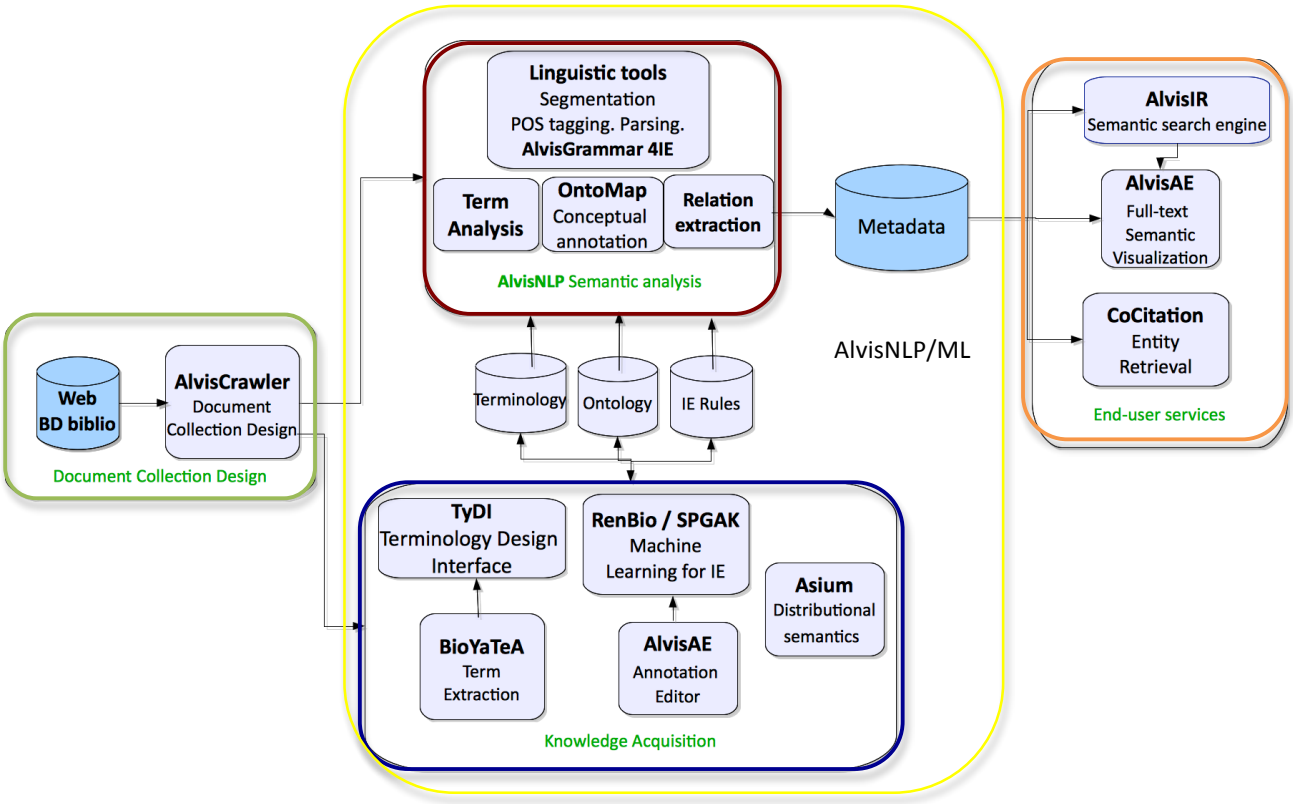


Figure 2. Architecture générale de la plateforme Alvis.

La figure 2 montre l'architecture générale de la plateforme et l'ensemble des modules développés par l'équipe sous ma direction. La chaîne AlvisNLP/ML (encadrée en jaune) reçoit un corpus constitué de documents au format texte constitué par le module d'aspiration de documents sur le Web, numérisés et éventuellement filtrés (encadré en vert). AlvisNLP/ML produit des métadonnées sémantiques (Metadata dans la figure) associées à des segments du texte qui sont utilisées par des modules destinés à l'utilisateur final (encadré en orange).

Les étapes d'annotation sémantique générique d'AlvisNLP sont adaptées à des tâches variées grâce des ressources déclaratives lexicales, syntaxiques et sémantiques, spécifiques au domaine étudié. Les modules du cadre bleu sont en charge de l'acquisition de ces connaissances. L'architecture d'Alvis NLP/ML permet d'intégrer et d'évaluer les logiciels d'acquisition de ces connaissances qui sont au cœur de notre recherche. La question critique de la production automatique d'exemples d'apprentissage dans une représentation adéquate est résolue par l'intégration étroite de la production et de l'acquisition.

L'évaluation d'AlvisNLP/ML et son extension par de nombreux modules ont été l'objet de notre participation au projet TIAE (*Text and Image Analysis Engine*) du programme Quaero en collaboration avec la société Jouve de 2008 à 2013 [r139, r140, r141]. Nous avons conçu plusieurs méthodes d'acquisition automatique ou semi-automatique de connaissances par apprentissage automatique à partir de corpus de documents. Les résultats d'apprentissage sont évalués à travers leurs performances en production et en interaction avec d'autres connaissances, ce qui permet de constater que des questions de recherche, considérées comme prioritaires pour une méthode prise isolément, se révèlent souvent secondaires une fois la méthode intégrée dans un processus complexe. De nouvelles questions de recherche sont apportées par l'interaction entre les étapes de traitement, par exemple, dans les boucles de rétroaction apprentissage / annotation.

Alvis NLP/ML représente donc un enjeu important sur les plans applicatifs et méthodologiques. Elle est maintenue par Robert Bossy et étendue par de nouveaux modules résultats des recherches de l'équipe. Elle est actuellement en cours de transfert chez des partenaires, dont les unités INRA Mét@Risk (*Méthodologies d'analyse et risque alimentaire*) et IATE (*Ingénierie des Agro-polymères et Technologies Émergentes*).

4. Pré-traitement du corpus, la classification de phrases

Il est fréquent que les connaissances spécialisées dont l'acquisition à partir de texte est automatisée soient dispersées dans de larges corpus de documents. Le filtrage ou la classification de paragraphes ou de phrases permet de focaliser les traitements sémantiques sur les parties pertinentes. La classification de documents appliquée à une représentation dite en sac de mots obtient des résultats satisfaisants dans le domaine général, mais n'avait été que très peu étudiée dans des domaines spécialisés et pour des petites collections de documents très courts comme le sont les phrases [Marcotte et al., 2001]. Pour traiter la question spécifique de l'extraction d'interactions géniques, j'ai encadré la thèse CIFRE de Mohammed Ould Abdel Vetah avec la société Valigen [269] puis le stage de DEA de Zhiu Quan [265] du LRI. Pour le filtrage, nous avons étudié l'apport de différentes informations linguistiques (lemmes, entités nommées, termes) et comparé différentes méthodes de sélection d'attributs et de classification discriminante, induction d'arbre de décision, bayésien naïf, SVM *etc*..

Le corpus d'évaluation porte sur les interactions géniques : la méthode doit prédire quelles sont les phrases qui mentionnent ou non une interaction. Philippe Bessières a sélectionné un corpus dit *transcript* de 2 500 références de la base bibliographique PubMed[3] sur la transcription de gène chez la bactérie modèle *Bacillus subtilis* [r164, c242]. Ce corpus est encore largement utilisé aujourd'hui comme *benchmark* comme nous verrons plus bas. Un ensemble de 1 023 phrases du corpus *transcript* ont été annotées par Philippe Bessières comme mentionnant ou non une interaction génique. Les phrases sélectionnées mentionnaient au moins deux gènes ou protéines. L'interaction génique devait impliquer des entités présentes dans la phrase. Les résultats de la méthode de classification ont été mesurés en terme de rappel et de précision en validation croisée. La méthode a obtenu un score de

[3] http://www.ncbi.nlm.nih.gov/pubmed

l'ordre de 85 % sur ce corpus [r165], élevé comparé à ceux de Marcotte et al. Les résultats ont été publiés à CAP [57], PKDD [12] et à JOBIM [54] en 2001 [53, 57, 69, s106].
Nous avons généralisé et implémenté la méthode avec Alain-Pierre Manine (en thèse ASC INRA bioinformatique), dans le logiciel en ligne STFilter[4] dédié au filtrage d'interactions géniques. Etant donné une espèce, STFilter télécharge automatiquement depuis MedLine l'ensemble des articles publiés sur cette espèce, segmente et sélectionne les phrases qui contiennent au moins deux noms de gènes ou de protéines et surtout, filtre automatiquement celles qui sont susceptibles de mentionner une interaction génique. Les utilisateurs biologistes de STFilter utilisent les classifieurs déjà appris ou annotent en ligne comme positives ou négatives les phrases proposées par le logiciel pour entraîner un nouveau classifieur (Figure 3). Nous avons entraîné STFilter sur divers corpus de références bibliographique MedLine annotés pour différentes espèces, principalement la bactérie modèle *Bacillus subtilis*, mais aussi la drosophile [57, 71, s104] en collaboration avec Bernard Jacq de l'IBDML (*Institut de Biologie du Développement de Marseille Luminy*), et sur le métabolisme des lipides chez l'homme, le rat et le poulet en collaboration avec Sandrine Laguarrigue de l'ENSAR (*École nationale supérieure agronomique de Rennes*) dans les projets Caderige I et II.

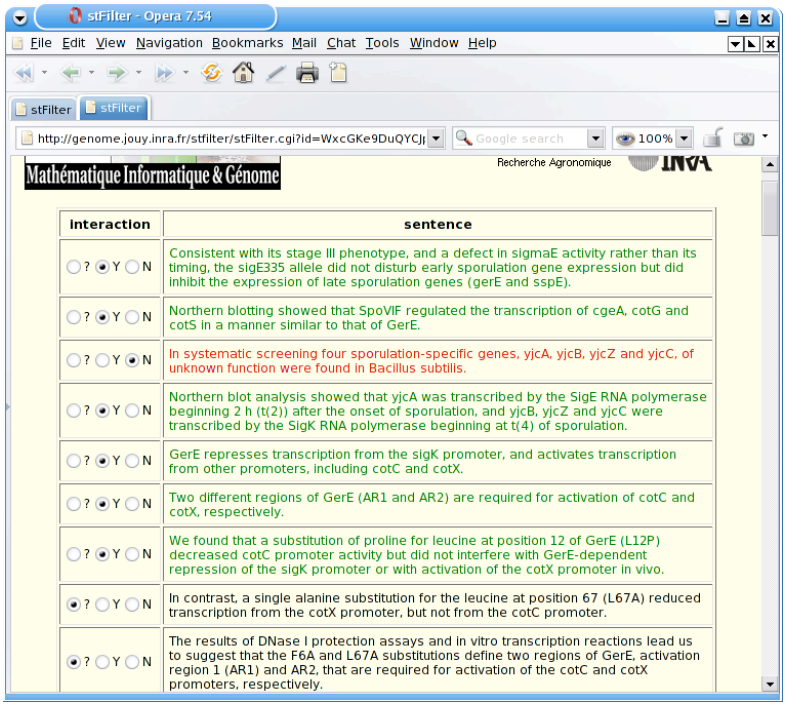


Figure 3. Fenêtre de classification de STFilter pour le filtrage de phrases.

La qualité de la classification a été améliorée en distinguant plusieurs classes selon une typologie biologique, selon que l'interaction était exprimée comme une liaison protéine-ADN (*binding*), un phénotype (c'est-à-dire une mutation empêche l'expression du gène), ou l'appartenance d'une protéines au regulon d'un gène (*regulon membership*). Ce dernier travail a été réalisé par Audrey Lemaçon en apprentissage (Master Egoist de Rouen) [271] et co-encadrée avec Philippe Bessières. Il est implémenté dans le logiciel CoCitation où la distinction du type d'interaction améliore les résultats de 8 % [l252].
L'intégration de CoCitation dans la chaîne Alvis est due à Julien Jourde et la finalisation de son interface est réalisée par Thomas Lacroix (Figure 4), sous ma direction. Le service est aujourd'hui accessible en ligne par un navigateur[5]. La figure 4 montre une fenêtre de résultats de CoCitation où trois phrases mentionnent des interactions géniques de types phénotype et régulation figurés par les ronds verts et rouges. L'identification des gènes et protéines *recA*, *addA*, *etc*. est le résultat d'un autre ensemble de méthodes décrit au paragraphe suivant.

[4] http://genome.jouy.inra.fr/stfilter/
[5] http://migale.jouy.inra.fr/CoCitations

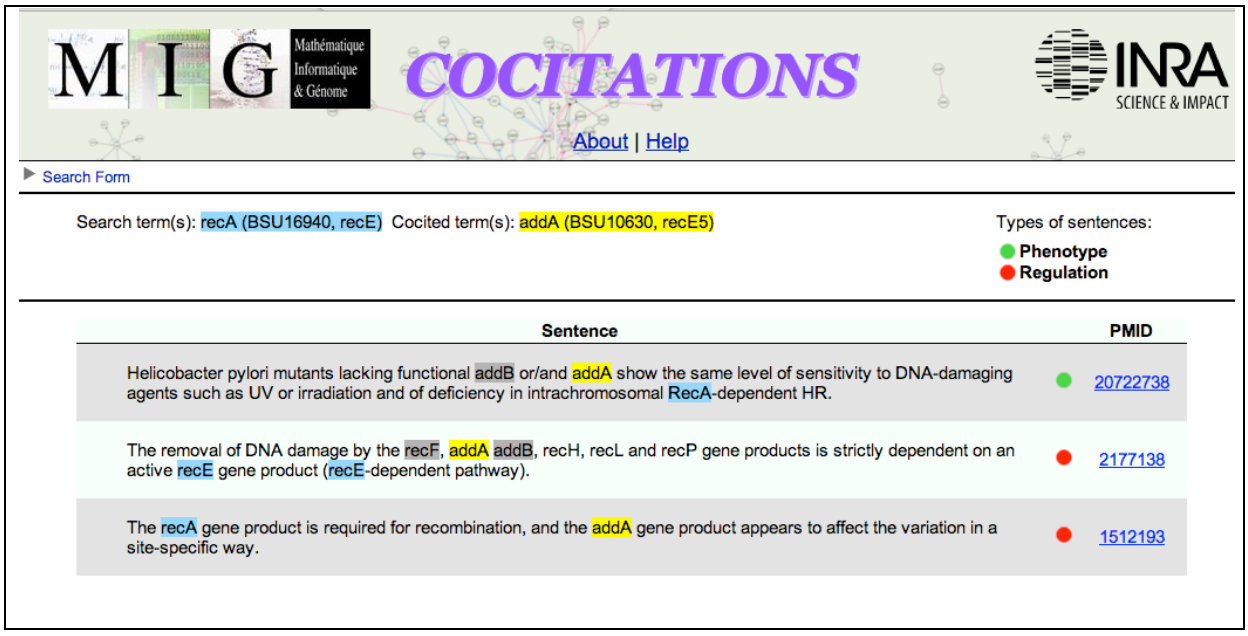


Figure 4. Fenêtre du logiciel CoCitation pour la détection d'entités et la classification de phrases.

5. La reconnaissance automatique des entités

La reconnaissance des entités est reconnue comme une étape critique du traitement documentaire dans les domaines scientifiques et techniques [Nadeau et Sekin, 2007]. Les méthodes diffèrent selon que les entités sont dénotées par des désignateurs figés tels que les noms propres ou les formules [Kripke, 1982], ou par des termes susceptibles de variations morpho-syntaxiques et peu référencés dans des nomenclatures. Une distinction classique veut que les entités nommées représentent des instances de concept, tandis que les termes dénotent des concepts du domaine. Dans le domaine, de la biologie moléculaire la distinction entre instances et concepts est rarement pertinente. Il est rare en effet qu'il soit fait référence à des instances qui seraient des individus particuliers. De manière générale, les entités mentionnées ne sont pas des instances, mais des concepts. Par exemple, la mention *Bacillus subtilis cotA gene* désigne les gènes *cotA* de tous les individus de l'espèce *Bacillus subtilis*, indépendamment même de la souche.

Un même objet biologique est fréquemment dénoté par par une forme figée et par un terme, par exemple, les deux formes *4'-phosphopantetheinyl transferase* et *sfp* désignent la même protéine. Nous entendons donc ici les entités nommées comme désignées par des formes figées, qu'il s'agisse d'instances ou de concepts.

5.1 Reconnaissance de gènes et protéines

Dans le domaine de la biologie comme dans tous les domaines scientifiques et techniques, la détection de nombreuses entités est essentielle à l'exploitation de la bibliographie. Les noms d'espèces, de gènes ou de protéines sont critiques pour l'extraction des réseaux de régulation génique dont ils sont les acteurs. La grande fréquence des homonymes et des synonymes et le manque de nomenclature stable et complète rendent la tâche complexe et d'autant plus nécessaire. Durant la décennie passée, une abondante littérature et des compétitions ont traité le sujet, principalement pour les eucaryotes [Krallinger et al., 2008]. Nous avons contribué à la reconnaissance des noms de gène et de protéine pour les procaryotes, c'est-à-dire les espèces dont les cellules n'ont pas de noyau. Le nommage des objets biologiques chez les procaryotes et leur contexte d'apparition suit des règles partiellement différentes. En particulier, les descriptions biologiques sont bien plus fréquentes chez les procaryotes et traitent des interactions moléculaires. Chez les eucaryotes, les descriptions sont plus fréquemment physiologiques ou phénotypiques, en terme d'impact sur le métabolisme ou sur des traits observables.

Nous avons conçu et évalué avec Frédéric Papazian et Robert Bossy la méthode de reconnaissance de gènes et de protéines *RenBio* [l245], qui utilise l'algorithme d'induction d'arbre de décision C4.5. Il est appliqué à des exemples représentés par l'analyse morpho-syntaxique des mots et un ensemble d'attributs typographiques (casse, présence de symboles non-alphanumériques, etc.) calculés par

AlvisNLP. RenBio a été évaluée sur un ensemble de résumés dit *Quaero_t3.2_gene* que j'ai annoté manuellement [c241, r136], en l'absence de corpus de gènes annotés pour les procaryotes. J'ai extrait du corpus *transcript* un ensemble représentatif de documents sur le phénomène de la sporulation et annoté les noms de gènes, de protéines et d'espèces en accordant une grande importance aux frontières de manière à produire une annotation homogène. Les résultats de la méthode RenBio sur ce corpus confirment l'importance des attributs typographiques et du voisinage, par rapport aux information linguistiques. Principes d'annotation et résultats sont publiés dans les actes du workshop *Data and text mining for integrative biology* associé à ECML que j'ai organisé en 2006 [29]. Une fonction de désambiguïsation endogène a été ajoutée à RenBio qui consiste à propager en cours d'apprentissage au reste des occurrences d'apprentissage du document l'information sur les décisions prises sur les occurrences les moins ambiguës [r147]. La méthode a été intégrée dans Alvis [r152, r154] et comparée plus récemment avec d'autres méthodes du projet Quaero sur le même corpus *Quaero_t3.2_gene*, où elle montre une précision élevée, mais un rappel médiocre par rapport à des méthodes plus récentes comme les CRF. Les résultats ont été publiés à LREC en 2010 [10].

5.2 Normalisation des noms de gènes et de protéines

La question de la normalisation des noms est importante en biologie [Lu et al., 2011]. Elle est particulièrement critique pour les gènes de la bactérie modèle *Bacillus subtilis*. Chaque gène connaît successivement plusieurs noms, un nom provisoire quand sa fonction n'est pas connue, par exemple, *yvud*, puis un nom reflétant sa fonction, par exemple, *gerE* pour *germination*. Plusieurs raisons peuvent ensuite justifier un nouveau changement de nom, soit un changement de fonction, l'identification d'un opéron (une séquence de gènes) à la place du gène, la découverte de l'identité avec un autres gène qu'on croyait distinct, *etc*. A cela s'ajoute vers 1994, l'alignement malheureusement non documenté des noms de gènes de *Bacillus subtilis* avec ceux de l'espèce modèle *Escherichia coli*. De façon dommageable, les noms devenus obsolètes sont fréquemment réutilisés, par exemple *DnaK* est un nom de protéine utilisé successivement pour nommer *DnaA* et *DnaN,* deux autres protéines tout à fait distinctes. Le référencement systématique des synonymes de noms de gènes et de protéines de *Bacillus* que j'ai fait avec Julien Jourde nous a montré l'incomplétude des nomenclatures existantes. Par exemple, seuls 51% des noms de gènes étaient communs aux trois principales nomenclatures, GenBank[6], SwissProt[7] et Subtilist[8]. De ce fait, les recherches bibliographiques des gènes concernés manquent les articles anciens, ou ramènent des articles citant un ancien homonyme. Le résultat de notre inventaire est utilisé par le logiciel CoCitation pour la reconnaissance de noms de gènes et de protéines de *Bacillus subtilis*.
S'ils sont fréquemment absents des nomenclatures, les renommages de gènes et de protéines sont par contre toujours documentés dans des articles quand la cause du renommage est d'origine biologique. L'extraction automatique de la littérature des mentions de renommage permet donc d'enrichir systématiquement la nomenclature des synonymes. Julien Jourde et Pierre Warnier, en thèse ont développé sous ma direction et celle de Philippe Veber, IR de l'équipe une approche à base de classification discriminante (SVM) et d'expressions régulières. Pour mesurer la qualité de la méthode et la comparer avec d'autres méthodes, Julien Jourde, avec Philippe Bessières et moi en collaboration avec l'équipe de Claire François à l'INIST avons annoté manuellement un corpus de référence de 461 renommages dans 1 836 documents [c241]. Nous avons organisé la tâche d'extraction de règles de renommage intitulée *Rename Task*[9] dans la *Bacteria Track* de la compétition internationale *BioNLP Shared Task* 2011 avec ce corpus [3].

Par exemple, à partir de

"*We propose to rename yusC, yusB and yusA as metN, mete and metO, respectively.* "

Il s'agit d'extraire les trois couples de synonyme (*yusC, metN*) (*yusB*, *mete*) et (*yusA*, *metO*). L'exemple est simple dans ce cas, mais le renommage n'est pas toujours explicite. Par exemple,

[6] http://www.ncbi.nlm.nih.gov/bioproject/PRJNA76
[7] http://www.uniprot.org/uniprot/?query=%28name%3Aprotein+AND+organism%3A%22bacillus+subtilis%22%29+AND+reviewed%3Ayes
[8] http://genolist.pasteur.fr/SubtiList/
[9] https://sites.google.com/site/bionlpst/home/bacteria-gene-renaming-rename

"Thus, a separate spoVJ gene as defined by the 517 mutation does not exist and is instead identical with spoVK."

Indique que les deux gènes *spoVJ* et *spoVK* que l'on croyait distincts sont en fait les mêmes. La conséquence est que *spoVJ* disparaît et est renommé *spoVK*. Pour mettre au point la méthode, nous avons défini une typologie des causes du renommage qui se manifestent sous forme linguistique, par exemple la division d'un gène en gènes d'opéron. La typologie a été mise au point lors d'une *training session* que j'ai organisée à Jouy-en-Josas avec dix étudiants en thèse de l'action européenne Marie Curie ITN *Transys* (*A systems approach to defining membrane protein networks and applications*). En perfectionnant la méthode, Pierre Warnier a obtenu des résultats élevés (73% F-mesure) [26] à la tâche *Rename Task* [2, 20]. Ses résultats sont inférieurs à ceux de la meilleure méthode (87%) qui utilise l'information des dépendances syntaxique. Cela montre la limite de l'approche sac de mots pour extraire de l'information relationnelle des publications, en particulier dans le cas des énumérations comme dans l'exemple ci-dessus.

Ces résultats de synonymie et d'homonymie sont interrogeables par le logiciel CoCitation pour la recherche de noms de gènes cocités dans la bibliographie. La figure 4 montre l'exemple de recherche de la cocitation des gènes dnaA, abrB et adeC et de tous leurs synonymes.

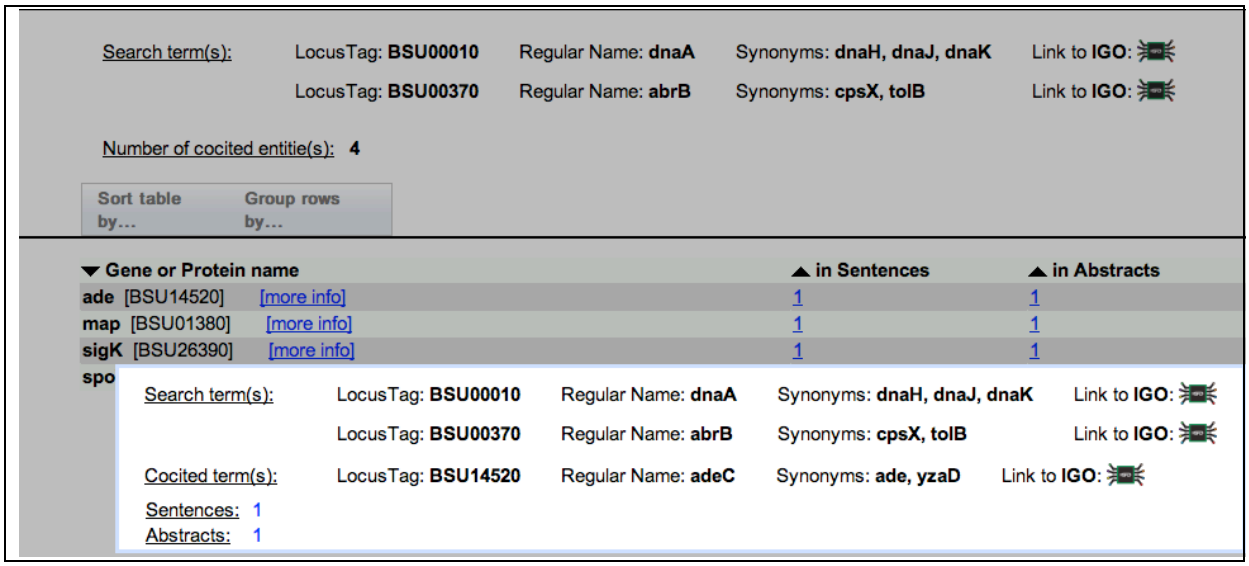


Figure 4. Fenêtre du logiciel CoCitation.

La chaîne AlvisNLP identifie d'abord les noms de gènes, de protéines et d'espèces dans les références de la base bibliographique PubMed. Les résultats sont stockés dans une base de données. Ils sont affichés à la demande, sous la forme d'extraits surlignés, de phrases ou de résumés. CoCitation est intégré dans IGO[10], le portail bioinformatique de l'unité MIG, de telle sorte que l'utilisateur a accès directement à toute l'information génétique.

5.3. Extraction de termes à partir de corpus

Avec les entités figées, l'identification de termes pertinents est à la base de toute analyse sémantique de texte. Un terme est une unité lexicale qui dénote un concept du domaine [Cabré, 1999] par exemple, *FHB resistance*. Les terminologies existantes sont en général peu adaptées à l'annotation automatique de documents spécialisés parce qu'elles sont incomplètes et que les termes, généralement choisis à des fins d'indexation manuelle, sont majoritairement différents des termes effectivement utilisés dans un texte rédigé. Dans le cas d'une terminologie, les termes doivent être compacts et compréhensibles hors contexte. Inversement, dans le texte, les termes courts sont préférés parce que plus compréhensibles, les répétitions sont à éviter et les paraphrases fréquentes. Dans les différents domaines de la biologie que nous avons traité, les termes des thésaurus Agrovoc, MeSH, ou les labels des ontologies, Gene Ontology ou EnvO sont insuffisants pour analyser les termes des textes. Seuls

[10] http://migale.jouy.inra.fr/IGO/

moins de 10 % des termes sont retrouvés et sont principalement des monotermes généraux et peu informatifs comme *activity*.

L'acquisition automatique de termes à partir de corpus est reconnue comme un moyen puissant d'assister la construction de terminologie dans des domaines spécialisés [Bourigault et al, 2004]. Les extracteurs de termes implémentent deux grandes classes de méthodes, statistiques basées sur des fréquences de coocurrences et linguistiques basée sur des informations syntaxiques et lexicales [Pazienza et al., 2005]. Je préfère la seconde qui permet l'extraction de termes rares et qui analyse la structure syntaxique interne des termes. Cette analyse permet d'identifier les sous-termes, de traiter les variations terminologiques et plus encore produit une information utile à la catégorisation conceptuelle comme nous verrons au paragraphe 7.

Depuis 2003, nous avons comparé plusieurs extracteurs de termes de l'état de l'art, notamment *Acabit* [Daille, 1994], *Nomino*[11], *Lingway*, *Syntex* [Bourigault, 1992] et *YaTeA* développé par Sophie Aubin et Thierry Hamon au LIPN dans le projet Alvis (2006). L'évaluation a été conduite en validant les termes extraits automatiquement et en comparant les termes extraits aux termes de référence annotés manuellement dans des documents. La comparaison a porté sur deux domaines et genres, des articles de biologie moléculaire et des brevets européens en agronomie. L'évaluation en biologie, avec Sophie Aubin pour la compétence terminologique et les biologistes Alain Kotoujansky et Philippe Bessières a montré l'importance des termes longs, incluant éventuellement des groupes prépositionnels [c242]. L'évaluation sur les brevets s'est faite dans le projet européen Epipagri (*Towards European Collective Management of Public Intellectual Property for Agricultural Biotechnologies*) à des fins d'indexation [s123, r148]. Annick Lacombe terminologue de la DV/IST (*Direction de la Valorisation, Information Scientifique et Technique*) et Franck Le Guerhier, chargé de valorisation d'INRA Transfert ont contribué à l'évaluation et la conception d'une terminologie pour l'indexation des brevets [r149, c243].

Nous avons finalement fait le choix de l'extracteur de terme YaTeA pour la qualité de ses extractions, mais surtout pour son adaptabilité à de nouveaux domaines. L'extracteur Syntex obtenait des résultats plus riches, mais moins précis. La méthode de YateA est purement linguistique. Elle consiste à appliquer itérativement des patrons morpho-syntaxiques sur le texte préalablement segmenté et analysé, entre des frontières définies par des mots ou des catégories syntaxiques. Par exemple, dans le fragment

During[ADV] sporulation[NOUN] of[PREP] Bacillus subtilis[P-NOUN], spore[NOUN]

Après analyse, l'adverbe *during*, la virgule et le verbe *encoded* sont identifiés comme frontières. Les syntagmes restant, *sporulation of Bacillus subtilis* et *spore coat proteins* vérifient les deux patrons [NOUN PREP P-NOUN] et [NOUN*]

During / sporulation of Bacillus subtilis /, / spore coat proteins / encoded by /

Un des points forts de YaTeA est sa capacité à calculer des sous-termes de termes complexes, à partir de leur occurrence dans le corpus et des termes certifiés donnés en entrée. Patrons et frontières sont définis de façon déclarative et sont modifiables. Un post-filtrage permet d'exprimer plus directement les contraintes.

Wiktoria Golik et moi avons significativement amélioré les performances de YateA en étendant les patrons et filtres dans une nouvelle version appelée *BioYateA*. Nous nous sommes attachées en particulier à traiter deux problèmes : le traitement des adjectifs verbaux et des attachements prépositionnels. Ils ont été identifiés par des études systématiques des termes extraits de plusieurs corpus de transcription moléculaire, brevets de pharmacologie et articles de physiologie animale. Les formes gérondives et participiales sont très fréquentes dans les articles décrivant des expérimentations. Par exemple,

(1) dans *Protein binding can influence*, *binding* doit être analysé comme un nom, et non comme un verbe,
(2) comme un adjectif dans *the binding protein*,
(3) et comme un verbe dans *when binding to a target*.

[11] version 4.2.22 updated the 25 July 2001. Available in http://www.nominotechnologies.com

La question de l'attachement des groupes prépositionnels est bien connue. Nous avons tout particulièrement traité les prépositions *at* et *to*, fréquentes dans nos documents. Par exemple dans *susceptibility to mastitis* ou *body weight at birth*, les groupes prépositionnels appartiennent au terme, ce qui n'est pas le cas dans *pigs at slaughter* extrait de *welfare of pigs at slaughter* où *at slaughter* est rattaché à *welfare*.
La démarche a consisté identifier le contexte d'apparition de ces formes et à les traduire en de nouveaux patrons morpho-syntaxiques et en règles de post-filtrage intégrés dans BioYaTeA. BioYaTeA améliore les résultats de YaTeA de plus de 20 points sur plusieurs corpus. Le détail de la méthode et des évaluations pourra être trouvé dans notre article publié à CICLing en 2013 et à paraître dans *International Journal of Computational Linguistics and Applications* en 2013 [1]. Les évaluations sont automatiques par rapport à une terminologie de référence et manuelles par validation des termes extraits.
Nous distribuons la méthode BioYateA sous forme d'un module public CPAN[12] [l250] depuis janvier 2013. Nous avons contribué à la nouvelle version de YateA développée par Thierry Hamon, disponible sous la même forme[13]. BioYateA a été retenu comme *supporting resource* de BioNLP-ST'13 [Stenetorp et al., 2013], c'est-à-dire que l'analyse terminologique des corpus des six tâches a été fournie aux participants de la compétition.
BioYateA est intégré dans la chaîne Alvis. La lemmatisation et les catégories morpho-syntaxiques nécessaires à BioYateA peuvent être calculées par GeniaTagger [Tsuruoka et Tsujii, 2005] ou TreeTagger [Schmidt, 1994] qui sont à sélectionner selon les domaines. Nos évaluations successives sur différents corpus ont confirmé la supériorité de GeniaTagger pour la biologie.
Les perspectives de recherche à court terme sur BioYaTeA portent sur une représentation de « macro-termes », c'est-à-dire sous forme de patrons lexico-syntaxiques qui permettrait une représentation plus concise d'ensembles de termes, afin de faciliter la création et la maintenance de terminologies en limitant l'explosion combinatoire. La représentation permettrait de représenter des sous-termes par des ensemble des termes appartenant à des catégories sémantiques de manière similaire à FastR [Jacquemin, 1996]. Elle permettrait également la représentation de sous-termes optionnels et sur des positions alternatives, tout en conservant de bonnes performances en terme de complexité. La représentation des termes dénotant des phénotypes du blé est le premier domaine expérimental traité. Par exemple, le macro-terme {*resistance* or *adaptation*} to {WHEAT-DISEASE} représente l'ensemble des termes de résistance et d'adaptation aux maladies référencées dans le dictionnaire WHEAT-DISEASE.

6. Normalisation et structuration terminologique

Les extracteurs de termes comme BioYateA produisent une grande quantité de termes candidats, non structurés sur le plan sémantique. Leur validation et leur structuration en terminologies, ou mieux encore, leur utilisation dans la construction d'ontologie à partir de corpus nécessite une interface homme-machine adaptée. L'importance du développement de terminologies pour nos applications en biologie a rendu nécessaire un outil intégré pour la validation, la structuration et la modélisation ontologique des termes extraits. Les solutions existantes privilégient la validation de terme sans inclure de structuration, comme *TermExtractor* [Selano et Velardi, 2007], ou l'édition d'ontologie, sans prendre en compte le niveau terminologique comme l'éditeur d'ontologie *Protégé* [Noy et al., 2001], ou enfin traitent ontologies et terminologies, mais de manière distincte comme *Terminae* [Aussenac-Gilles et al., 2008] ou *Dafoe* [Charlet et al., 2010]. Leurs interfaces homme-machine s'adressent à des ingénieurs de la connaissance ou à des terminologues. A l'inverse, nous faisons l'hypothèse que les tâches de validation et de structuration terminologique et de modélisation conceptuelle puissent être partagées entre l'ingénieur de la connaissance et les experts du domaine. Ces derniers doivent pouvoir directement interagir avec l'interface sans l'intermédiation de l'ingénieur de la connaissance.
Dans ce but, nous avons spécifié le logiciel TyDI (*Terminology Design Interface*) [l246] avec Sophie Aubin et Wiktoria Golik. TyDI permet d'explorer efficacement de grands volumes de termes extraits, de les valider et de les structurer par des relations de synonymie et d'hyperonymie (de généralité) et de

[12] http://search.cpan.org/~bibliome/Lingua-BioYaTeA/
[13] http://search.cpan.org/~thhamon/Lingua-YaTeA-0.5/

les relier aux concepts d'une ontologie sous forme hiérarchique. La figure 5a présente les principales fenêtres de TyDI avec un exemple de l'ontologie ATOL (*Animal Trait Ontology*) qui porte sur les traits des animaux de rente. La fenêtre encadrée en vert montre les termes extraits automatiquement qui vérifient les conditions exprimées par l'utilisateur dans la fenêtre encadrée en bleu. La fenêtre encadrée de parme affiche les contextes d'apparition du terme sélectionné dans les documents. La fenêtre encadrée de rouge affiche l'ontologie. La fenêtre qui affiche la structure de la classe sémantique d'un terme avec son terme vedette, ses synonymes et ses hyper- et hypoymes n'apparît pas ici. Par exemple *whithdrawal response* est synonyme du terme *withdrawal reaction* et hyponyme du terme *pain response*.

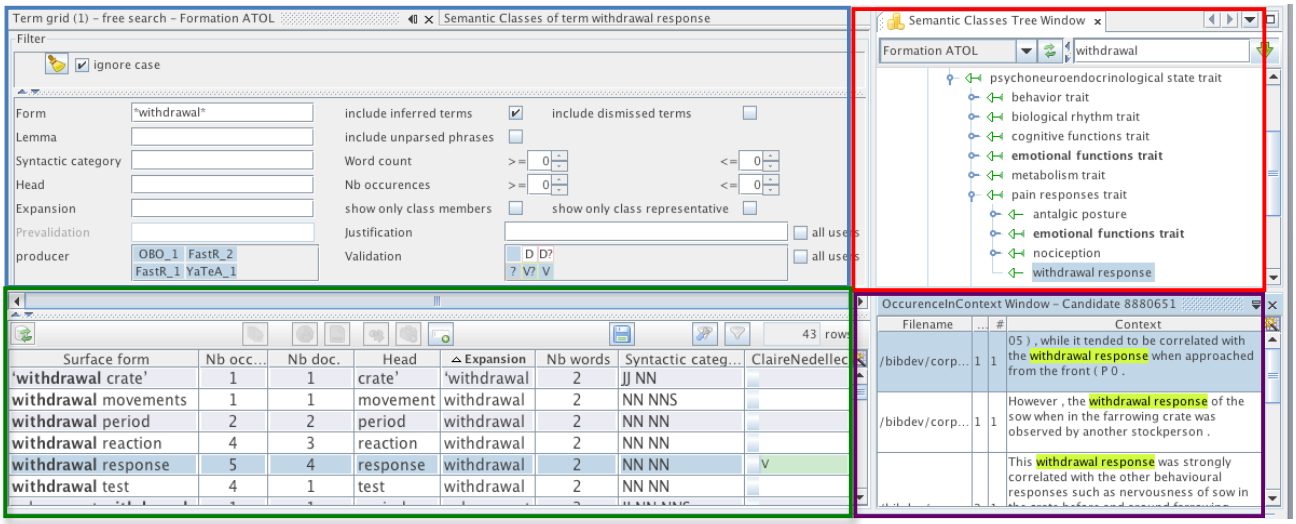
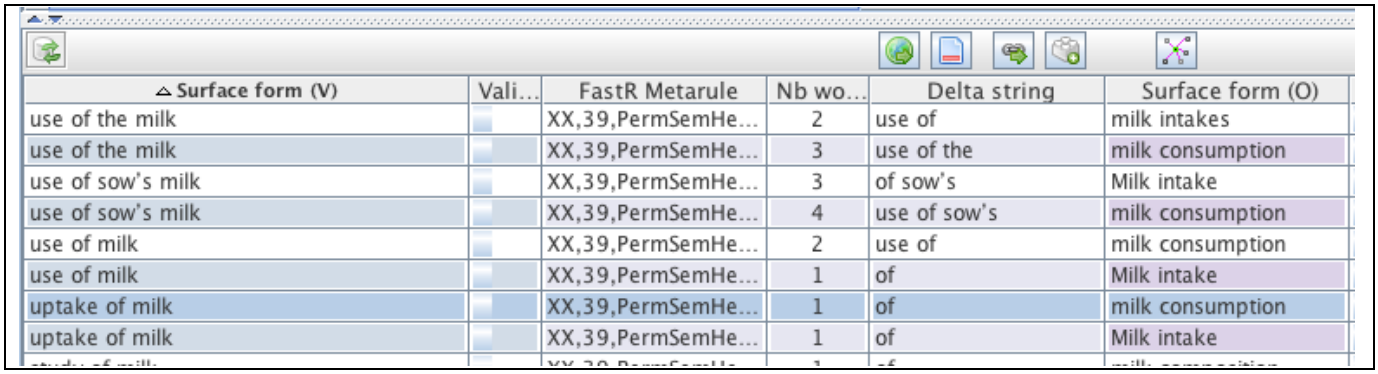


Figure 5. TyDI (Terminology Design Interface).

TyDI prend en entrée des candidats termes extraits de corpus ou des termes certifiés, ainsi que des variations terminologiques produites par l'outil *FastR* [Jacquemin, 1996] qu'il permet de valider et structurer. La figure 5b présente un exemple où TyDI affiche les différentes variations morphosyntaxiques *milk intake*, *milk consumption*, etc. (colonne de droite) du terme *use of the milk* (colonne de gauche). Le thésaurus WordNet a été utilisé ici pour inclure des proximités sémantiques comme entre *intake* et *consumption*. La figure 5c présente le graphe de variations plus facile à interpréter que la liste, quand les couples de variations sont nombreux dans une même classe d'équivalence.



Surface form (V)	Vali...	FastR Metarule	Nb wo...	Delta string	Surface form (O)
use of the milk		XX,39,PermSemHe...	2	use of	milk intakes
use of the milk		XX,39,PermSemHe...	3	use of the	milk consumption
use of sow's milk		XX,39,PermSemHe...	3	of sow's	Milk intake
use of sow's milk		XX,39,PermSemHe...	4	use of sow's	milk consumption
use of milk		XX,39,PermSemHe...	2	use of	milk consumption
use of milk		XX,39,PermSemHe...	1	of	Milk intake
uptake of milk		XX,39,PermSemHe...	1	of	milk consumption
uptake of milk		XX,39,PermSemHe...	1	of	Milk intake

Figure 5b. Exemple de variations calculées par FastR et visualisées dans TyDI.

TyDI prend également en entrée des ontologies ou des terminologies existantes, dans différents formats à des fins d'intégration et de réutilisation. Il produit des terminologies et des ontologies dans les langages standard du web sémantique.

Notre objectif était de concevoir un outil qui permette un usage collectif de manière à mobiliser des groupes d'experts dans une construction collaborative. Avec Frédéric Papazian qui a implémenté TyDI, nous avons fait le choix technique d'une architecture client-serveur avec une base de données côté serveur. Le client interagit en permanence avec le serveur de telle sorte que toutes les modifications soient enregistrées et visibles par tous à tout moment. Un système de verouillage empêche les accès concurrents. Le client riche permet une représentation ergonomique des différentes connaissances en jeu [r134].

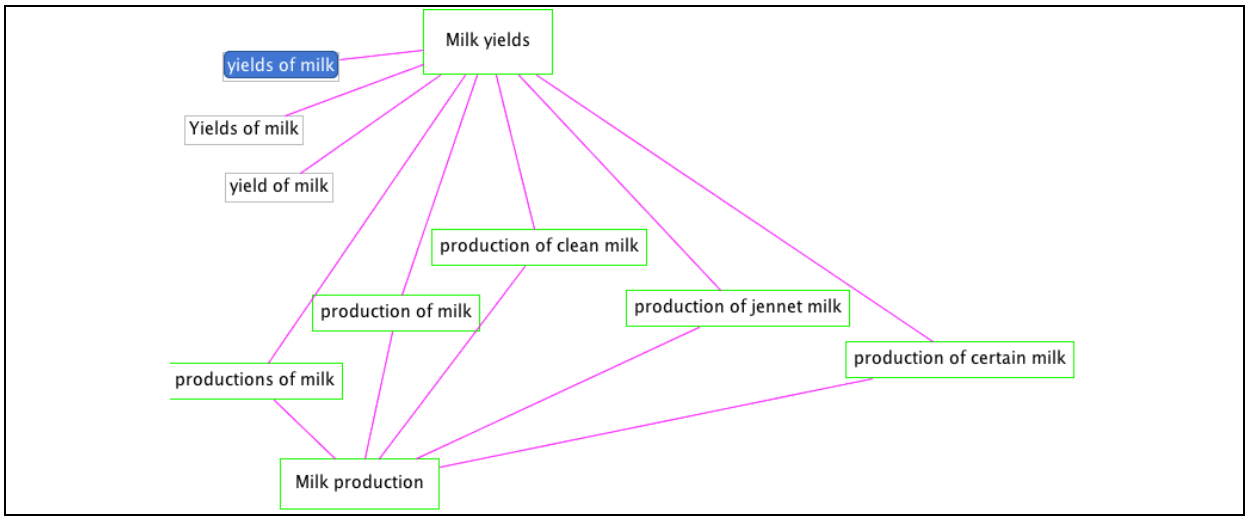


Figure 5b. Exemple de variations calculées par FastR et visualisées dans TyDI.

Comme détaillé dans les articles publiés à EKAW [9] et TIA en 2011 [68], TyDI est à la fois plus complet qu'une interface de validation de terme, puisqu'il permet la structuration terminologique, et plus complet qu'un éditeur d'ontologie, puisqu'outre la modélisation des concepts de l'ontologie, il assure leur lien au texte à travers le niveau terminologique. Ce lien explicite entre les niveaux lexical et conceptuel est indispensable à l'annotation sémantique de texte [s86].
TyDI ne permet pas de modéliser d'autres relations que la relation de généralité, pour des raisons ergonomiques, de compréhensibilité des tâches de structuration terminologique et de modélisation par des experts. Nos différentes expériences dans diverses applications ont montré la difficulté de modéliser de façon unifiée et systématique des relations spécialisées à partir de termes. Cette fonction relève d'un éditeur d'ontologie dédié et nécessite une démarche de modélisation plus encadrée.
Les choix méthodologiques de TyDI ont été validés dans plusieurs applications pour différents besoins [r135]. Elles ont contribué à affiner la démarche en fonction des objectifs et des compétences impliquées. Une interaction étroite avec les spécialistes des domaines des applications a fait évoluer notablement l'ergonomie de l'interface. J'en cite ici trois par ordre croissant de complexité qui impliquent Wiktoria Golik, terminologue et ingénieur de la connaissance. Une première utilisation par un expert seul, Bernard Teyssendier du département INRA Biologie Végétale dans le projet VegA sur l'indexation de brevets pour l'exploitation de la biomasse [s207] a montré tout l'intérêt de concevoir terminologie et ontologie dans un aller-retour constant. Les détails de cette expérience pourront être trouvés dans notre article à EKAW en 2011 [9].
Notre contribution au projet *ATOL* (*Animal Trait Ontology*) du département INRA Phase (*Physiologie animale et systèmes d'élevage*) porte sur la lexicalisation des concepts et l'enrichissement d'une ontologie existante par de nouveaux concepts de manière à l'utiliser pour l'indexation automatique de documents [48]. Un corpus représentatif du journal *Animal* a constitué la source de termes candidats. Les termes extraits par BioYaTeA ont été validés et structurés par des experts travaillant indépendamment les uns des autres sur des domaines disjoints et coordonné par Wiktoria Golik sous ma direction et Léa Joret de l'unité INRA Scribe. L'amélioration de l'ontologie mesurée par sa couverture sur le corpus à indexer a été considérable, démontrant la pertinence de l'approche. La méthode et les résultats sont détaillés dans un article à MTSR'12 [8] [66]. L'ontologie ATOL ainsi

enrichie est accessible en ligne[14]. La figure 6 montre les niveaux les plus généraux de l'ontologie. Ils se divisent en traits décrivant les produits (œuf, viande ou chair, lait), le comportement et le bien-être (*welfare*) et deux grandes fonctions, la nutrition et la reproduction. ATOL est actuellement utilisé par le département Phase, par exemple dans le projet européen AquaExcel (*Aquaculture infrastructures for excellence in European fish research*), où notre approche contribue également à étendre l'ontologie *EOL* (*Environment Ontology for Livestock*) [19]. EOL décrit l'environnement d'élevage.

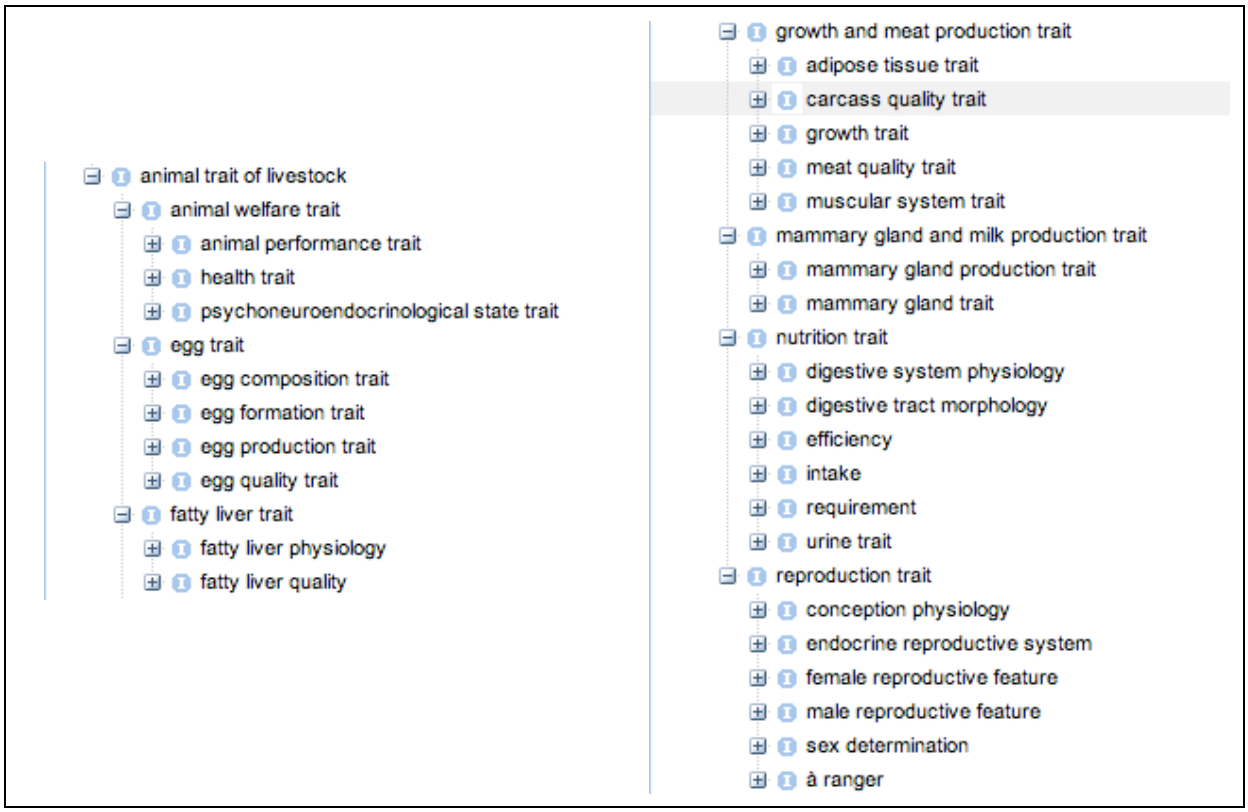


Figure 6. ATOL (Animal Trait Ontology), version du 05/03/2013.

TriPhase, un troisième projet en cours avec TyDI est ambitieux par l'étendue du domaine et le nombre de personnes impliquées. Il vise à concevoir une ontologie de la physiologie animale et des systèmes d'élevage pour l'analyse stratégique du département PHASE, à partir de ses publications. L'ontologie indexera également les publications internationales du domaine dans notre moteur de recherche sémantique pour les chercheurs. Ce projet sous ma direction et celle de Wiktoria implique sept documentalistes du département et des curateurs d'ATOL. Nous utilisons ici à la fois la fonction collaborative de TyDI et la possibilité d'intégrer plusieurs sources de connaissance : principalement des éléments du thesaurus Agrovoc, les ontologies ATOL/EOL et le squelette de l'ontologie cible définit par les documentalistes. Les termes candidats sont extraits des publications des chercheurs du département. J'attends de cette expérience des résultats méthodologiques sur l'organisation du travail, les rôles des acteurs en fonction de leurs compétences, les modalités d'interaction et les différentes étapes de la conception dans une démarche outillée par TyDI. Un des défis du projet particulièrement intéressant est celui de la formation des documentalistes à l'ingénierie des connaissances à partir de leurs compétences en Sciences de l'Information. Leur implication et la pertinence de leur contribution montrent que la démarche est non seulement réaliste, mais généralisable à d'autres documentalistes spécialisés comme ceux du département Phase. Elle sera présentée aux FRéDoc [s210], formation des réseaux de la documentation.

7. La catégorisation sémantique

L'analyse terminologique identifie automatiquement les termes du texte sans produire d'information sémantique. La *catégorisation sémantique* consiste à associer aux termes, des types comme *gène* ou

[14] http://www.atol-ontology.com/index.php/en/

protéine ou plus généralement des concepts d'une ontologie. La catégorisation fait souvent l'objet d'un traitement indépendant de la détection. Deux approches automatiques prévalent pour calculer un appariement entre les termes prédits dans le texte et les labels des concepts de l'ontologie : (1) utiliser les structures internes des labels des concepts de l'ontologie et des termes [Aronson et Lang, 2010], ou (2) utiliser les fréquences des mots des termes et des labels ou de leur contexte dans le corpus [Verspoor et al, 2005 ; Claveau, 2013]. Nous développons des méthodes de catégorisation de termes suivant ces deux axes.

7.1 Catégorisation sémantique par appariement syntaxique

La première méthode *OntoMap* développée avec Pierre Warnier en thèse et Wiktoria Golik, consiste à utiliser la structure morpho-syntaxique des termes et des labels des concepts de l'ontologie pour les apparier. Elle est une extension de la méthode *MetaMap* qui utilise le thésaurus UMLS dans le domaine médical [Aronson et Lang, 2010]. Notre méthode assigne au terme le label de l'ontologie qui partage la même tête syntaxique. Dans le cas où la tête n'est pas informative, c'est la tête du sous-terme qui est traitée. Dans le cas où plusieurs labels partagent la même tête, le label dont les sous-termes sont les plus proches est choisi en fonction d'expressions régulières. Cette méthode permet à l'opposé des méthodes qui comparent simplement les mots, éventuellement ordonnés, de traiter les variations morpho-syntaxiques, en particulier celle des groupes prépositionnels. Par exemple, les termes *milk with chocolate* et *chocolate milk* sont synonymes, mais *milk chocolate* et *chocolate milk* sont différents bien qu'ils partagent les mêmes mots. L'analyse de la structure syntaxique permet de distinguer ces deux cas.
La méthode a été appliquée à la tâche *Bacteria Biotope* de BioNLP-ST'11 [25] pour catégoriser les entités habitat de bactérie avec l'ontologie OntoBiotope [c237] après que BioYaTeA ait identifié les candidats termes et produit leur analyse en constituants. Cette combinaison reconnaissance – catégorisation a obtenu le meilleur score de la compétition. Les détails de la méthode et des résultats pourront être trouvés dans BMC Bioinformatics [283] et [24, 50]. Une variante de la méthode consiste à ne plus considérer la structure dans le cas d'ambiguités, mais seulement des sacs de mots. Elle est évidemment plus efficace quand l'ordre des mots diffère, mais génère des ambiguités. La méthode a été appliquée à la préannotation du corpus Bacteria Biotope de BioNLP-ST'13 [21]. Dans cette expérience l'ontologie OntoBiotope contient plus de 2 000 catégories et les entités du texte à détecter et annoter par les catégories sont au nombre de 507. La catégorisation seule obtient un F-score très élevé de 90%. La détection de termes puis la catégorisation obtiennent un score de 63,3%, significativement plus élevé que celui des participants à cette tâche [21].

7.2 Apprentissage d'ontologie et sémantique distributionnelle

a. *Sémantique distributionnelle*

La méthode *OntoMap* qui est efficace dans le cas où les termes partagent des mots communs, en particulier les mots principaux, ne permet pas d'apparier des termes complètement différents, à l'opposé des méthodes basées sur les similarités de contexte. La deuxième classe de méthodes appartient à la *sémantique distributionnelle* et utilise les contextes d'apparition des termes. J'ai encadré les thèses de David Faure (1996-2000) [268], puis de Pierre Warnier sur différents aspects.
La sémantique distributionnelle consiste en une classification non supervisée de type classification conceptuelle (*conceptual clustering*) de mots ou de termes du texte en langage naturel dans le but de former des classes d'éléments sémantiquement proches. On voit tout l'intérêt pour la catégorisation de termes et la construction d'ontologie, de méthodes capable d'assister la sélection et la structuration des termes du domaine. Selon la théorie de Zeillig Harris sur les langues de spécialité [Harris, 1951], le contenu et la structure des langues de spécialités forment un sous-langage dont les relations entre les éléments peuvent être analysés à partir de leur distribution. Plus particulièrement travaux en sémantique distributionnelle se basent sur sa théorie selon laquelle les mots qui apparaissent dans les mêmes contextes tendent à avoir des significations proches [Harris, 1954]. Deux types de contextes sont étudiés, les contextes syntaxiques et les contextes graphiques. La sémantique distributionnelle appliquée à la simple cooccurrence de termes dans des fenêtres de mots comme entre *pré* et *vache* dans la phrase *La vache broute dans le pré* est préférée pour la classification de documents [Brown, et al., 1992], mais elle produit des proximités sémantiques difficiles à interpréter pour la construction de modèles de connaissance. L'utilisation de la cooccurrence syntaxique, c'est-à-dire de

mots qui sont arguments des mêmes prédicats est plus précise, mais moins productive et surtout, elle est sensible aux erreurs d'analyse syntaxique. Classiquement, de gros corpus sont utilisés pour pallier à la dispersion et aux erreurs des données. Par exemple, à partir de recettes de cuisine, la classe sémantique *{graisse, beurre, matière grasse}* est construite grâce à la cooccurrence de ces termes comme arguments objets des verbes, *faire fondre*, *étaler*, *mélanger, ajouter un peu de*.
Les travaux initiaux de Hirshman et Grishman (78) puis Hindle (92), Grishman et Sterling (94), et Grefenstette (94) ont posé le cadre méthodologique, celui de la classification conceptuelle, et en linguistique computationnelle, celui de l'acquisition de restrictions de sélection pour la désambiguïsation des attachements adjectivaux et verbaux. Ces méthodes produisent des mesures de similarités entre paires de mots, généralement pour former des classes plates et non hiérarchisées.
Au début des années 90, les résultats des méthodes de l'état de l'art étaient limités par plusieurs facteurs :

- les méthodes de calcul de similarité utilisées étaient souvent élémentaires pour des raisons de complexité, par exemple, le cosinus;
- les exemples d'apprentissage étaient bruités par des erreurs d'analyse syntaxique ;
- et les inductions étaient erronées, du fait de la polysémie des dépendances syntaxiques. Par exemple, à partir des exemples (*conduire un(e)* {*train*, *voiture*, *tracteur*}) (*voyager en* {*voiture, train / été, juin*}) la sémantique distributionnelle produit des similarités intéressantes entre les véhicules {*train*, *voiture*, *tracteur*} grâce à l'intersection importante {*train*, *voiture*} commune aux deux classes d'arguments, mais également des similarités entre véhicules et arguments temporels {*été*, *juin*} qui ne sont pas distinguables ici sans autres exemples.
- La formalisation de classes sémantiques sous forme de concepts d'une ontologie requiert plusieurs étapes complémentaires, la sélection de termes dénotant des concepts parmi les termes extraits, la qualification de la relation sémantique entre les termes à partir de leurs similarités et une structuration hiérarchique plus précise que l'inclusion entre classes.

David Faure et moi avons travaillé sur ces questions à partir de 1996 financés par le projet européen Esprit IV "Long Term Research" ILP 2 (*Inductive Logic Programming*). Nos résultats avec les analyseurs *Shallow Parser* du *Xerox Research Center* à Grenoble pour l'anglais [Aït-Mokhtar et Chanod, 97], et *Sylex* [Constant, 1991] pour le français étaient très prometteurs. Nous avons fait le pari à moyen terme, gagné depuis, que les performance de l'analyse syntaxique pour des textes scientifiques et techniques progresseraient rapidement. L'apprentissage coopératif apparaissait comme une réponse particulièrement adaptée et originale pour corriger les classes induites à partir des similarités. Nous avons conçu la méthode Asium, qui prend en entrée l'analyse en dépendances syntaxiques du texte. Elle est basée sur un algorithme de classification hiérarchique ascendante. Il construit une hiérarchie de classes sémantiques. Un seuil appliqué aux distances entre termes et classes, détermine les frontières des classes. Les détails de la méthode et des exemples pourront être trouvés dans la thèse de David [Faure, 2000]. La méthode se distinguait de l'état de l'art par quatre points :

(1) L'interaction utilisateur-système permet de réviser selon des modalités variées, les classes sémantiques apprises, au fur et à mesure de leur formation, ce qui évite la propagation des erreurs,
(2) La méthode génère un graphe acyclique plutôt qu'une hiérarchie stricte, capable de rendre compte de différents rôles sémantiques pour un même terme,
(3) La distance entre mots est à la fois intuitive et originale, le seuil des classes est ajustable en cours d'apprentissage,
(4) La méthode construit des classes d'arguments, au lieu de classes de prédicats, ce qui accélère considérablement l'apprentissage et permet de construire des classes plus pertinentes et plus larges que des classes binaires.

David Faure a implémenté l'algorithme d'Asium, il a été déposé par le CNRS à l'agence de protection des programmes (APP). Différents aspects de la méthode et des résultats expérimentaux ont été publiés en acquisition des connaissances et fouille de texte à EKAW [13] et dans les conférences TALN et JFA [58, 59] et dans des workshops ECML [34] et LREC [35]. Ils ont été présentés dans des séminaires invités [s90, s97, s98, s100, s110, s189].

Le logiciel Asium a été évalué principalement en interne dans plusieurs tâches applicatives. Nous avons montré un gain considérable de 30 % pour la désambiguïsation des attachements syntaxiques sur un corpus de référence constitué de recettes de cuisine, étiqueté manuellement. L'acquisition de hiérarchies conceptuelles a été évaluée dans ce même domaine et dans le domaine des brevets sur les oxy-brûleurs [s99] avec la société Air Liquide en collaboration avec Eunika Mercier-Laurent (EML Conseil).
Nous avons formalisé la méthode d'Asium dans un cadre de PLI pour expliciter l'ensemble des opérations [33, 72, r172]. Pour améliorer sa reproductibilité, sa réutilisation et surtout la comparaison et l'expérimentation, nous avons développé l'outil *Mo'K* avec Gilles Bisson (INRIA, projet Helix). De façon similaire à Haïku, nous avons défini un cadre à la fois formel et opérationnel sous forme d'opérateurs configurables, qui permet la comparaison des différentes méthodes de sémantique distributionnelle avec des configurations différentes des exemples, des distances ou des opérateurs de fusion de classes. Nous avons également défini un ensemble de mesures d'évaluation des classes, spécifiques à ce domaine basées sur la mesure de la couverture du corpus par les classes. L'évaluation des résultats était manuelle, il n'existait pas de cadre d'évaluation quantitatif général ni pour la classification conceptuelle, ni pour la sémantique distributionnelle. La comparaison avec une ontologie existante est souvent non pertinente dans les domaines spécifiques considérés. Grâce à l'implémentation de Mo'K par Gilles, nous avons montré la pertinence des nouvelles mesures d'évaluation et la supériorité de la distance proposée dans Asium par rapport aux méthodes de l'état de l'art, en particulier celles de Hindle (91), Dagan et al. (99) et Grishman et Sterling (94), ceci sur *Amaryllis*, un corpus public de l'INIST [32, 56].
Les résultats d'Asium forment des cadres de sous-catégorisation dont les restrictions de sélection sont les classes sémantiques apprises. En ce sens, ils constituent une première étape dans l'apprentissage de cadres prédicatifs (*predicate-argument structure*) purement sémantiques, pour exprimer des règles d'extraction d'information de relations spécialisées. Cela fit l'objet d'un article prospectif dans *IEEE Intelligent Systems* en 2002 [178] et de séminaires [s80, s94, s99].

b. *Sémantique distributionnelle et ontologie*

J'ai repris ces travaux en 2010 avec la thèse de Pierre Warnier en cours, où nous évaluons l'intérêt d'une méthode de *co-clustering* pour la sémantique distributionnelle capable d'exploiter une ontologie comme connaissance *a priori* pour guider l'acquisition de classes sémantiques et enrichir l'ontologie. Pierre adapte la méthode de calcul de similarité *X-Sim* développée par Gilles Bisson son co-encadrant, et Clément Grimal (2012). La matrice de similarité est initialisée avec des mesures de similarité entre les nœuds de l'ontologie calculée par la distance de Wang (2007). Nous visons le traitement de gros corpus tel que les pages de *Wikipedia*. La grande taille et la dispersion des données posent un problème de complexité algorithmique qui est traité par une méthode de réduction de dimension avec perte (*random projection*) [Bingham et Mannila, 2001]. Les premières expériences portaient sur le corpus Bacteria Biotope [c235] et l'ontologie OntoBiotope. Le corpus contient 600 000 références de la base bibliographique PubMed, analysés à l'aide de CCG et d'AlvisGrammar décrits ci-dessous. Les premiers résultats décevants ont confirmé l'importance de la prise en compte des termes composés plutôt que des mots simples, en raison de la fréquence élevée de têtes identiques dans les labels de l'ontologie. L'analyse des dépendances syntaxique portant sur les termes devraient produire de meilleurs résultats.
Parallèlement, une évaluation comparative avec différentes méthodes et mesures de similarité, cosinus, LSA, et Random Projection pour la sémantique distributionnelle est réalisée sur un sous-ensemble de Wikipédia. Nous évaluons la capacité des méthodes à retrouver des couples de nominalisation verbe – nom appartenant au lexique NOMLEX (*NOMinalization Lexicon*)[15], grâce à la similarité de leurs prédicats et de leurs arguments dans le corpus. Par exemple, dans le cas du couple (*to express*, *expression*), certains objets du verbe *to express* et et certains compléments en *of* de *expression* devraient être identiques et la similarité entre *to express* et *expression* d'autant plus élevée. Ces expériences permettent en particulier d'explorer les seuillages optimaux de la sélection d'attributs pour limiter la dispersion élevée des données. La meilleure configuration mesurée sur un appariement

[15] http://nlp.cs.nyu.edu/meyers/nombank/nombank.1.0/NOMLEX-plus-clean.1.0

exact sans interprétation manuelle produit 40% des appariements espérés, ce qui est très prometteur par rapport aux travaux précédents [Gabor et al., 2012].

8. Analyse syntaxiques

De nombreuses tâches d'acquisition de connaissances, requièrent une analyse syntaxique en dépendances des textes. Nous avons vu plus haut les exemples de la sémantique distributionnelle (§7) et de l'extraction d'information relationnelle (§1).
Les analyses en dépendances sont plus aisées à exploiter pour des algorithmes d'apprentissage automatique relationnel ou distributionnel que les analyses en constituants du fait de l'homogénéité des représentations entre l'analyse et l'algorithme d'apprentissage.

8.1 Comparaison expérimentale

Durant la dernière décennie, la qualité et le nombre d'analyseurs syntaxiques disponibles ont évolué très vite. Notre choix des analyseurs syntaxiques s'est fait sur la base de plusieurs études comparatives au fur et à mesure des nouveaux développements. En 2002 l'évaluation comparative des performances de *Link Parser* [Sleator et Temperley, 1993] avec d'autres analyseurs de l'état de l'art, comme *Minipar*[16] lors du stage de DEA de Sophie Aubin a en particulier porté sur les relations verbales et les formes passives qui sont importantes dans l'extraction d'interactions géniques [r163, r264, 280]. En 2005, l'équipe a corrigé manuellement l'analyse par LinkParser du corpus du challenge LLL (*Learning Language in Logics*) composé de phrases d'interactions géniques [c242]. Ces analyses de référence constituent encore aujourd'hui un ensemble adapté à l'évaluation des analyseurs représentatif de nos besoin en biologie, bien que limité en taille. Le jeu de données est accessible en ligne et le détail de sa construction est décrit dans [40].
En 2011, la comparaison quantitative détaillée de plusieurs analyseurs de l'état de l'art conduite par Zorana Ratkovic (thèse co-encadrée avec Thierry Poibeau du LaTTICe / ENS-CNRS) a utilisé ce corpus LLL. Nous avons conclu à la supériorité de l'analyseur *HPSG* (*Head-Driven Phrase Structure Grammar*) *Enju* de *Tsujii Laboratory* [Miyao and Tsujii, 2008] et de *CCG* (*Combinatory categorial grammar*) de *Cambridge University* [Clark et Curran, 2017] sur *BioLG*, l'adaptation de *Link Parser* à la biologie. Zorana a en particulier examiné le traitement de phénomènes linguistiques très fréquents dans nos corpus que sont les formes passives, les coordinations, les appositions et les énumérations. Une évaluation comparative de l'impact de ces analyseurs dans des tâches d'extraction d'information relationnelle à partir de corpus est en cours.

8.2 Adaptation des analyseurs au domaine de la biologie

Les analyseurs généralistes ont montré leurs limites dans le traitement de corpus spécialisés. Pour adapter automatiquement un analyseur syntaxique au domaine de la biologie pour des tâches d'acquisition de connaissance, nous étudions deux voies qui ne requièrent pas de coûteux corpus annotés : la simplification des analyses en ne conservant dans une analyse donnée, que les dépendances syntaxiques pertinentes et l'utilisation de connaissances du domaine pour ordonner les analyses concurrentes alternatives.

a. *Grammaire simplifiée*

La première voie porte sur la simplification de la grammaire. Dans la perspective de l'extraction d'information dans des domaines scientifiques, il est raisonnable de faire l'hypothèse que certaines distinctions entre des étiquettes de dépendances et certaines dépendances ne sont pas critiques, par exemple, la dépendance entre le déterminant et le nom, ou entre la préposition et le déterminant. Inversement, la richesse de la grammaire augmente considérablement le nombre d'analyses générées et leur complexité, en rendant pour l'apprentissage, la recherche de régularités entre les exemples d'apprentissage plus difficile. Sophie Aubin avec le RCLN et moi avons développé *AlvisGrammar* en 2005, un nouveau jeu d'étiquettes pour les dépendances syntaxique, qui répond aux besoins en EI [r282]. Nous avons fait *a priori*, le choix des 17 dépendances les plus pertinentes, dépendances verbales, attachement prépositionnels, *etc.,* en éliminant les dépendances potentiellement inutiles, pour généraliser et réduire le jeu d'étiquettes des grammaires existantes. Les étiquettes comportent les informations sur le type de dépendance, suivi éventuellement de la préposition, puis les catégories

[16] http://webdocs.cs.ualberta.ca/~lindek/minipar.htm

grammaticales des deux arguments. Par exemple, *COMP(by):V PASS-N* désigne la relation entre le verbe au passif et le nom qui est la tête du syntagme agent introduit par la préposition *by*, comme *COMP(by):V PASS-N (yfhS, E sigma E)* dépendance analysée dans la phrase, yfhS *is transcribed by* E sigma E.
Les analyses de référence des données LLL sont représentées dans cette grammaire. De façon plus générale, *AlvisGrammar* est la grammaire que nous exploitons pour nos recherches en sémantique distributionnelle et en extraction de relations. Un premier traducteur de Link Parser en AlvisGrammar, *LP2LP*, avait été développé par Erick Alphonse.
Zorana Ratkovic en thèse l'a étendu aux trois analyseurs, Enju, CCG et BioLG. Grâce aux nouveaux traducteurs qu'elle a implémentés, elle évalue l'intérêt d'AlvisGrammar par rapport aux sorties brutes des analyseurs pour l'extraction d'information de relation. La méthode d'extraction est basée sur un *Shortest Path Global Alignment Kernel* développé par Dialekti Valsamou en thèse sous ma direction. La méthode consiste à représenter les exemples sous la forme du chemin syntaxique entre les arguments à classer comme étant ou non en relation. Les premiers résultats avec les données LLL montrent une amélioration notable de la qualité des extractions de 6 points, entre l'utilisation des sorties directes des analyseurs et leur traduction en AlvisGrammar. L'adaptation de la tokénisation et de l'étiquetage morpho-syntaxique à la biologie améliorent encore ces résultats. Ils seront soumis à EACL cette année. Ce résultat devra être amélioré en traitant les cas malheureusement fréquents où aucun chemin syntaxique n'est trouvé entre les arguments candidats, le classifieur donnant la classe par défaut. Ce cas se produit quand l'analyseur n'a pas trouvé d'analyse, mais aussi quand la réécriture en AlvisGrammar coupe les chemins. Un premier examen des données montrent que les chemins ainsi perdus résultent d'analyses erronnées. Une analyse plus approfondie devrait permettre d'améliorer AlvisGrammar et de réduire le nombre de chemins perdus.

b. *Utilisation d'information lexicale*

Une deuxième voie d'adaptation de l'analyse syntaxique à la biologie vise à ordonner les analyses alternatives par des connaissances lexicales spécifiques au domaine. Plus précisémenent, la méthode exploite les termes du domaine pour ne retenir que les analyses qui sont cohérentes avec les termes. Par exemple, si le terme *interaction with the control region* appartient au lexique, pour l'extrait,

by preventing sigmaD-driven transcription of degR through interaction with the control region.

l'analyse où *with the control region* est attachée à *interaction* sera sélectionnée parce qu'elle est cohrente avec le terme, et les analyses où *with the control region* est attachée à *transcription* ou à *preventing* seront supprimées. Suivant ce principe, *BioLG* une première adaptation de Link Parser à la biologie a été conçue et implémentée. Les détails de la méthode et des résultats seront trouvés dans [11] publiée à RANLP. Sampo Pyysalo (Université de Turku) l'a ensuite étendue [Pyysalo et al., 2006].
Zorana Ratkovic a repris ce concept pour AlvisGrammar en l'étendant. L'appariement des termes certifiés et des syntagmes du corpus analysé n'est pas toujours direct. Il faut fréquemment prendre en compte les attachements d'arguments supplémentaires par rapport aux termes du lexique. Par exemple, dans les phrases de l'exemple (1) ci-dessous, le terme du lexique est *sea sand*, l'attachement de l'adjectif supplémentaire *black* à *sand* puis de *sand* à *isolate from* ne pose pas de problème, une seule analyse est possible ici.

(1) *an agarolytic bacterium isolated from sea sand*
an agarolytic bacterium from black sea sand on Jeju Island

(2) *isolated from the Korean tidal flat sediment*
isolated from tidal flat sediments in Korea

Exemples du corpus de Bacteria Biotope.

Dans l'exemple (2), le terme du lexique est *tidal flat sediment*. Dans la première phrase, l'attachement de l'adjectif *Korean* à *tidal flat sediment* n'est pas ambigü. Par contre dans la deuxième phrase, le groupe prépositionnel *in Korea* peut être soit attaché au verbe *isolated* ou bien au terme *tidal flat*

sediment. Sans autre information grammaticale ou terminologique, les deux analyses sont proposées par l'analyseur syntaxique.
Pour ordonner les analyses les plus probables pour un domaine donné, Zorana et moi définissons une typologie des variations et ajouts dans les phrases par rapport à un ensemble de termes certifiés. Pour cette étude, nous avons construit un corpus spécialisé de paraphrases dans le domaine des biotopes bactériens. Les paraphrases ont été alignées automatiquement en sélectionnant des phrases du journal JSEM (*International Journal of Systematic and Evolutionary Microbiology*) et des entrées du champ *isolation source* de GeneBank portant sur les mêmes bactéries et qui contiennent les mêmes termes. Notre hypothèse est que si nous pouvons automatiquement construire un alignement grammatical et lexical, il pourra être utilisé pour sélectionner des analyses syntaxiques correctes et les ordonner, mais aussi pour l'étiquetage sémantique à l'aide d'ontologie. En effet, dans de nombreux cas, les fragments à traiter ne résultent pas de simples variations morpho-syntaxiques. Outre l'ajout de précisions comme *in Korea*, nous avons identifié plusieurs phénomènes linguistiques combinés comme l'hyperonymie. Par exemple, dans

1. *isolated from a cow jaw*
2. *discovered from human and animal sources*

cow est généralisé en *animal*. Ce travail en cours ouvre des perspectives très intéressantes sur la normalisation à l'aide d'ontologie dans l'annotation sémantique de textes.

9. Extraction d'information relationnelle.

9.1 Introduction

L'identification de relations sémantiques entre les entités du texte est une tâche essentielle de l'extraction d'information. L'extraction de relations n-aires, ou d'événements à partir de texte a été formalisée en particulier grâce aux compétitions MUC (*Message Understanding Conferences*). Un large ensemble de travaux portant sur l'extraction d'information à partir de dépêches ou de textes de journaux a été publié dans les années 90. Pendant longtemps, les approches à base d'apprentissage automatique se sont opposées aux approches à base de patrons. Les premières obtiennent de bonnes performances de rappel, mais nécessitent des exemples annotés. Les secondes sont plus rapides à mettre en œuvre, avec de bonnes performances en précision, mais moins générales et adaptables. Au début des années 2000 l'application de l'extraction d'informations relationnelles à des domaines spécialisés, en particulier dans le domaine biomédical a mis en évidence, d'une part la nécessité d'une analyse linguistique profonde et d'autre part l'intérêt de l'apprentissage automatique quand les tâches d'extraction se multiplient et que les experts du domaine sont rares [179]. L'apprentissage automatique est devenu l'approche dominante aujourd'hui en extraction d'information relationnelle à partir de textes pour la biologie.
La question de l'extraction de relations est formulée comme un problème d'apprentissage artificiel supervisé. Elle consiste à apprendre automatiquement un classifieur à partir d'exemples de textes où les relations sont annotées manuellement. Le classifieur appliqué à de nouveaux textes prédit pour chaque ensemble d'arguments candidats préalablement identifiés, s'ils sont effectivement en relation. L'exemple ci-dessous tiré des données LLLL, porte sur l'extraction de relation d'interactions géniques entre des protéines d'une part qui sont les agents de l'interaction et des gènes qui sont les cibles, d'autre part. La relation dans l'exemple est binaire et orientée.

Dans la phrase,
GerE *stimulates* ***cotD*** *transcription and inhibits* ***cotA*** *transcription in vitro by* ***sigma K*** *RNA polymerase, as expected from in vivo studies, and, unexpectedly, profoundly inhibits in vitro transcription of the gene (****sigK****) that encode* ***sigma K****.*

6 gènes et protéines sont mentionnés formant 30 couples ordonnés. En fait seuls 5 d'entre eux sont positifs : (GerE, cotD), (GerE, cotA), (sigma K, cotA), (GerE, SigK) et (sigK, sigma K), c'est-à-dire que dans ces couples le premier membre de la relation est une protéine qui active le deuxième membre de la relation qui est un gène. Les 25 autres couples sont des exemples négatifs de la relation.

Exemple**.** Information d'interaction génique à extraire en biologie.

La variabilité des formulations textuelles pour exprimer une même connaissance impose une étape d'analyse linguistique qui associe aux exemples d'apprentissage issus du texte, un ensemble d'informations linguistiques qui mettent en évidence des régularités, implicites dans le texte lui-même [179]. Dans l'exemple, l'agent de l'interaction génique *GerE* est loin de la cible *sigK*, il est le sujet du verbe *inhibits transcription of the gene (**sigK**)*, mais il n'est pas repris dans cette deuxième proposition. Sans l'information de l'analyse syntaxique qui relie le verbe *inhibits* à son sujet *GerE*, l'extraction automatique est difficile.

Les premières approches à base de patrons sélectionnaient comme positives, les paires d'arguments qui encadraient le verbe (ou sa nominalisation) qui portait l'action d'interaction [Ono et al., 2001]. Tant le rappel que la précision étaient médiocres. De nombreuses relations d'interaction ne sont pas signifiées par des actions, comme l'appartenance à un régulon et inversement, des arguments candidats peuvent être en positions de sujet ou d'objet, encadrant le verbe, sans être en relation et inversement. L'analyse syntaxique permet d'identifier les rôles des arguments par rapport au verbe. Nous avons contribué avec le challenge LLL et ses analyses de références à populariser l'utilisation de l'information syntaxique, devenue aujourd'hui majoritaire en biologie. Par exemple, dans la tâche GRN (*Gene Regulation Network*) de la compétition BioNLP-ST en 2013, tous les participants utilisent cette information [20].

Plus généralement, la recherche de l'algorithme d'apprentissage le plus performant pour l'extraction d'information est devenu aujourd'hui indissociable de la recherche sur la représentation des exemples la plus appropriée. La question complexe de l'appariement de la représentation linguistique et de la représentation des données d'apprentissage n'est pas traitée de manière approfondie en raison de la nouveauté de la disponibilité d'analyses linguistiques de qualité et d'algorithmes d'apprentissage performants pour des représentations structurées comme les arbres et les graphes. Les informations produites par l'analyse linguistique appartiennent à différents niveaux interdépendants de normalisation et d'abstraction : lemmes, étiquettes morpho-syntaxiques, dépendances syntaxiques, termes et variations, coréférences, classes sémantiques, concepts, rôles sémantiques.

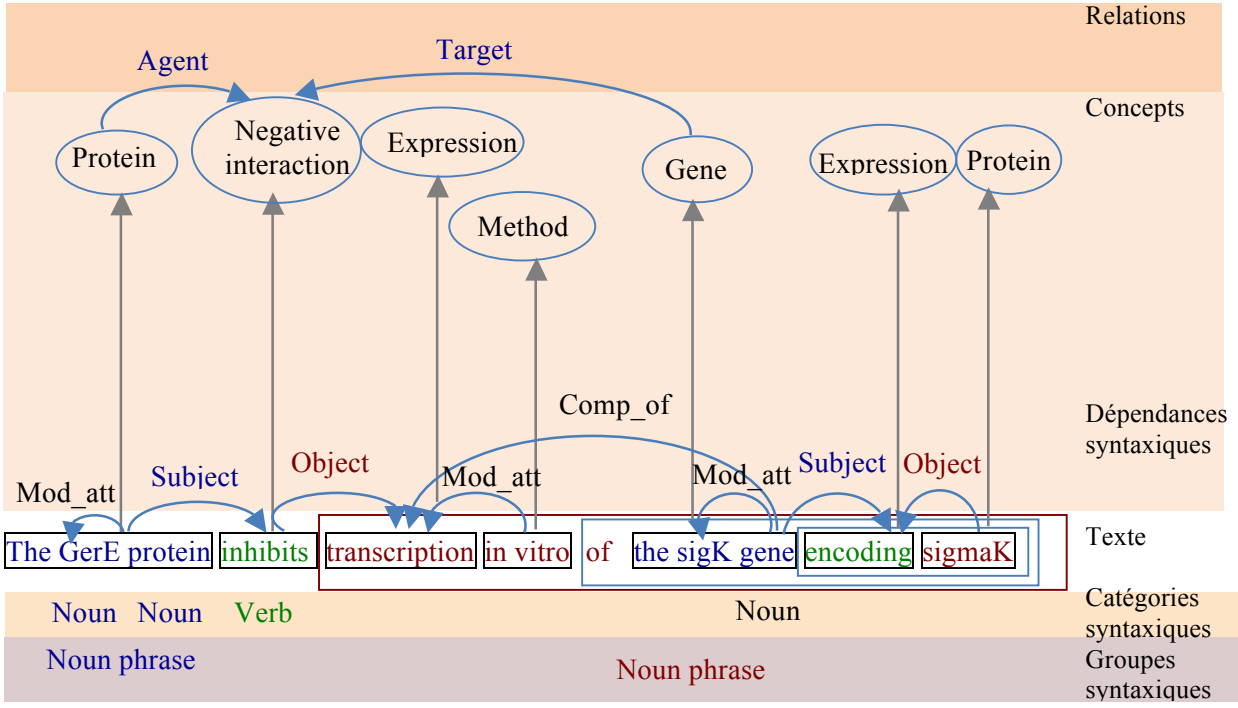


Figure 7. Représentation graphique d'un exemple d'analyse syntaxique et sémantique.

La figure 7 montre un exemple de l'analyse de la phrase *the GerE protein inhibits transcription in vitro of the sigK gene encoding sigmaK* extraite des données LLL. La table 1 détaille les différents niveaux d'analyse.

Catégorie sémantique	is_a(GerE protein, protein), is_a(inibit, negative_interaction), …
Dépendance syntaxique	sujet(GerE protein, inhibit), obj(transcription, inhibit), …
Terme	terme(GerE protein), terme(in vitro) terme(sigK gene)
Catégorie syntaxique	cat(the, det), cat(Ger_protein, term), cat(inhibit, verb), …
Entité nommée	entité(GerE), entité(sigK), entité(sigma K)
Texte segmenté	mot(the), mot(Ger_protein), mot(inhibit), mot(transcription), …

Table 1. Exemple d'analyse syntaxique et sémantique.

La question est alors, étant donné la richesse de l'information linguistique, de proposer une représentation des exemples et une technique d'apprentissage aptes à en tirer parti, tout en conservant des propriétés calculatoires raisonnables.

9.2 Chemins syntaxiques pour l'extraction de relations

Deux types de connaissances linguistiques sont reconnues comme particulièrement critiques, les dépendances syntaxiques comme marqueur de rôles sémantiques et donc des relations cibles, et les classes sémantiques permettant de généraliser les descriptions des exemples. Par exemple, comme mentionné ci-dessus, la protéine, sujet d'un verbe d'interaction dont l'objet direct est un gène, est probablement l'agent de l'interaction dont le gène objet est la cible. Diverses représentations des informations syntaxiques ont été étudiées : sous forme de séquences (Culotta *et al.*, *ACL* 2006 ; Bunescu & Mooney, *NIPS* 2006), de *shallow parsing* (Pustejovsky, *PSB* 2002 ; Zelenko, *JMLR* 2003), d'arbres (Zhang, *et al* ; ACL 2006 ; Liu *et al*, *NAACL* 2007) ou de graphes (Culotta & Sorensen, *ACL* 2004 ; Fundel, *Bioinformatics* 2006).

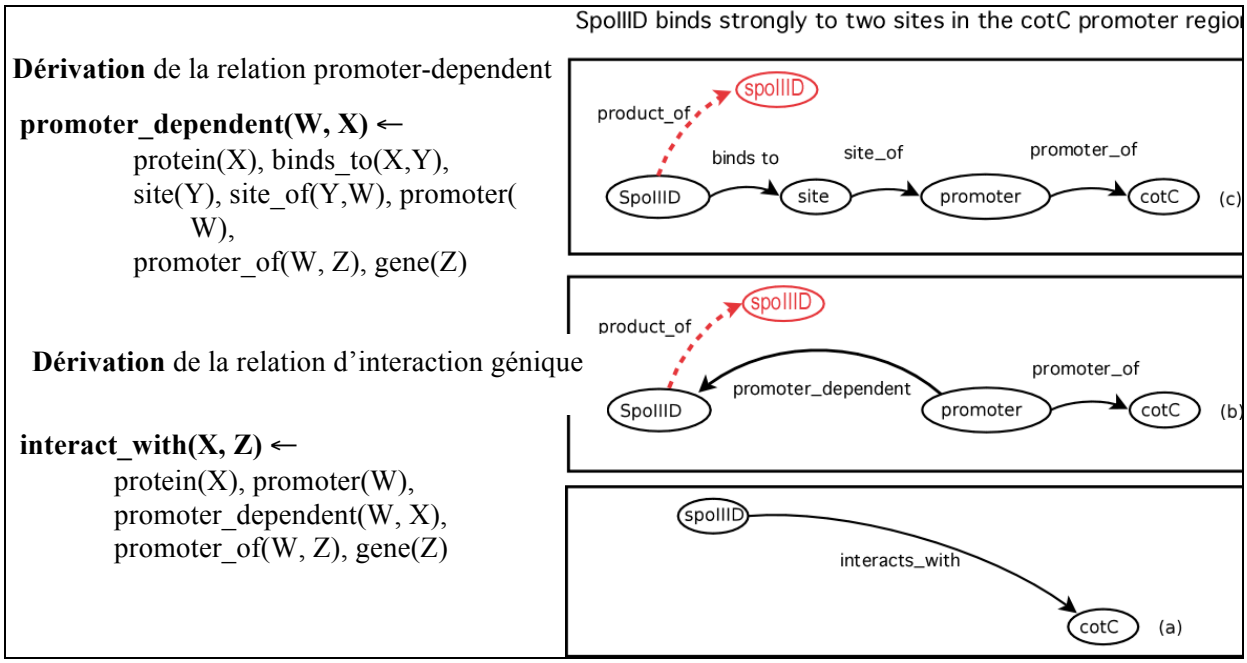


Figure 8. Dérivation d'une relation d'interaction à partir de la liaison de la protéine sur le site promoteur du gène.

Nous travaillons depuis le début des années 2000 sur cette question en particulier grâce aux thèses de Mohammed Ould Abdel Vetah et d'Alain-Pierre Manine et aux travaux d'Erick Alphonse en post-doctorat. Nous cherchions à exploiter directement la représentation des dépendances syntaxiques, sous forme de relations dans un algorithme d'apprentissage relationnel [51, 52]. L'approche que nous avons choisie initialement en Programmation Logique Inductive (PLI) consistait à représenter les exemples d'apprentissage par le chemin syntaxique entre les arguments potentiels de la relation (gènes et protéines dans le cas de l'interaction génique) de façon similaire à [Daraselia et al., 2004]. Cette idée novatrice est maintenant largement utilisée en EI. Si l'exploitation des informations linguistiques sémantiques permet l'apprentissage de règles d'extraction puissantes et compréhensibles, elle est aussi

sensible aux erreurs d'analyse, en particulier des dépendances syntaxiques. Pour limiter les effets des erreurs d'analyse syntaxique, la méthode élaguait les syntagmes qui ne sont pas dit "d'intérêt". Par exemple dans,

> *sigmaK-dependent transcription of gerE initiates a negative feedback loop in which GerE acts as a repressor to limit production of sigmaK*.

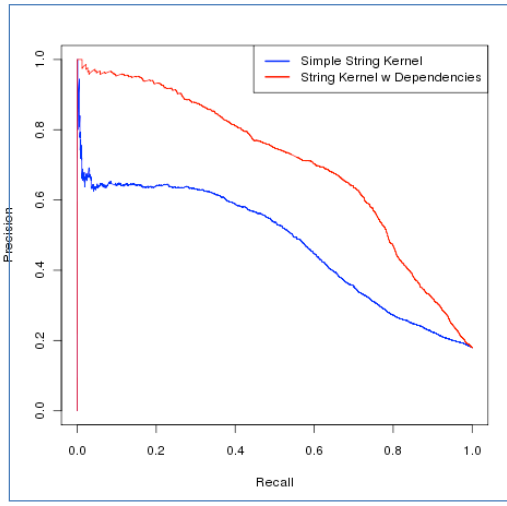
le chemin entre *gerE* et *sigmaK* inclut *a negative feedback loop in which* que la méthode supprime. Cela suppose que soit défini déclarativement en amont l'ensemble des termes et relations d'intérêt. La méthode d'apprentissage sous-jacente était LP-Propal développée par Erick pendant sa thèse. Les résultats de la thèse d'Alain-Pierre en 2006 [270] montraient la supériorité de l'approche par rapport à l'état de l'art mesurés sur la tâche LLL [r166, r167, r168.

Les travaux d'Erick et d'Alain-Pierre en post-doctorat se sont poursuivis au LIPN. Ils ont en particulier montré l'intérêt de la PLI pour apprendre un programme plutôt qu'une relation simple. Par exemple, à partir de la liaison d'une protéine avec le promoter d'un gène, et de l'appartenance du promoteur au gène, on peut déduire une interaction directe de la protéine avec le gène comme représenté dans la Figure 8. Ces résultats ont été publiés à ICTAI [Manine et al., 2008a] et dans *Int. J. Med. Inform.* [Manine et al., 2008b] en 2008.

9.3 Approche à base de noyau

Inspirée par ces travaux, la représentation biologique des données LLL a été étendue sous forme d'événements complexes décrivant les rôles des différents objets et événements biologiques. Nous avons utilisé les données LLL réannotées dans cette représentation pour organiser deux tâches d'extractions d'information géniques *Bacteria Interaction* dans BioNLP-ST'11 [284, 2] et *Gene Regulation Network* dans BioNLP-ST'13 [20]. Le détail de la représentation est sur le site Web[17].

Pour traiter cette représentation biologique complexe, nous avons adopté une approche différente de la PLI à base de noyau (*kernel*), moins sensible aux erreurs dans les exemples et maintenant largement utilisée par l'état de l'art [Björne et al., 2012]. Ces travaux ont été réalisés par Philippe Veber, IR de l'équipe, puis Dialekti Valsamou en thèse co-encadrée avec Pierre Zweigenbaum du LIMSI. Comme précédemment, les exemples sont représentés par la séquence des mots et des dépendances sur le chemin syntaxique. Les expériences menées sans dépendances montrent tout leur intérêt (Figure 9).



R	P	F
76,9 (±2,8)	55 (±2,1)	64,4 (±1,8)

Figure 9. Performance de l'extraction de relations de LLL avec un *string kernel*.

La figure 10 montre un exemple de la représentation syntaxique pour la méthode d'apprentissage. La phrase *A low level of transcription of GerE activates transcription of cotD RNA polymerase in vitro*, est segmentée, puis lemmatisée et enfin analysée dans la représentation AlvisGrammar. Les arguments sont remplacés par *Agent* et *Target* pour éviter la surspécialisation de l'apprentissage.

Les dépendances syntaxiques sont figurées en bleu. La catégorie syntaxique est indiquée sous les mots de la phrase. Le chemin entre les deux arguments de la relation cible, ici l'interaction génique, est

[17] https://sites.google.com/site/bionlpst2013/tasks/gene-regulation-network

représenté dans le deuxième vecteur. Il relie *Agent* et *Target* à travers quatre dépendances successives de complément de nom, de sujet, d'objet et de complément de nom. Contrairement à la représentation des exemples en PLI, les exemples sont ramenés à une séquence ordonnée qui limite arbitrairement le nombre d'appariement entre les éléments de la séquence.

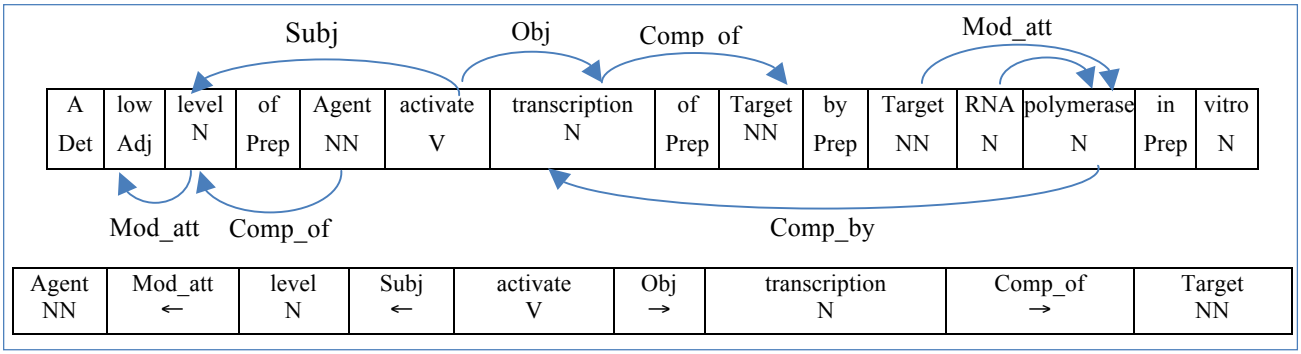


Figure 10. Représentation vectorielle du chemin syntaxique.

La méthode à base noyau *Shortest Paths Global Alignment Kernel* (SPGAK) calcule un alignement entre les paires d'exemples qui minimise la distance globale entre l'exemple à prédire et les exemples de l'ensemble d'apprentissage. La classe prédite est la classe des exemples d'apprentissage les plus proches. La figure 11 montre un exemple de l'alignement entre deux exemples.

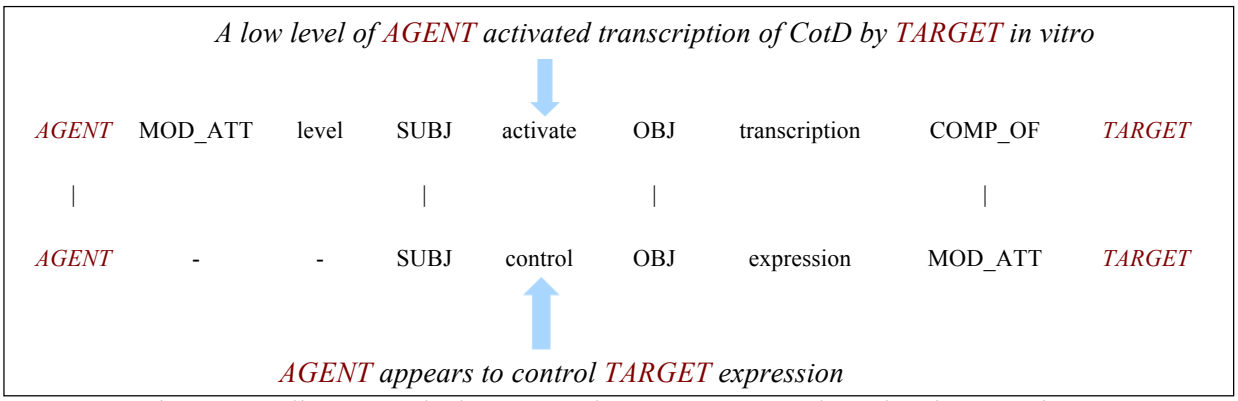


Figure 11. Alignement de deux exemples représentés par leur chemin syntaxique.

Pour que la mention de la quantité (*low level*) de la protéine du premier exemple ne pénalise pas trop l'appariement, il faut que la pénalité du *gap* ne soit pas trop élevée. Inversement, la pénalité doit être modérée si les deux éléments sont de même catégorie morpho-syntaxique et élevée dans les autres cas. Les évaluations de Philippe puis de Dialekti sur les *benchmarks* LLL et BI, sont très prometteuses et au niveau de l'état de l'art (Table 2).

Global Alignment Kernel		Analyse syntaxique		
		Sans	Syntaxe automatique	Syntaxe manuelle
Cl. sémantiques	sans	-	61.0 ± 4.1	77.0 ± 2.4
	avec	-	59.4 ± 5.4	**79.1 ± 2.8**

Table 2. Résultats expérimentaux avec la méthode SPGAK sur les données LLL.

String Kernel		Analyse syntaxique		
		Sans	Syntaxe automatique	Syntaxe manuelle
Cl. sémantiques	sans	52.2 ± 3.1	64.4 ± 1.8	69.0 ± 2.3
	avec	52.4 ± 3.7	68.4 ± 2.3	75.4 ± 2.6

Table 3. Résultats de *String kernel* sur les données LLL.

Elles montrent en particulier l'importance de la qualité de l'analyse linguistique en amont, tokenisation, lemmatisation et analyse syntaxique. Le travail en cours mesure en particulier l'impact sur la qualité de l'extraction de l'utilisation de termes composés en lieu et place des mots simples. Les résultats obtenus avec un *string kernel* sont inférieurs (Table 3).

9.3 Classes sémantiques pour l'extraction de relations

La méthode SPGAK inclut l'exploitation de similarités sémantiques pour prendre en compte des distances entre mots pour le calcul de similarité entre les exemples qui ne soient pas strictement booléennes, de façon similaire à [Plank et Moschitti, 2013]. Dans l'exemple, sans mesure de similarité sémantique entre *activate* et *control* et entre *expression* et *transcription* les mots sont considérés comme aussi différent que *appears* et *level*. Plusieurs voies ont été explorées pour utiliser une information de similarité sémantique. Nous avons montré avec Philippe sur les données LLL que des classes sémantiques conçues manuellement augmentaient les performances de classification. Ces résultats ont été consolidés ensuite par Zorana et Dialekti.

Des expériences comparatives avec des ressources non adaptées à la biologie améliorent les résultats de quelques points. Elles utilisent une similarité de chaînes de caractères, une similarité basée sur le thesaurus WordNet WS4J (*WordNet Similarity for Java*) et enfin la sémantique distributionnelle par simple coocurrence dans une fenêtre de mots sur PubMed.

Pour adapter la similarité à la biologie, les travaux en cours en collaboration entre Pierre Warnier et Dialekti Valsamou étudient l'apport de la similarité distributionnelle calculée sur un corpus du même domaine, en prenant en compte le contexte syntaxique de la classe. Notre hypothèse est que les voisins sémantiques différent selon le rôle syntaxique. C'est le cas en particulier des termes polysémiques, et en biologie, les métonymies sont fréquentes. Des améliorations évidentes de la représentation devraient également porter sur la généralisation des étiquettes des dépendances syntaxiques.

10. Recherche documentaire.

Avec l'extraction d'information détaillée ci-dessus, la recherche d'information est une des grandes classes d'application textuelle. C'est celle choisie dans le projet Alvis (*semantic search engine*) en 2005 pour démontrer les bénéfices d'une annotation sémantique fine pour la recherche bibliographique. Le principe consiste à annoter la collection de documents puis indexer avec les unités sémantiques normalisées (termes et entités nommées), puis par les concepts de l'ontologie du domaine structurés en hiérarchie de généralisation, les concepts sont désignés par les termes canoniques. Ensuite, l'utilisateur formule des requêtes sous forme de combinaisons de concepts de l'ontologie, au niveau de généralité souhaité. Par exemple, la requête *disease resistance* renvoie tous les documents qui mentionnent la résistance à différentes maladies comme la fusariose, sans qu'il soit nécessaire que le terme de la requête *pest resistance* soit présent dans le document (Figure 12).

L'annotation sémantique du texte, structurée et normalisée, apporte une puissance d'interrogation considérable par rapport à une approche *à la Google* qui recherche les mots exacts ou approchés de la requête. Elle est particulièrement adaptée aux domaines spécialisés comme la biologie, où une termino-ontologie spécifique peut être construite [39, s83, s88]. Elle est particulièrement utile dans les domaines transversaux peu structurés pour des utilisateurs qui ne connaissent *a priori* pas les mots-clefs pertinents.

Depuis 2005, Robert et moi avons fait évoluer *AlvisIR*, le premier moteur de recherche du projet Alvis, dont les fonctions d'indexation et d'interrogation avaient été développées par nos partenaires HIIT (*Helsinki Institute for Information Technology*) et les sociétés Index Data et Exalead. Il a été complètement intégré à la chaîne AlvisNLP/ML [27]. Ludovic Patey et Frédéric Papazian ont réimplémenté AlvisIR [247] en 2009 sous ma direction, de telle sorte que ce logiciel devenu complètement générique puisse être instancié pour une application particulière en quelques heures [l248]. Les ressources ontologies et terminologies sont gérées par une base de données pour de meilleures performances.

Notre représentation de l'annotation sémantique est particulièrement performante parce qu'elle évite l'extension de la requête, très coûteuse pour des ontologies volumineuses [r150, r155, r156]. Elle consiste à annoter les termes des documents par les chemins des concepts jusqu'à la racine. L'appariement des termes des requêtes et des terms des documents se fait ensuite par une simple comparaison de chaînes de caractères des chemins.

Figure 12. Exemple de requête du moteur de recherche sémantique AlvisIR Sam Blé.

Les interfaces de ce type de moteurs sont complétées par des facettes qui visualisent les termes des documents réponse (Figure 12, colonne de gauche), pour guider le raffinement de la requête. Les nombreux projets applicatifs où des instances d'AlvisIR ont été déployées, ont affiné la définition des besoins. Outre de meilleures performances, et le développement de fonctions documentaires classiques (gestion des utilisateurs, sauvegarde des résultats, ajouts de documents), AlvisIR propose une recherche documentaire originale par apport à l'état de l'art en RI, par navigation dans l'ontologie comme alternative aux requêtes et une fonction de raffinement de requête par concepts dont le principe a été validé par les utilisateurs. Aujourd'hui, il existe de nombreux équivalents intégrant étroitement des fonctions d'analyse sémantique basées sur l'apprentissage et de RI. A la différence d'AlvisIR, les classes sont calculées sur des critères statistiques. AlvisIR continue également à supporter la comparaison en terme de temps de réponse avec les moteurs comparables, comme celui du moteur de la société Exalead, que nous avons mesuré.
Les extension d'AlvisIR aujourd'hui concernent l'interrogation et la visualisation de relations d'une part et d'autre part l'intégration avec des fonctions de tableau de bord pour l'aide à la décision.
Les principaux usages d'AlvisIR s'organisent selon la typologie suivante (sont mentionnés entre parenthèses, les projets décrits ci-dessous) :

1. Recherche exploratoire d'information (*MAFOR*)
2. Recherche de documents contenant des informations particulières (*OntoBiotope, FSOV Sam Blé*)
3. Caractérisation du contenu de collections de documents (*Analyse stratégique du département Phase*)
4. Exploitation de grandes collections de documents hétérogènes (*GIS GC HP2E*)
5. Aide à la construction d'ontologie par l'évaluation de sa couverture d'une ontologie pour l'indexation de texte (*TriPhase, OntoBiotope*)

Pour chaque usage, plusieurs projets applicatifs peuvent être cités en exemple. Nous prendrons un pour chacun d'entre eux.

10.1 Exemple de recherche exploratoire d'information

Le projet d'expertise collective MAFOR (*Valorisation des matières fertilisantes d'origine résiduaire sur les sols à usages agricole ou forestier*) est confronté à la difficulté de réunir et synthétiser les

informations sur les épandages d'effluents dans un but de fertilisation agricole à partir d'un ensemble de références de quelques centaines de référence. La démarche bibliographique consiste à d'abord construire un large corpus de plusieurs centaine de milliers de références par des requêtes de mots-clefs avec le moteur de recherche classique de *Web of Science*[18]. Les termes du corpus sont annotés automatiquement par BioYateA, puis AlvisIR est utilisé itérativement pour identifier les documents hors sujet qui ont été sélectionnés par erreur et corriger les requêtes en conséquence pour ajuster le corpus. Les fonctions de surlignage et de raffinement de requête sont particulièrement utiles ici. Cette approche s'est révélée efficace pour ajouter aux requêtes des critères que les experts n'avaient pas identifiés dans un premier temps, par exemple exclure les contaminations de milieux marins, mais également pour préciser les termes trop généraux ou ambigüs [s112]. Le projet est conduit par la DEPE (*Délégation à l'Expertise Collective, Prospective et Etudes*.) sur commande des pouvoirs publics.

10.2 Recherche d'informations particulières dans les documents

Le projet *Sam Blé* (*Sélection du blé par marqueur génétique*) utilise le moteur AlvisIR pour indexer l'information sur les marqueurs génétiques des gènes qui contrôlent le phénotype d'intérêt d'une variété donnée dans une collection de 5 000 articles[19] [239]. Toutes ces informations sont annotées automatiquement par Alvis. L'organisation hiérarchique des phénotypes dans une ontologie de 400 concepts permet aux sélectionneurs de poser des requêtes sur un phénotype général et d'identifier les documents rapportant des conclusions précises sans avoir à décrire toutes les formulations du phénotype. Par exemple, le terme de la requête *pest resistance* est automatiquement étendu en des centaines de termes particuliers, comme *resistance to Puccinia triticina* qui est un champignon. Outre l'INRA, le projet financé par FSOV implique les semenciers français et les instituts techniques concernés (Union Française des Semenciers, Arvalis).

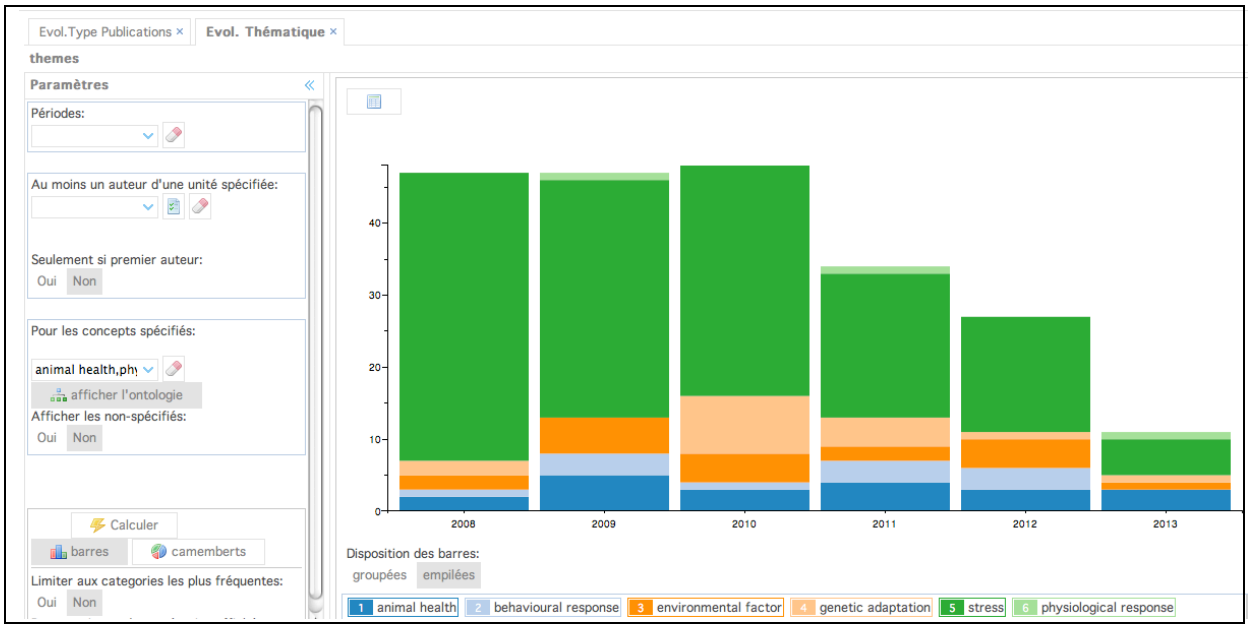


Figure 13. Exemple de visualisation de la distribution des publications par concept.

10.3 Caractérisation du contenu de collections de documents

Le projet *TriPhase*, outre un moteur de recherche bibliographique en physiologie animale mentionné ci-dessus, vise à indexer l'ensemble des 9 397 publications en français et en anglais du département scientifique INRA Phase à des fins d'analyse stratégique[20]. Le moteur sémantique indexe les métadonnées bibliographiques, y compris des informations de structure comme les unités de recherche et les types de publication (rapport, article de congrès, de journal, facteur d'impact), mais surtout les

[18] http://apps.webofknowledge.com
[19] http://bibliome.jouy.inra.fr/test/alvisir/FSOV.
[20] http://bibliome.jouy.inra.fr/test/alvisir/webalvis/triphase2

concepts des documents grâce à l'ontologie de 1600 concepts et 2 200 synonymes et traductions que nous développons actuellement pour ce domaine.
AlvisIR avec ses fonctions de tableau de bord permet au chef de département de comparer visuellement les volumes de publications en croisant différents critères et à différents niveaux de granularité thématique. Par exemple, comparer au cours du temps l'évolution respective des thèmes dans le département, ou pour des unités données. La figure 13 montre l'évolution au cours du temps des publications des principaux thèmes du département, *Animal Health*, *Behavioral response*, *Environmental Factor*, *etc*.

10.4 Exploitation de collections de documents très hétérogènes et nombreux

Le projet Système d'Information du GIS (*Groupement d'Intérêt Scientifique*) GC HP2E (*Grandes Cultures à Hautes Performances Economiques et Environementales*)[21] porte sur l'implantation des grandes cultures.) Il implique principalement INRA Transfert (la filitale de transfert de l'INRA), la DV/IST (*Direction de la Valorisation / Information Scientifique et Technique de l'INRA*), Arvalis, le CETIOM (*Centre Technique des Oléagineux*) et l'ITB (*Institut Technique de la Betterave*). L'ensemble des partenaires du GIS expoite une documentation très riche et diverses, composée de documents en français et anglais, scientifiques, techniques, de vulgarisation, d'articles et de rapports. Notre solution de moteur de recherche sémantique permet sans structuration documentaire dont le GIS n'a pas les moyens, d'indexer et fouiller cette collection de manière efficace [l230][22]. Les 847 documents recensés ont été numérisés et indexés sémantiquement par une une ontologie bilingue de 4 919 concepts et termes, développée par Paul Bui-Quang sous ma direction [c234]. La figure 14 montre les 13 principaux niveaux de cette ontologie.

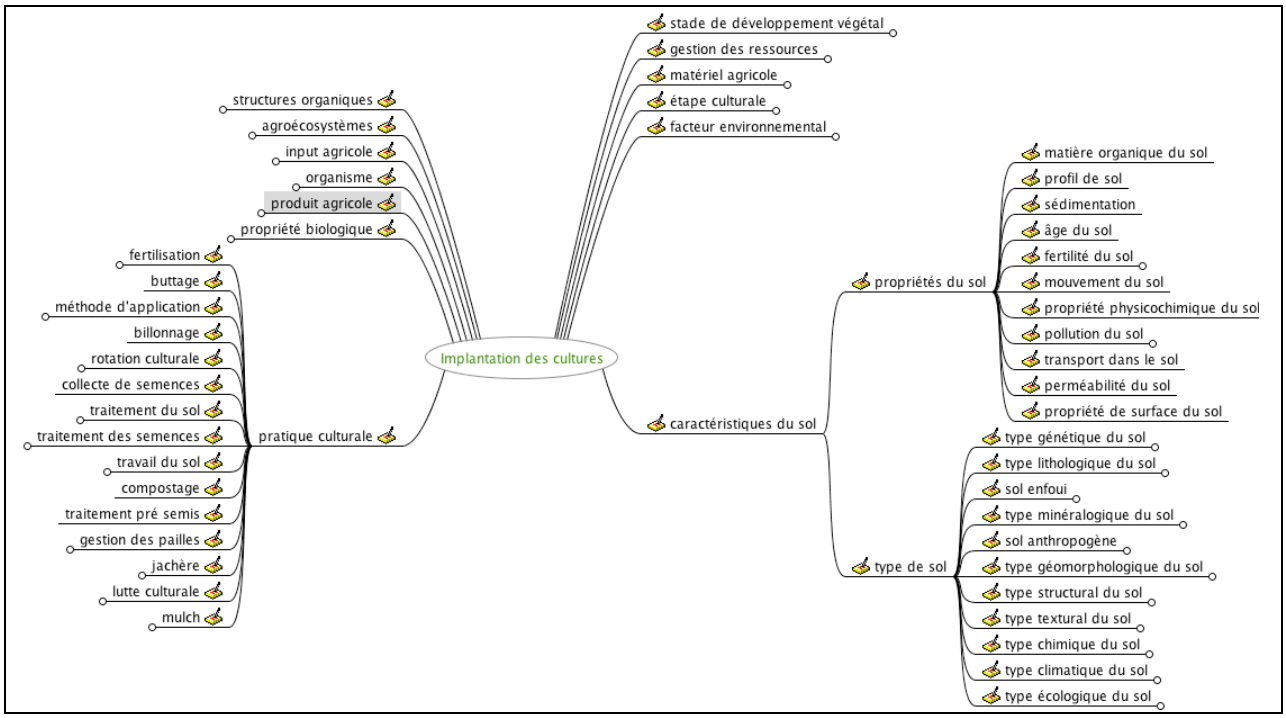


Figure 14. Niveaux généraux de l'ontologie implantation des cultures.

10.5 Aide à la construction d'ontologie par l'évaluation de sa couverture d'une ontologie pour l'indexation de texte

Le projet *OntoBiotope* vise à extraire et normaliser les descriptions d'habitats bactériens mentionnées dans la littérature scientifique. Pour cela, j'ai défini l'ontologie OntoBiotope qui décrit actuellement 2 120 concepts et termes [c237]. Un des moyens productifs d'étendre l'ontologie par de nouveaux habitats est d'utiliser AlvisIR pour sélectionner les références qui mentionnent le type d'habitat à

[21] http://www.gchp2e.fr/
[22] http://bibliome.jouy.inra.fr/test/alvisir/gis_hp2e_NEW

compléter et dans ces références, d'identifier les habitats qui ne sont pas indexés. Par exemple pour étendre le sous-arbre des aliments, on interroge avec le terme *rice*. AlvisIR ramène le document de PMID 15341665 dont l'extrait : *Interviews with the chef revealed that many eggs were used in the preparation of egg-fried rice, which was left at room temperature for seven hours and was used in the preparation of the other two rice dishes.* (PMID 15341665), avec ses habitats identifiés grâce à OntoMap et l'ontologie OntoBiotope surlignés en bleu. Les entités en orange inconnues de l'ontologie sont alors ajoutées. Le moteur de recherche est couplé à l'éditeur d'annotation sémantique AlvisAE (*Alvis Annotation Editor*) qui permet de visualiser l'ensemble des annotations sémantiques d'un document complet, ce que les courts extraits (*snippets*) du moteur ne permettent pas. L'extension de l'ontologie impliquera les 30 microbiologistes du réseau OntoBiotope du métaprogramme INRA MEM (*Méta-omiques des écosystèmes microbiens*).
La version actuelle du moteur, non publique, indexe 1,2 million de mentions de microorganismes et 1,6 million de mentions de milieux dans les 600 mille références pertinente de PubMed. Son usage dépassera le cadre de la construction d'ontologie pour devenir un outil pour la recherche en écologie bactérienne.

Ce bref panamorama illustre la diversité des usages du moteur de recherche sémantique AlvisIR. Son intégration étroite avec Alvis NLP permet très rapidement de mettre en œuvre la séquence (1) Construction du corpus de documents, (2) acquisition de termino-ontologie (3) indexation sémantique (4) mise en ligne d'une instance du moteur de recherche sémantique. Notre technologie d'indexation sémantique permet donc de construire rapidement et automatiquement un système d'information pour explorer efficacement un grand nombre de documents variés, propriété particulièrement utile pour exploiter un fond documentaire riche et transversal comme celui du GIS.

11. Editeur d'annotation

11.1. Visualisation d'annotation

De nombreux éditeurs d'annotation sémantique ont été proposés ces dernières années, qui permettent d'associer automatiquement ou manuellement à des fragments de texte en langage naturel, visuellement et formellement des catégories ou concepts et des relations préalablement définis dans un schéma formel. La figure 15 en montre un exemple. Les entités sont surlignées, les couleurs distinguent les différents types. Les relations entre entités sont figurées par des lignes. La figure 16 montre un extrait du schéma d'annotation associé, qui est un sous-ensemble de l'ontologie du domaine. Les nœuds représentent des racines des types d'entités. Les relations, ici binaires, sont représentées par les arcs orientés.

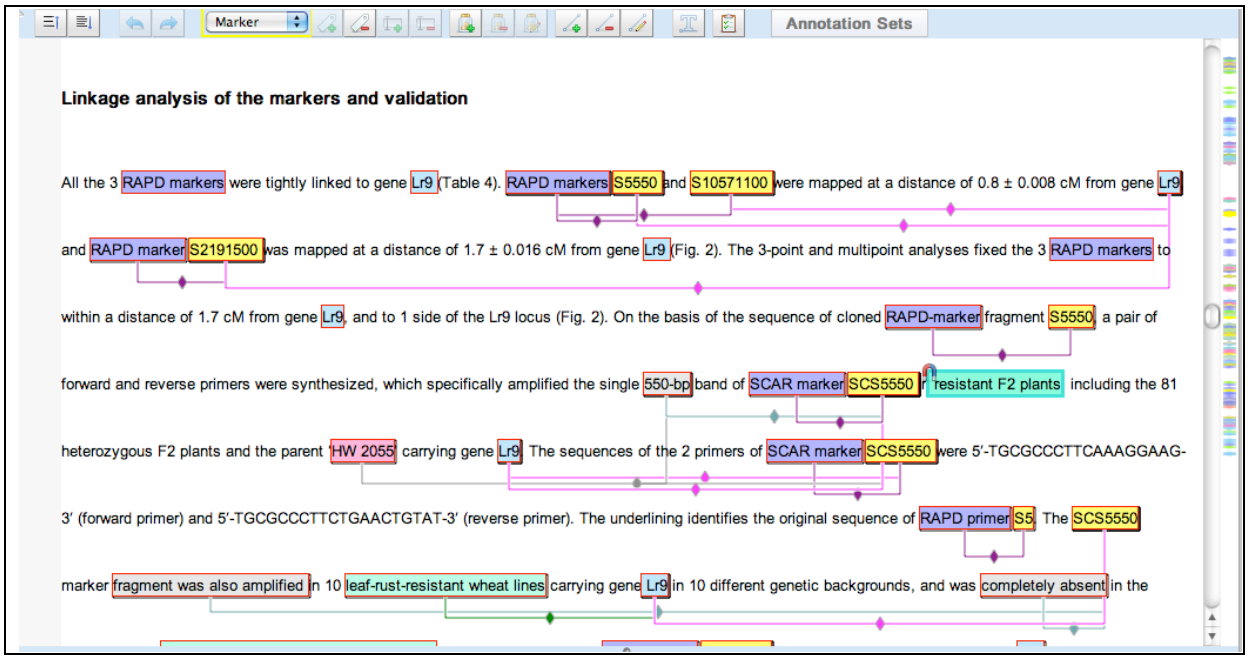


Figure 15. Exemple de visualisation d'annotations sémantiques sur les marqueurs génétiques du blé.

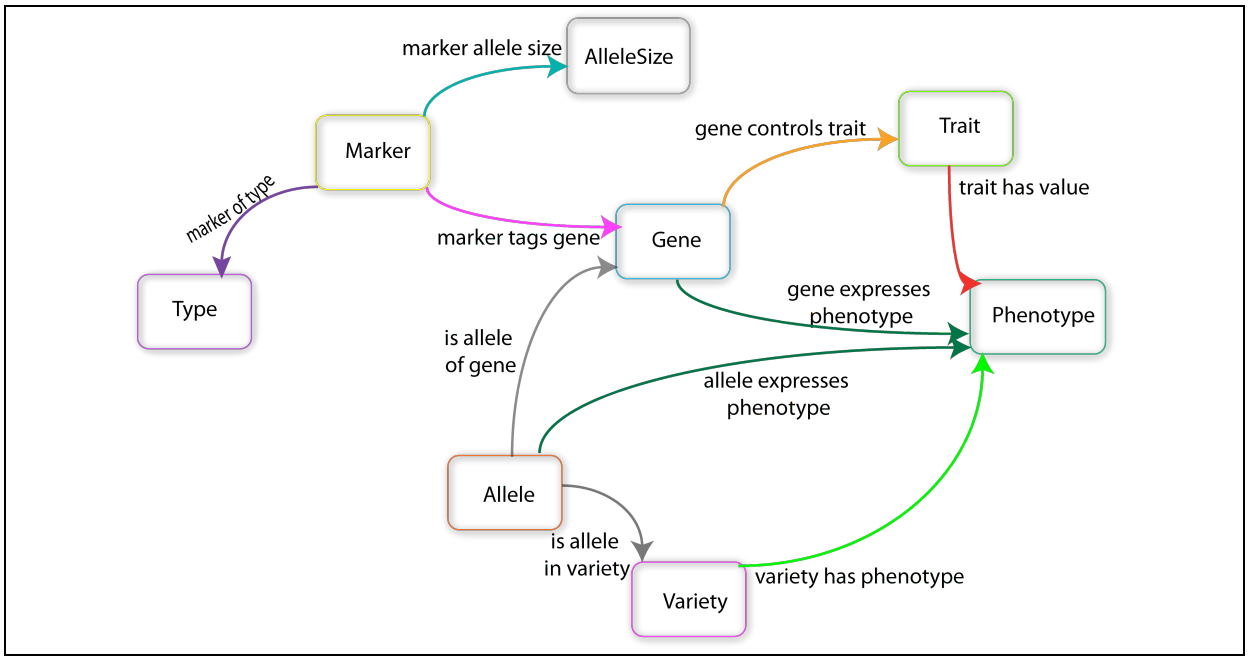


Figure 16. Schéma d'annotation pour les marqueurs génétiques du blé.

La visualisation de textes annotés automatiquement est précieuse dans de nombreuses applications d'extraction d'information, où le retour au texte est nécessaire à partir de l'information structturée, soit parce que l'information doit être vérifiée, ou parce que le contexte contient des informations nécessaires. C'est le cas par exemple des marqueurs génétiques mentionnés plus haut. Les détails de l'expérimentation qui a conduit aux résultats publiés dans les articles et extraits automatiquement sont nécessaires aux sélectionneurs pour évaluer la fiabilité du marqueur. La visualisation graphique d'information riches dans de grands documents pose de nombreuses questions scientifiques et techniques qui sont traitées par exemple dans le workshop d'ACL LAW (*Linguistic Annotation Workshop & Interoperability with Discourse*) qui attire une communauté grandissante. Ces questions portent sur la représentation des connaissances, l'interface homme-machine et l'ergonomie, la visualisation graphique d'éléments mobiles, *etc*. Depuis 1999 les éditeurs d'annotation Caderige[23] puis AlvisAE (*Alvis Annotation Editor*) [l249] ont été développés sous ma direction.
Notre éditeur AlvisAE propose une solution originale qui est décrite dans [23] et [67]. Son développement est principalement dû à Frédéric Papazian et Robert Bossy sous ma direction scientifique. Après une étude approfondie des technologies et des éditeurs existants, nous avons fait le choix technique d'une application Web utilisable avec un navigateur dont le serveur référence les informations dans une base de données. L'interface est en GWT et le serveur en Scala Lift. Le schéma d'annotation est décrit en JSON. Les types de connaissances sont

- Les **entités** attachées à des sections de texte éventuellement non contigües,
- Les **relations** nommées, n-aires entre tout type de connaissances,
- Et les **groupes** non ordonnés de tous types de connaissances.

Dans l'exemple de l'application de sélection du blé, les gènes, les marqueurs et les variétés sont des entités ; les relations sont par exemple *gene_expresses_phenotype_in_variety* qui relie les trois types d'entités gène, phénotype et variété ; le groupe *synonyme* est un exemple de groupe qui permet de créer des ensembles de termes du texte qui sont équivalents sémantiquement comme *barley yellow dwarf* et *BYD*. Ce dernier type groupe est original et il a montré sa pertinence dans plusieurs applications, où il permet de synthétiser l'information en représentant des classes d'équivalence. Par exemple, une relation entre deux éléments de deux groupes est valide pour l'ensemble des éléments des deux groupes.

[23] http://caderige.imag.fr/

Un des points très originaux d'AlvisAE est la visualisation d'annotation d'entités par les concepts d'une ontologie (figure 17). L'entité sélectionnée dans le texte annoté ou dans le tableau des annotations structurées est associée de manière synchronisée au concept surligné dans l'ontologie affichée dans le panneau gauche de la copie d'écran. Dans l'exemple, l'entité *sheep* du texte est associée au concept *sheep* de l'ontologie, dont l'ancêtre est *farm animal*.

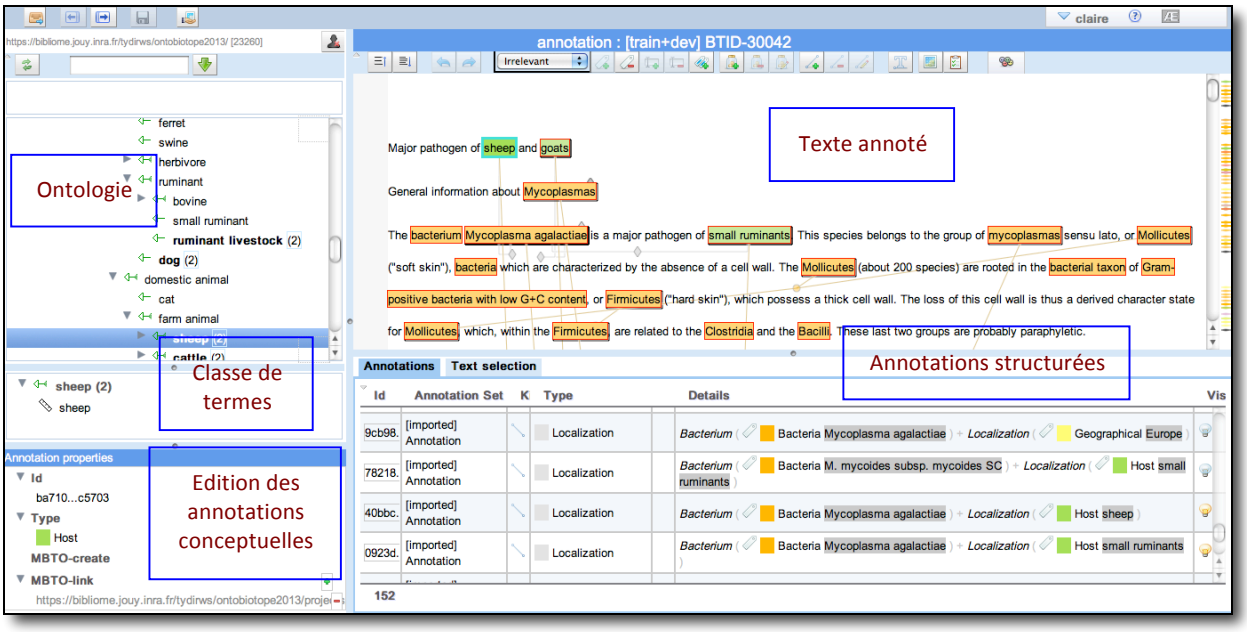


Figure 17. Visualisation d'annotation d'entités par une ontologie, exemple des biotopes bactériens.

AlvisAE est à notre connaissance le seul éditeur qui permet de représenter des entités dont les éléments du texte ne sont pas contigüs. Ce besoin est pourtant fréquent, en particulier dans le cas de coordination ou d'énumération dont la tête (l'élément principal) est factorisé. La figure 18 en donne quelques exemples.

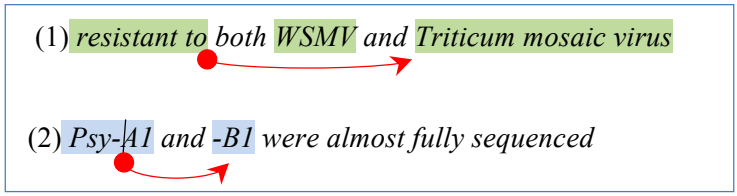


Figure 18. Exemples d'annotation discontinue.

Dans l'exemple (1) le mot *resistant* est commun aux deux entités *resistant to WSMV et resistant to Triticum mosaic virus***.** Dans l'exemple (2), la protéine *Psy* est rattachée aux deux chromosomes *A1* et *B1*. La représentation d'annotations discontinues ou partiellement recouvrantes exclut de représenter les annotations dans le texte en XML. Notre choix de représentation est celui d'annotations dites déportées où les positions des balises sont indiquées explicitement par leur position en caractères. Les annotations peuvent contenir plusieurs fragments de tailles quelconques. La synchronisation avec les éléments graphiques non typographiques comme les lignes pose des problèmes algorithmiques non triviaux, par exemple pour la fonction de zoom.

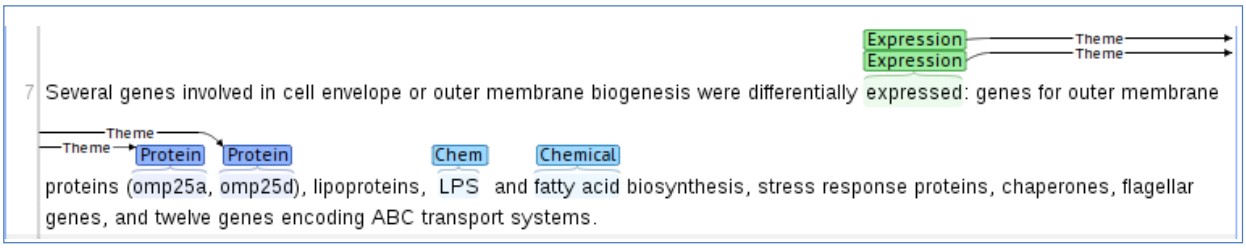


Figure 19. Exemple d'affichage de relation avec l'éditeur d'annotation Brat.

Enfin, la visualisation des relations entre entités distantes pose un problème d'ergonomie que nous avons traité au moyen de lignes de couleur sur une grille horizontale et verticale (Figure 15). Elles sont masquables toutes ou séparément pour préserver la lisibilité du texte. Comparé à d'autres éditeurs récents comme Brat [Stenetorp et al., 2012] où les lignes horizontales rendent les origines et fin de lignes sont indiscernables (Figure 19), cette solution a montré de bonnes propriétés ergonomiques.

11.2 Annotation manuelle

Plus que pour la visualisation, les éditeurs d'annotation sont utilisés manuellement pour formaliser de nouvelles connaissances à partir de textes. Nous avons vu dans les paragraphes précédents que l'analyse sémantique de texte est exploitée pour (1) l'acquisition d'ontologie à partir de corpus de texte et (2) pour l'annotation sémantique de texte. L'annotation manuelle de texte par des experts est une étape critique de ces deux tâches :

- L'annotation sémantique manuelle est très utilisée pour la production de données, soit directement pour nourrir des bases de connaissance structurées, soit pour acquérir des données d'apprentissage pour l'entraînement des méthodes d'annotation sémantique automatique. Dans ce cadre, les éditeurs d'annotation assistent l'annotation sémantique manuelle pour une ontologie donnée et fixée.
- *L'annotation manuelle pour l'acquisition d'ontologie* est largement utilisée et reconnue comme une source conceptuelle et lexicale très pertinente. Dans une perspective de recherche et d'extraction d'information, elle est un moyen utile de guider les choix de modélisation de l'ontologie par l'application visée et d'associer à l'ontologie le niveau terminologique nécessaire à son exploitation.

a. *Annotation manuelle et extraction de relation*

Le premier cas d'utilisation, l'acquisition de données d'apprentissage pour l'entraînement des méthodes d'annotation sémantique automatique est le plus représenté. Il consiste à annoter manuellement des exemples représentés dans le schéma d'annotation, ou à réviser des annotations automatiques de manière à construire un ensemble d'annotations de référence. La figure 20 montre le rôle de l'éditeur AlvisAE dans l'apprentissage de relations : les préannotations sont réalisées par la chaîne AlvisNLP qui utilise les règles d'extraction apprises à l'aide des exemples annotés par AlvisAE, ceci de façon itérative.

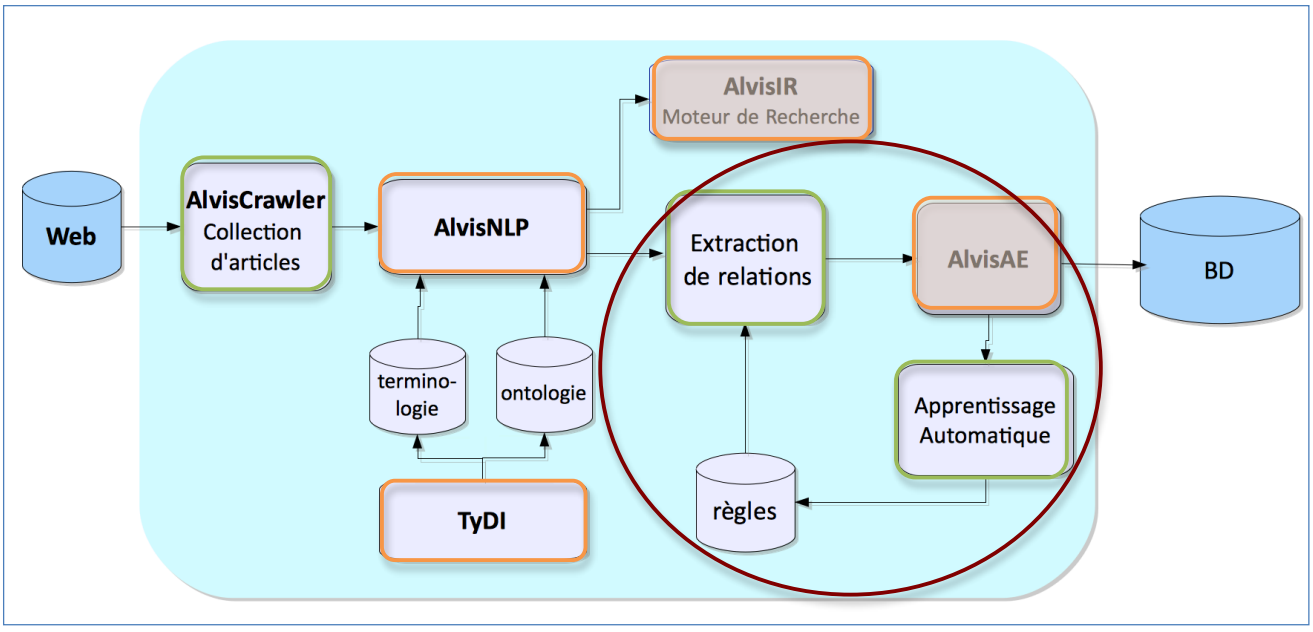


Figure 20. Rôle de l'éditeur d'annotation dans l'extraction de relation - architecture logicielle.

Pour permettre l'annotation, la révision et la publication de différents niveaux de connaissances par différents annotateurs, nous avons défini avec Robert Bossy, une notion de tâche et de *workflow* qu'il

a implémenté dans AlvisAE. Pour chaque campagne d'annotation, les annotateurs se voient attribuer des tâches (annotation, révision, arbitrage), des connaissances à annoter (entités, relations ou groupes spécifiques) et des documents. L'annotation peut être réalisée en double-aveugle ou non, selon les besoins. La figure 21 donne un exemple du tableau de bord de campagne.

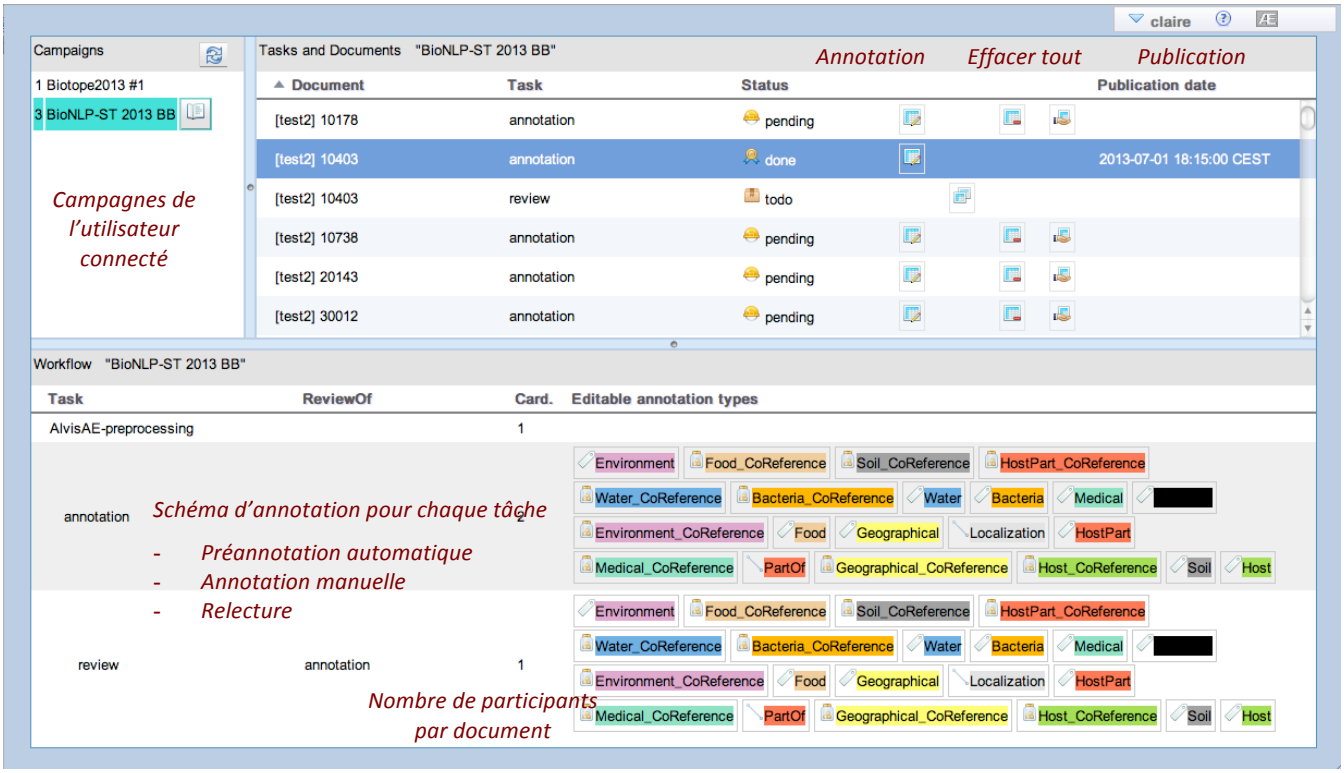


Figure 21. AlvisAE. Gestion de campagne d'annotation.

Dans l'exemple, pour la campagne `BioNLP-ST 2013 BB` (panneau de gauche), selon la liste du panneau central, l'utilisateur doit annoter le document `[test2] 10178`, il a fini d'annoter le document `[test2] 10433` et il doit vérifier les annotations du document `[test2] 10403`. Son schéma d'annotation pour ces différentes tâche est affiché dans le panneau du bas. L'expressivité de la représentation des schémas et l'ergonomie de l'éditeur sont ici des facteurs importants.

b. *Annotation manuelle et construction d'ontologie*

Le développement de l'annotation sémantique automatique à l'aide d'ontologies a créé un besoin d'éditeurs plus riches, avec des interfaces graphiques intuitives, permettant des annotations collectives et concurrentes, proposant des fonctions de résolution de conflits et de calcul d'accord inter-annotateurs [Widlöcher et Mathet, 2009 ; Ogren, 2006].
La figure 22 montre un exemple où les deux ensembles d'annotations figurés dans les deux fenêtres du haut diffèrent par l'annotation de *FUS3*. La fenêtre du bas permet de construire l'annotation consensuelle après arbitrage.
AlvisAE appartient à une nouvelle génération d'éditeur d'annotation qui accroît le potentiel applicatif en permettant conjointement l'annotation de texte *et* la révision collective de l'ontologie. Ce type d'outil s'inscrit dans la double perspective d'annotation sémantique et de population d'ontologie. Le texte est vu comme à la fois comme une source de nouveaux concepts pour l'ontologie et l'objet de l'interprétation formelle guidée par l'ontologie. Quelle que soit la qualité de l'ontologie, l'annotation sémantique manuelle révèle des concepts improprement définis ou absents de l'ontologie qu'il est nécessaire de modéliser dans une perspective d'accès à l'information. C'est particulièrement vrai dans les domaines scientifiques et techniques qui sont caractérisés par l'innovation et la diversité. Réaliser l'annotation de texte et la révision de l'ontologie en deux temps de façon itérée jusqu'à convergence est évidemment improductif.

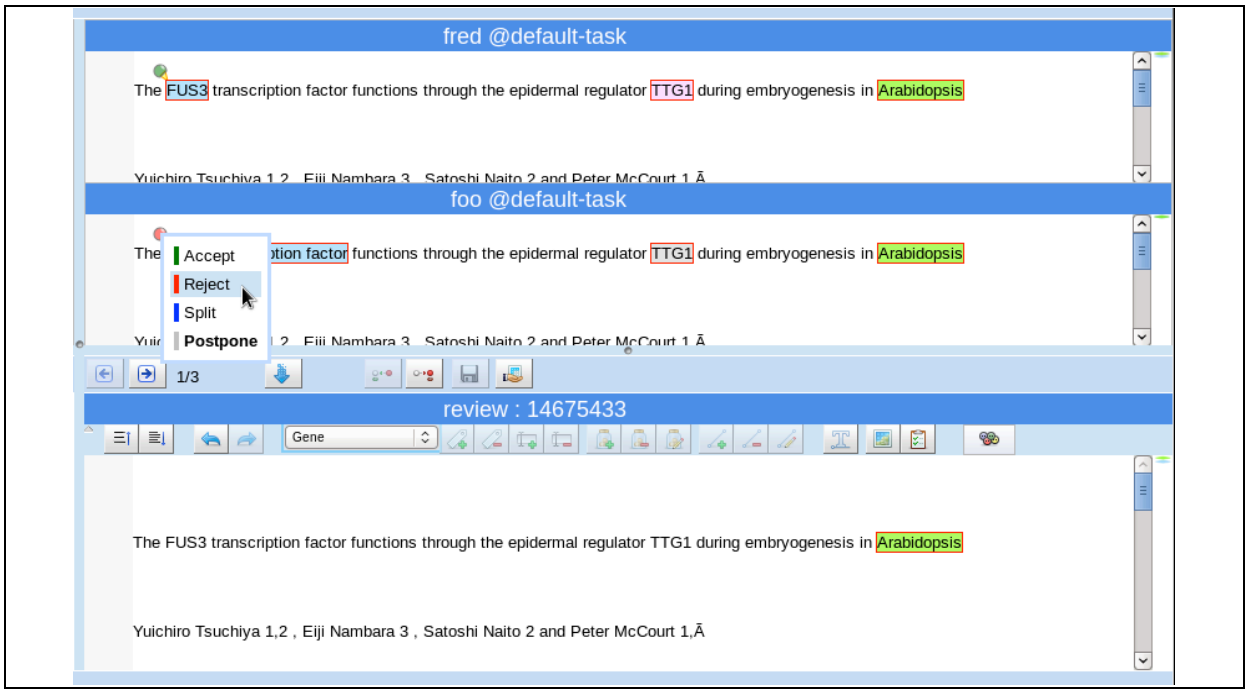


Figure 22. Exemple de résolution de conflit d'annotation avec AlvisAE.

Des équipes pluridisciplinaires, experts métiers, ingénieurs de la connaissance, terminologue, impliquées dans des campagnes d'annotation manuelle de documents sont à même d'identifier les manques de l'ontologie et d'y remédier collectivement. Cette activité révèle les défauts de l'ontologie et en suggère des révisions opérationnelles. Un gain considérable d'efficacité est obtenu grâce à la possibilité de modifier l'ontologie en cours d'annotation et d'identifier parmi les annotations passées celles qui sont à réviser en conséquence. AlvisAE est le seul éditeur d'annotation qui permet la révision l'ontologie au cours de l'annotation sémantique.

Notre utilisation d'AlvisAE dans de nombreux projets d'annotation sémantique dans le domaine des Sciences et la Vie a montré la faisabilité d'applications où l'ontologie est complétée collectivement au fur et à mesure que le texte à annoter révèle de nouveaux concepts. Par exemple, l'annotation des 131 documents de référence de la tâche *Bacteria Biotope* de la *BioNLP Shared Task* en 2013 par l'ontologie *OntoBiotope* a été réalisée par 7 annotateurs[24] en double aveugle [r130] et l'ontologie a été augmentée d'une vingtaine de concepts et termes. La préannotation des catégories réalisée automatiquement par OntoMap a été corrigée en quelques heures par les annotateurs malgré la taille de l'ontologie de plus de 2 000 concepts.

L'architecture d'application Web permet l'utilisation d'AlvisAE par un navigateur et le partage instantané et collectif des évolutions de l'ontologie. Son ergonomie intuitive pour des utilisateurs non informaticiens et la pré-annotation automatique immédiatement visualisable et modifiable ont apporté un gain de temps considérable par rapport aux versions précédentes de l'éditeur.

Outre de nouvelles perspectives applicatives, cette proposition ouvre de nouvelles pistes de recherche sur l'automatisation du maintien de la cohérence des annotations et de l'ontologie, d'un point de vue à la fois formel et social. Le travail collectif provoque des divergences incompatibles entre des versions concurrentes des annotations et de l'ontologie produites par les différents acteurs qu'il faut gérer.

La version actuelle du module de détection d'incohérences entre annotateurs et avec l'ontologie est basé sur une typologie fine des causes de divergence. La typologie des différences entre annotateurs inclut les différences de frontières des unités sémantiques, de concepts associés aux unités sémantiques, d'arguments des relations et de type de la relation. En cas de révision de l'ontologie, sont à valider par ordre de priorité, les annotations qui sont annnotées par

- des concepts supprimés,
- des concepts dont le label a été modifié,
- des concepts dont un des termes non canonique a été modifié,

[24] https://sites.google.com/site/bionlpst2013/tasks/bacteria-biotopes

- des concepts déplacés dont le sens a pu changer,
- des concepts dont un des ancêtres a été déplacé.

Les premières expérimentations sur l'annotation des données de la tâche Bacteria Biotope sont très positives. Les méthodes développées ont permis de maintenir la cohérence entre annotateurs et au cours du temps. L'accès à la demande aux fonctions de comparaison et de validation permet aux utilisateurs de détecter rapidement un risque d'incohérence.

11.3. Applications

AlvisAE est actuellement utilisé dans d'autres projets par des biologistes, sur le développement de la graine, les marqueurs génétiques du blé déjà mentionné et la régulation génique chez *Bacillus subtilis* et. Nous développons ici les deux premiers exemples.

Le projet sur le développement de la graine est conduit par deux biologistes moléculaires des plantes, spécialistes reconnus, Bertrand Dubreucq et Loïc Lepiniec de l'IJPB (*Institut Jean-Pierre Bourgin*) dont l'objectif est de modéliser les réseaux de régulation impliqués dans le développement de la graine d'*Arabidopsis thaliana* à partir des connaissances extraites des textes et d'autres connaissances expérimentales, en particulier d'imagerie. La partie méthodologique est assumée par Dialekti Valsamou en thèse sous ma direction et celle de Pierre Zweigenbaum du LIMSI, financée par l'IDEX (*Initiative d'Excellence*) Paris-Saclay. Elle porte sur l'extraction d'information par des méthodes d'apprentissage et de traitement de la langue. La première partie du travail a consisté à définir un modèle biologique qui est d'une grande richesse (Figure 23).

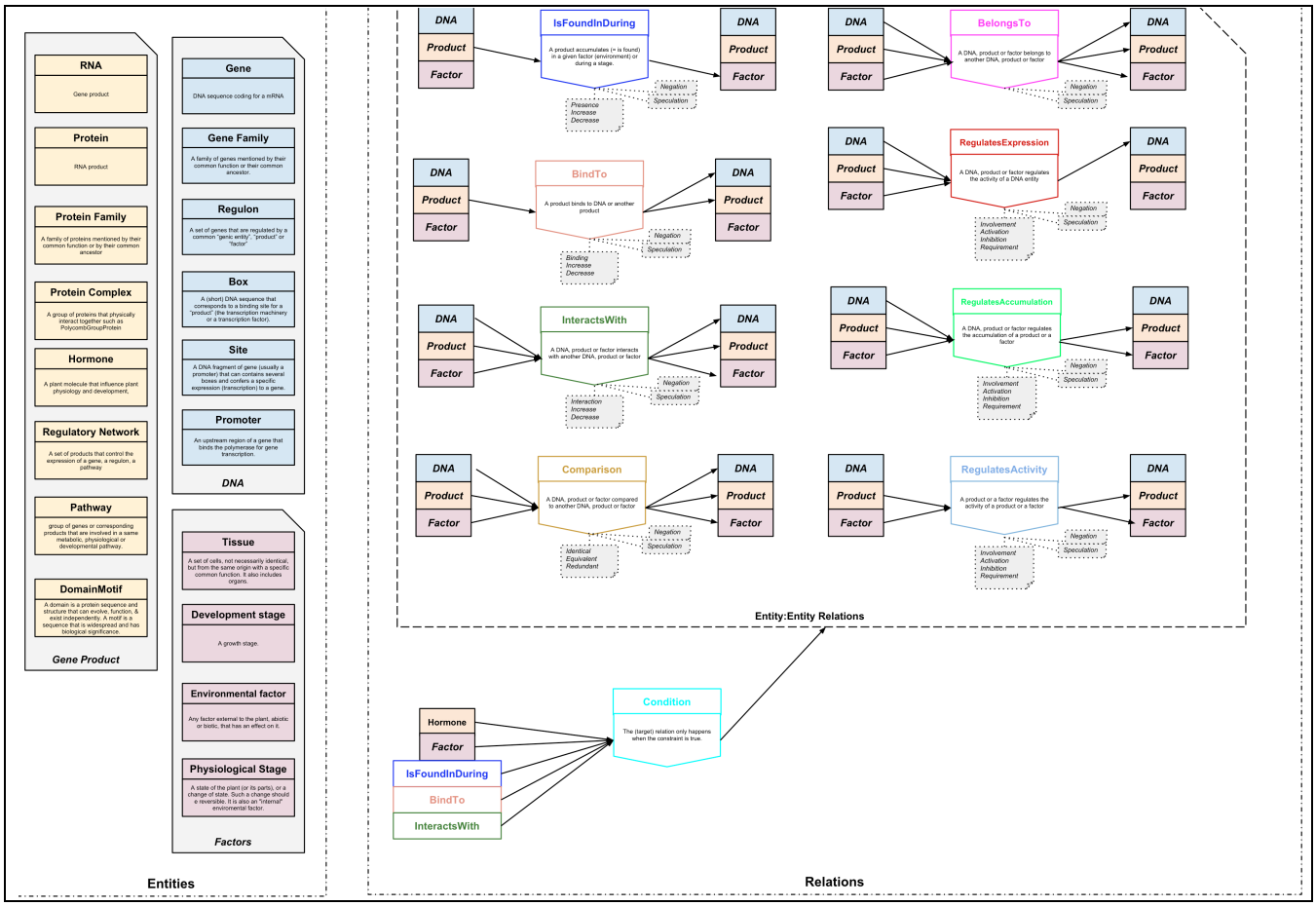


Figure 23. Schéma d'annotation des réseaux de régulation impliqués dans le développement de la graine d'*Arabidopsis thaliana*

Il inclut la formalisation des voies métaboliques, des facteurs environnementaux et des phénotypes dans 16 familles d'entités, 8 relations avec des modalités de négation et de spéculation [r125]. La complexité du modèle soulève des questions nouvelles sur l'annotation d'exemples par les biologistes, en particulier la gestion du risque qu'il y ait plusieurs manières d'annoter la même connaissance,

malgré la précision des consignes d'annotation des *Guidelines*. La figure 24 donne un exemple d'annotation. Les annotations de la première ligne du texte formalisent trois entités *Postembryonic organ formation* (en bleu) qui est un réseau de régulation, *stem cell niche*, *shoot meristem* et *root meristem* (en rose) qui sont des tissus. L'entité *Stem cell niches* régule le réseau de régulation (ligne verte) sous condition d'être localisée dans un des deux méristèmes (lignes rouges). Ces annotations graphiques sont reprises dans le tableau sous le texte. Le panneau de gauche indique les modalités de la relation de régulation, c'est-à-dire ici si le fait est nié (*negation*), s'il est certain (*speculation*) et quel est le rôle de l'entité agent, (*involvement*, *activation*, *inhibition*, *requirement*). La précision du schéma d'annotation et les nombreux exemples et définition du document de Guidelines permettent de limiter les risques d'incohérences.

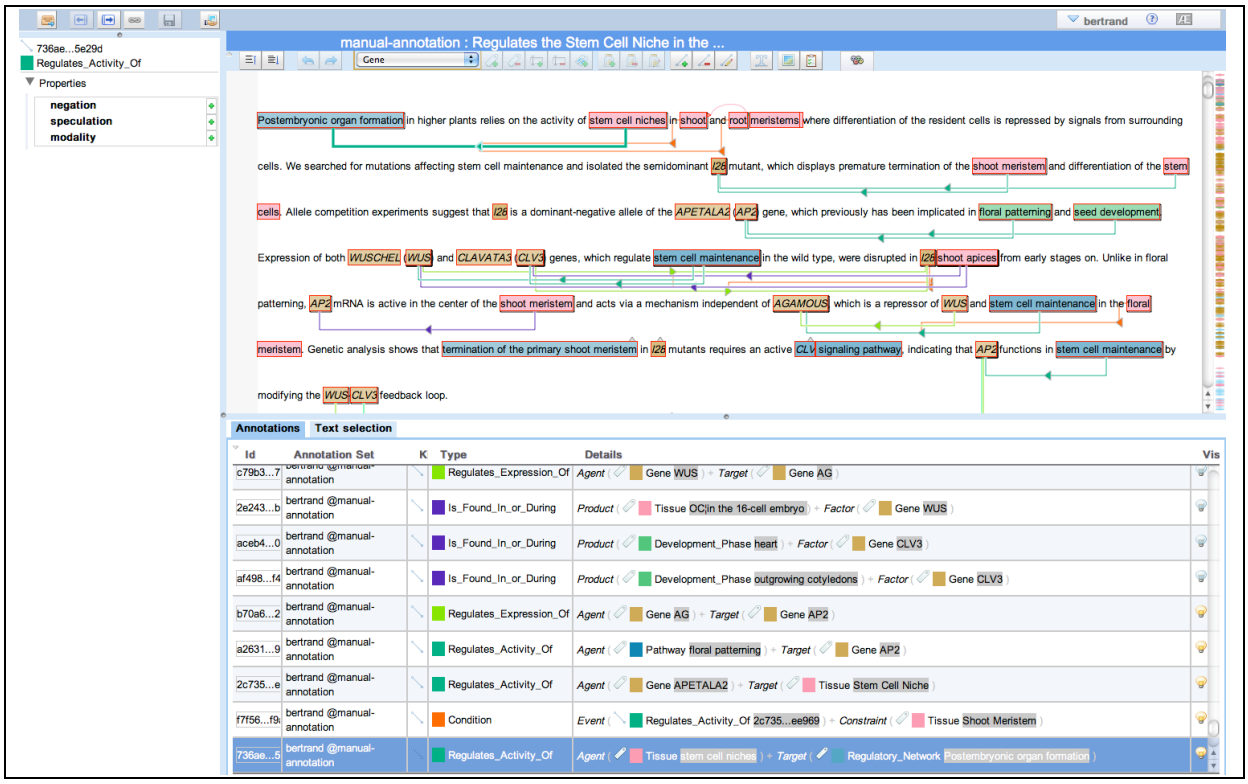


Figure 24. Exemple d'annotation par AlvisAE d'article sur le développement de la graine.

L'utilisation d'AlvisAE dans le projet *Sam Blé* (*Sélection du blé par marqueur génétique*) est assez similaire dans sa première partie : nous avons défini le schéma d'annotation de la figure 16 avec Pierre Sourdille et Marion Ranoux biologistes de l'unité INRA GDEC. Il a été validé et complété grâce à l'annotation par Pierre et Marion d'un ensemble représentifs de documents. J'ai organisé une sesssion de formation à l'utilisation d'AlvisAE avec les biologistes de l'INRA et les 10 partenaires sélectionneurs du projet. Ils annotent chacun une dizaine d'articles pour produire à la fois une base de connaissance mutualisée et les exemples nécessaires à l'induction de règles d'extraction d'information. Outre la visualisation des annotations par les sélectionneurs, le projet prévoit la révision des prédictions erronnées par les persones habilitées, une fois l'application en production. En améliorant de façon continue la qualité des annotations, la communauté permettra non seulement de maintenir une base de connaissance de qualité, mais aussi d'améliorer les performances de la prédiction en réentraînant l'algorithme d'apprentissage avec les exemples révisés. La figure 25 illustre cette boucle de rétroaction.

L'outil AlvisAE associé à notre méthodologie d'annotation de corpus nous a permis de développer des corpus de qualité grâce à la collaboration étroite de biologistes, de spécialistes du TALN et de l'apprentissage automatique [r137, r138]. Ils sont largement diffusés et pérennisés, en particulier à travers l'organisation de compétitions internationales en extraction d'information. Notre objectif à court terme est de valoriser les ensembles de données de qualité que nous produisons pour notre

propre évaluation et d'attirer les chercheurs en extraction d'information pour le développement de nouvelles méthodes adaptées à des questions de recherche pertinentes en biologie.

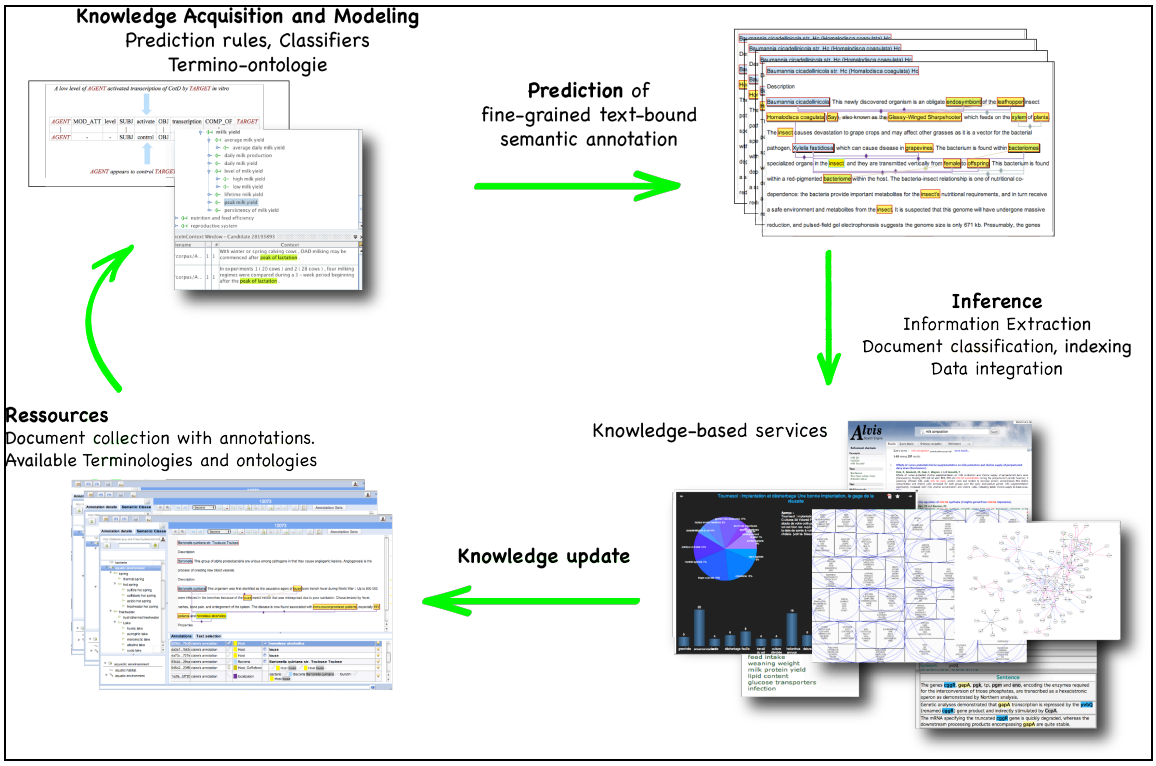


Figure 25. Itération acquisition de connaissance / prédiction / révision.

Activités scientifiques transversales

La présentation des méthodes qui précède a donné de nombreux exemples de la motivation des questions méthodologiques et des résultats obtenus dans le domaine de la biologie. Dans cette partie, les questions biologiques sous-jacentes à ces exemples sont explicitées. Les diverses applications citées relèvent d'un même cadre en biologie : (1) au niveau de la cellule : la modélisation de réseaux de régulation génétiques et du métabolisme, (2) la relation génotype – phénotype et plus largement, (3) la description des phénotypes multi-espèces et l'adaptation au milieu. Ces thèmes sont des objets d'étude de l'unité MIG à laquelle j'appartiens et grâce à laquelle j'ai créé les collaborations dont les motivations en biologie sont décrites ici.

1. Réseaux de régulation génique et métabolisme

Le volume d'informations génétiques a rendu impossible la maîtrise par un biologiste de l'ensemble des connaissances nécessaires à la compréhension du fonctionnement de la cellule [178]. La majorité de ces connaissances est décrite uniquement dans des articles et pas dans des bases de données. C'est un des enjeux de la biologie prédictive que d'intégrer les connaissances expérimentales et bibliographiques dans un modèle cohérent et multi-échelle [s124]. Une des connaissances essentielles est celle de la régulation des gènes à travers les interactions géniques. Pour être exploitables pour la modélisation des réseaux de régulation et du métabolisme, les interactions géniques doivent être extraites des articles sous une forme structurée [s117]. Cette question biologique a été historiquement la première étudiée en extraction de connaissance.

1.1 Bactéries

Les enjeux de la connaissance fondamentale des mécanismes biologiques chez les bactéries sont nombreux et bien connus du grand public. Elles sont étudiées pour leur pathogénécité, mais aussi pour leur contribution à la transformation des produits, agro-alimentaires et non agro-alimentaires (par exemple, la fabrication du saucisson, la production d'éthanol), leur contribution à la croissance des plantes, par exemple en symbiose, la dépollution des effluents, *etc.* la liste est longue. Le choix d'espèces modèles permet de concentrer les études fondamentales sur des espèces simples à manipuler sans risque et de reproduction rapide.

J'étudie l'extraction de réseaux de régulations chez la bactérie modèle *Bacillus subtilis* en collaboration étroite avec Philippe Bessières, microbiologiste de MIG. Avec l'équipe Bibliome, nous avons développé des méthode d'EI et nous avons construit les nomenclatures, terminologies et ontologies nécessaires à l'étude des procaryotes. Nous avons annoté manuellement des articles de référence avec des schémas détaillés d'une grande pertinence biologique de manière à produire des exemples d'apprentissage indispensables à la recherche dans ces domaines [c242].

Nous utilisons ces corpus d'évaluation pour organiser des compétitions internationales. Depuis le début des années 90 avec MUC et TREC, la comparaison formelle des méthodes de "text-mining " est devenue incontournable : il est difficile de publier sans évaluer et comparer les résultats sur des données standards (*benchmarks*) publiées à l'occasion de compétitions internationales. Les principales en "text-mining " pour la biologie sont *BioCreative* et *BioNLP shared tasks*, aux comités de programme ou d'organisation desquels j'appartiens. Initialement ces compétitions portaient sur la santé humaine, sans véritable analyse des questions biologiques. La conséquence en est que l'énorme majorité des publications de "text-mining " pour la biologie concernent aujourd'hui l'homme et la compréhension de la problématique biologique reste superficielle. Cela s'est longtemps traduit par des données annotées de qualité médiocre. Cela nous a conduit en 2005 à organiser le challenge *Learning Language in Logic* (LLL) avec James Cussens (Univ. York) [40, 255], sur l'extraction d'interaction génique associé à la conférence ICML. Il a contribué à faire émerger ce sujet devenu populaire. LLL a attiré 13 participants, ce qui était très appréciable en 2005, grâce à la qualité et à la richesse des données qui sont encore aujourd'hui des données de référence.

Depuis 2010, associés au *Tsujii laboratory* (Université de Tokyo), au DBCLS (*Database Center for Life Science in Japan*) et au NACTEM, avec Robert Bossy nous organisons des tâches de *BioNLP*

shared task dans deux éditions, *Bacteria Interaction* en 2011[25] [2, 284] et *Gene Regulation Network* en 2013[26] [20]. J'étais *chair* de BioNLP-ST'13 [22]. La dernière tâche, GRN, a pour objectif l'extraction du réseau de régulation et c'est par rapport au réseau de référence que les méthodes d'extraction sont évaluées. Le progrès accompli en moins de 10 ans, à la fois dans la formalisation de la question biologique et dans les performances des méthodes permet aujourd'hui d'utiliser les méthodes à grande échelle et de l'étendre à d'autres espèces. Si l'objectif d'automatisation complète de l'extraction des réseaux n'est pas encore atteint, la qualité est telle que les prédictions sont utilisables dans des outils destinés aux biologistes, comme l'est CoCitation pour la visualisation de phrases d'interaction et de leurs entités [s91, s92, s93, 252].

1.2 Réseaux de régulation chez l'arabette

Depuis 2012, la collaboration avec Bertrand Dubreucq et Loïc Lepiniec, biologistes de l'IJPB porte sur l'extraction d'information des réseaux de régulation impliqués dans le développement de la graine chez *Arabidopsis thaliana* (arabette). Les enjeux de la compréhension de la croissance des plantes sont majeurs, en agro-alimentaire, comme en écologie. L'arabette est une espèce modèle depuis les année 80 et la masse de connaissance accumulée est considérable, mais éminemment dispersée dont beaucoup sous forme d'articles [Van Landeghem et al., 2013]. Les méthodes mises au point sur l'arabette sont généralisées aux plantes d'intérêt.

L'étude d'*Arabidopsis* met à l'épreuve la généralisation des méthodes d'EI développées pour les bactéries, que nous adapterons à d'autres plantes comme le blé. Elle requiert la prise en compte des facteurs environnementaux et des phénotypes avec leurs interactions avec le métabolisme. Chez les bactéries, ces éléments sont plus simples : les principaux facteurs sont des facteurs de stress principalement thermique et de présence ou non de nutriment, les principaux phénotypes sont décrits au niveau biologique comme la mutation d'un gène knock-out et son impact sur l'expression des gènes. Inversement, les noms de gènes sont normalisés. La prise en compte des différents niveaux de complexité chez les plantes, biologique, physiologique et phénotypique requièrent de nouveaux traitements linguistiques. Ne citons que l'analyse terminologique pour détecter les phénotypes dans toute leur diversité et le traitement des anaphores et de la métonymie fréquentes en raison de la multiplicité des objets biologiques mentionnés dans une même article. Inversement, la grande taille de la littérature concernée rend l'automatisation de l'extraction bien plus critique pour les biologistes.

2. Relation génotype – phénotype

Les enjeux de production et de qualité des plantes cultivées sont tels que les objectifs à long terme de connaissance fondamentale des mécanismes biologiques tels que décrit pour l'arabette sont incompatibles avec les objectifs économiques à court terme. La sélection des caractères d'intérêt est une pratique aussi ancienne que l'agriculture. Elle associe variété et phénotype sans connaissance des mécanimes biologiques sous-jacents et de la fonction des gènes. Elle bénéficie aujourd'hui des outils de génotypage et de phénotypage qui permettent d'associer, sinon un gène ou un groupe de gènes à un caractère d'intérêt (par exemple, la résistance à la rouille), mais au moins un marqueur ou une région (QTL), sans modélisation des réseaux biologiques. L'hypothèse est que la présence de l'allèle du marqueur pour une variété donnée est un bon indicateur de la présence de l'allèle du gène qui contrôle le phénotype, c'est-à-dire la valeur du trait recherché. La recherche de ces marqueurs et leur validation, en l'absence du génome complet de l'espèce est une tâche très fastidieuse et coûteuse. L'information sur les marqueurs génétiques associés aux phénotypes d'intérêt est devenue une ressource majeure pour la sélection génétique, d'autant plus qu'elle est peu mutualisée, contrairement aux animaux comme la vache laitière. Le manque d'investissement dans des études de grande ampleur met en péril les semenciers de petite taille, contribuant à la concentration du secteur.

Dans l'exemple du blé, la majorité des marqueurs ne sont décrits que sous forme textuelle dans des articles publiés et accessibles à tous. Le projet SAM (*Sélection Assistée par Marqueurs*) Blé (2010-2014) est financé par le Fond d'Obtention Végétale (FSOV). Il est coordonné par Pierre Sourdille de l'unité INRA GDEC (*Génétique, Diversité et Ecophysiologie des Céréales*), il implique les principaux semenciers français, des instituts techniques (Arvalis, UFS). Outre de nouvelles expériences de

[25] https://sites.google.com/site/bionlpst2013/tasks/gene-regulation-network
[26] https://sites.google.com/site/bionlpst/home/bacteria-gene-interactions

phénotypage *in vitro* et en plein champs, le projet a pour objectif l'extraction des informations de relations marqueur – variété – phénotype à partir de la bibliographie, qui seront capitalisées et exploitées par les obtenteurs dans une base de données croisées avec les autres informations génétiques. Les phénotypes de résistance aux maladies, de développement de la plante et de qualité boulangère sont privilégiés ici. Avec Dialekti, nous adaptons et évaluons la méthode SPGAK pour prédire les information à extraire à partir des exemples de référence [r131, c240] annotés par les sélectionneurs comme décrit plus haut.
J'ai développé une ontologie de 400 concepts qui décrit l'ensemble des traits et phénotypes, du plan moléculaire (teneur en protéine) au plan agroalimentaire (épaisseur de la croute du pain). Thierry Marcel (IJPB) a contribué à la description des maladies fongiques [c238]. Le moteur de recherche sémantique *AlvisSamBlé* [l232] indexe une collection bibliographique construite automatiquement par *AlvisCrawler* [l251] sur la base d'une requête *marker gene wheat* à partir de la base bibliographique WoS. *AlvisCrawler* développé par Dialekti pour le blé a été généralisé à toute la bibliographie scientifique.

3. Description des phénotypes multi-espèces

3.1 Les animaux

L'analyse et la normalisation de traits et phénotypes sont également au cœur du projet ATOL (*Animal Trait Ontology*)[27] du département PHASE qui porte sur les traits des animaux de rente (2010-). La variabilité des descriptions des traits pour une même espèce et entre espèces est un obstacle à l'exploitation de la littérature et des bases de données de phénotypes. L'ontologie ATOL répond à ce besoin en proposant un ensemble normalisé et structurés de termes de référence. Si en biologie animale, la démarche est déjà très avancée pour plusieurs espèces académiques comme la souris, aucune de ces ontologies n'est spécifiquement dédiée aux animaux d'élevage (poissons, volailles, mammifères) pour lesquels pourtant de nombreux phénotypes d'intérêt économique sont étudiés. ATOL vise à définir et à organiser les caractères phénotypiques des animaux d'élevage, en prenant en compte les préoccupations sociétales et les grands types de production (lait, œuf, viande, fertilité, alimentation). Il implique une quarantines de chercheurs en physiologie animale coordonnés par 10 curateurs, un par domaine étudié.
Par ailleurs ATOL utilisé par des outils de recherche sémantique permettra une recherche de référence bibliographique plus puissante et pertinente sur les phénotypes des animaux d'élevage à partir du corpus de références présent sur le web [s96]. Comme décrit ci-dessus (§5) notre collaboration avec les 10 curateurs d'ATOL formés à l'utilisation de nos outil terminologiques en particulier TyDI a permis d'enrichir l'ontologie ATOL et de développer une première instance du moteur de recherche [l231] pour la revue *Animal* [8, s95]. Il est en cours d'extension à toute la littérature du domaine dans le projet *TriPhase* décrit ci-dessus. L'ontologie EOL (*Environment Ontology for Livestock*) a été complétée de manière similaire pour les milieux aquatiques et plus largement les milieux d'élevage dans le projet européen AquaExcel [66].

3.2 Les bactéries

L'étude systématique des spécificités génétiques qui contrôlent les phénotypes adaptés aux milieux est bien plus riche chez les microorganismes pour lesquels les chercheurs disposent d'abondantes informations génétiques, mais la description des milieux reste exclusivement en langage naturel. Sans mise en relation entre espèces et milieux et sans normalisation sémantique par une ontologie, il n'y a pas de comparaison automatique possible entre les biotopes exprimés différemment, par exemple, *French cheese* et *Camembert*. Les enjeux ici sont de différents ordre. Une meilleure connaissance des environnements des bactéries avec leurs propriétés physico-chimiques (température, salinité, présence d'oxygène) permet de caractériser les spécificités génétiques impliquées dans l'adaptation à l'environnement. Concrètement, si un *cluster* de gène n'est identifié que dans des espèces d'un milieu d'un certain type et pas ailleurs, il est probable que ce *cluster* soit impliqué dans l'adaptation des espèces à ce milieu. Au-delà d'une meilleure compréhension de l'interaction de la bactérie avec son environement, les applications potentielles sont considérables dans tous les domaines de la

[27] http://www.atol-ontology.com/index.php/fr/structure-atol-fr/objectifs-fr

microbiologie, pour la compréhension des mécanismes de pathogénicité dans le domaine de la santé, l'identification de bactéries plus efficaces pour certaines fonctions (dépollution, fermentation, *etc*.).
Pour cela, j'ai défini l'ontologie OntoBiotope de 2 120 concepts et termes [c237]. J'ai identifié les concepts pertinents pour l'ontologie à partir de plusieurs sources documentaires, en particulier les champs de description des milieux d'isolement des bactéries dans les bases de génomes complets comme GOLD (*Genomes OnLine Database*)[28]. L'application de BioYaTeA a produit un grand nombre de termes pertinents que j'ai validés et structurés. L'ontologie OntoBiotope diffère d'EnvO[29] qui a les mêmes objectifs de normalisation du vocabulaire, principalement pour les eucaryotes, par ses choix de modélisation résolument orientés pour répondre à la tâche d'extraction d'information des milieux bactériens. Les concepts d'OntoBiotope sont choisis en fonction de leur pertinence pour regrouper des milieux qui ont des proprités similaires pour les bactéries. Par exemple, du point de vue de OntoBiotope, deux aliments d'origine végétales vont être plus similaires du fait du traitement habituel, par exemple la cuisson, que de leur proximité taxonomique. Par exemple, le concombre et la courgette appartiennent tous deux à la même famille des *Cucurbitaceae*. Le premier se consomme cru, le second se consomme cuit. Des mentions de présence de bactéries sans précision sur l'état du légume sont à interpréter dans l'état d'aliment prêt à consommer. Les bactéries comme *E. coli* sont plus vraisemblablement les mêmes dans les concombres et les épinards crus (*Amaranthaceae*) qui n'appartiennent pas à la même famille. C'est mon choix de modélisation pour OntoBiotope. Par ailleurs, le choix du lexique dans OntoBiotope est celui du vocabulaire utilisé par les biologistes pour décrire l'origine de leur prélèvements car c'est dans ce contexte que l'ontologie sera utilisée. EnvO abonde de termes techniques spécialisés pour chacun des domaines par exemple la pédologie pour l'étude des bactéries du sol.
Avec cette ontologie, 1,2 million de mentions de microorganismes et 1,6 million de mentions de milieux sont référencés dans les 600 mille entrées pertinente de PubMed. L'intégration de ces nouvelles données avec les données génétiques du portail IGO permettra dès cette année de rechercher en ligne, et à très grande échelle, quelles sont les spécificités génétiques d'une espèce ou d'une famille qui sont exclusives d'un milieu [s94].
Notre méthode de prédiction des relations entre les espèces et de leurs milieux associés intègre des ressources lexicales (OntoBiotope et notre taxinomie des espèces) et les méthodes originales d'extraction d'information développées par l'équipe, AlvisNLP, BioYateA, OntoMap, la reconnaissance d'espèces et la résolution d'anaphore. Pour évaluer la méthode et plus généralement promouvoir ce sujet auprès de la communauté BioNLP, Robert et moi avons organisé les tâches *Bacteria Biotope* de la compétition *BioNLP Shared Task* en 2011 [25, 2] et en 2013 [21]. Le corpus collectivement annoté rassemble des textes de référence de centres de séquençage bactériens [c236]. Notre méthode, publiée dans *BMC Bioinformatics* [283], a obtenu le meilleur score de la compétition en 2011.
J'anime avec Philippe Bessières et Maarten van de Guchte de Micalis le réseau OntoBiotope du métaprogramme MEM (2012-13) [s114]. Il regroupe des microbiologistes INRA de toutes les communautés pour étendre et exploiter ces premiers résultats. Une première réunion a consolidé les objectifs [s116], la prochaine réunion formera les biologistes à l'utilisation de nos outils pour valider le modèle de OntoBiotope, l'étendre et annoter les exemples de références nécessaires à l'apprentissage [r130].

Ces projets de bioinformatiques sont en cours. Les logiciels et les bases de connaissances cités évoluent rapidement avec les progrès des méthodes et la collaborations des spécialistes des domaines. Plus généralement, les applications évoluent de l'extraction d'information fine, vers la modélisation d'ontologie pour l'intégration de données hétérogènes comme dans le cas du projet ATOL [s118, s119, s120].

[28] http://www.genomesonline.org/cgi-bin/GOLD/index.cgi
[29] http://bioportal.bioontology.org/ontologies/1069

Conclusion

L'introduction annonçait mon projet de recherche comme *relevant de [..] deux grandes questions, celle de l'apprentissage automatique de connaissances à partir d'exemples et celle de l'acquisition manuelle de connaissances expertes et de leur modélisation, [..] dans un objectif de conception de systèmes assistants, c'est-à-dire automatiques lors que cela est possible et interagissant avec l'utilisateur si c'est nécessaire ou utile [..]*.

Ce projet, général quand j'ai commencé ma thèse il y a 23 ans, s'est poursuivi et précisé au fil du temps. La problématique de l'articulation de l'apprentissage et de l'acquisition de connaissance avec ses questions de modélisation des traitements et de coopération homme-machine s'est concrétisée avec le développement et l'intégration d'outils multiples dans la plateforme Alvis.

L'*Expert*, figure mythique des années 90, à qui je promettais assistance a pris forme humaine comme biologiste, chargé de valorisation, documentaliste, terminologue et ingénieur de la connaissance. Surtout comme biologiste. C'est cet expert-là, avec ses compétences et ses limites, qui définit les tâches et pose les bonnes questions.

La modélisation de la *théorie du domaine*, initialement en logique d'ordre un – comme dans le reste du monde – a évolué en logique de description pour modéliser des *ontologies*. Si le langage de représentation s'est contraint, le projet de modélisation de connaissance à partir d'exemples reste le même. Le relativisme du modèle en fonction de la tâche reste essentiel.

La vraie grande révélation a été celle de la source infinie de connaissances à interpréter et à modéliser que sont les documents en langue naturelle, avec la linguistique computationnelle comme un formidable outillage pour réaliser ce projet.

C'est dans la convergence de l'apprentissage automatique, du traitement automatique de la langue naturelle et du Web sémantique que le projet pluridisciplinaire initial se précise et se réalise pour répondre à des problématiques biologiques concrètes.

Comment le modèle, ses concepts et ses termes rendent compte de la variabilité des signifiés dans le texte, le mystère continue de s'épaissir.

Jouy-en-Josas, 17 octobre 2013

Références

Salah Aït-Mokhtar, Jean-Pierre Chanod. Incremental finite-state parsing. *Proceeding of the fifth conference on Applied natural language processing*, Pages 72-79 , Association for Computational Linguistics, 1997.

Alan R. Aronson, François-Michel Lang. An overview of MetaMap: historical perspective and recent advances. *J Am Med Inform Assoc*, 2010.

Sophie Aubin and Thierry Hamon. Improving Term Extraction with Terminological Resources. In *Advances in Natural Language Processing* (5th International Conference on NLP, FinTAL 2006), number 4139 in LNAI Springer:380–7.7. August 2006.

Aubin S., Nazarenko A., Nédellec C.: Adapting a General Parser to a Sublanguage. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP'05)*. Edited by Angelova G, Bontcheva K, Mitkov R, Nicolov N, Nikolov N. Borovets, Bulgaria; 2005:89-93, 2005.

Aussenac-Gilles, N., Després, S., Szulman, S.: The TERMINAE Method and Platform for Ontology Engineering from Texts. In: Buitelaar, P. & Cimiano, P. (eds.): *Bridging the Gap between Text and Knowledge - Selected Contributions to Ontology Learning and Population from Text*, pp. 199--223. IOS Press, 2008.

M. Madan Babu (ed.), *Bacterial Gene Regulation and Transcriptional Networks*, Caister Academic Press, 2013.

Ella Bingham et Heikki Mannila. Random projection in dimensionality reduction: applications to image and text data. Proceedings of the seventh ACM SIGKDD international conference on Knowledge discovery and data mining (KDD '01), pages 245-250, ACM New York, USA 2001.

Steven Bird, Ewan Klein, and Edward Loper, *Natural Language Processing with Python*, O'Reilly Media, 2009.

G. Bisson, C. Grimal, "Apprentissage multi-vue de co-similarités pour la classification", *Conférence Francophone sur l'Apprentissage Automatique*, LORIA, Nancy, 23 - 25 May 2012.

Kalina Bontcheva, Valentin Tablan, Diana Maynard, Hamish Cunningham: Evolving GATE to meet new challenges in language engineering *Journal Natural Language Engineering*, Volume 10 Issue 3-4, pages 349-373, 2004.

Jari Björne, Filip Ginter and Tapio Salakoski, University of Turku in the BioNLP'11 Shared Task. *BMC Bioinformatics*, , 13(Suppl 11):S8, 26, June 2012.

Didier Bourigault: Surface Grammatical Analysis For The Extraction Of Terminological Noun Phrases. *COLING 1992: 977-981*, 1992.

Didier Bourigault, Nathalie Aussenac-Gilles, Jean Charlet: Construction de ressources terminologiques ou ontologiques à partir de textes - Un cadre unificateur pour trois études de cas. *Revue d'Intelligence Artificielle* 18(1): 87-110, 2004.

Brown P. F., Della Pietra V. J., de Souza P. V., Lai J. C. and Mercer R. L. "Class-based n-gram models of natural language.", in *Computational Linguistic* 18(4), pp.283-298, 1992.

Buntine Wray. Generalized Subsumption and Its Application to Induction and Redundancy, *Artificial Intelligence*, Vol. 36, 149–176, 1998.

Maria Teresa Cabré, Terminology: Theory, Methods, and Applications, *Volume 1 de Terminology and Lexicography Research and Practice Series*, Éditeur John Benjamins Publishing Company, 1999.

David Chapman. Planning for conjunctive goals. *Artificial Intelligence* Volume 32, Issue 3, Pages 333–377, July 1987.

Jean Charlet, Sylvie Szulman, Nathalie Aussenac-Gilles, Adeline Nazarenko, Nathalie Hernandez, Nadia Nadah, Éric Sardet, Jean Delahousse, Henry Valéry Téguiak, Audrey Baneyx: DaFOE : une plateforme pour construire des ontologies à partir de textes et de thésaurus. *EGC 2010*: 631-632, 2010.

Stephen Clark and James R. Curran. Wide- coverage efficient statistical parsing with CCG and log-linear models. *Computational Linguistics*, 33(4):493–552, 2007.

Vincent Claveau. IRISA participation to BioNLP-ST 2013: lazy-learning and information retrieval for information extraction tasks. In *Proceedings of the BioNLP 2013 Workshop Companion Volume for Shared Task,* Sofia, Bulgaria. Association for Computational Linguistics, 2013.

Constant Patrick, *Analyse syntaxique par couches*. Thèse de doctorat, ENST, Paris, 1991.

M. Craven & J. Kumlien. Constructing Biological Knowledge Bases by Extracting Information from Text Sources. *Proceedings of the 7th International Conference on Intelligent Systems for Molecular Biology (ISMB'99),* 77-86, AAAI Press, 1999.

Dagan I., Pereira F., and Lee L., "Similarity-Based Estimation of Word Coccurrence Probabilities", in proceedings of the *32nd Annual Meeting of the Association for Computational Linguistics, ACL'94,* New Mexico State University, June 1994.

Dagan, Ido, Lillian Lee and Fernando Pereira. Similarity-based models of cooccurrence probabilities, *Machine Learning*, Vol. 34(1-3) special issue on Natural Language Learning, pp. 43-69, 1999.

Béatrice Daille, Study and Implementation of Combined Techniques for Automatic Extraction of Terminology, In: *The Balancing Act: Combining Symbolic and Statistical Approaches to Language. Workshop at the 32nd Annual Meeting of the ACL (ACL'94),* Las Cruces, New Mexico, USA, 1994.

Daraselia, A. Yuryev, S Egorov, S Novichkova, A Nikitin, and I Mazo. Extracting human pro- tein interactions from MEDLINE using a full-sentence parser. *Bioinformatics*, 20(5):604–611, 2004.

Faure D., *Conception de méthodes d'apprentissage symbolique et automatique pour l'acquisition de cadres de sous-catégorisation de verbes et de connaissances sémantiques à partir de textes : le système Asium*., Thèse en informatique de l'Université Paris-Sud, soutenue le 20 décembre 2000.

Fellbaum, Christiane. WordNet and wordnets. In: Brown, Keith et al. (eds.), *Encyclopedia of Language and Linguistics*, Second Edition, Oxford: Elsevier, 665-670, 2005.

Ronen Feldman. *Probabilistic revision of logical domain theories*. Doctoral Dissertation, Cornell University Ithaca, NY, USA, 1993.

David Ferrucci and Adam Lally. UIMA: an architectural approach to unstructured information processing in the corporate research environment. *Journal Natural Language Engineering*, Volume 10 Issue 3-4, Pages 327 – 348, September 2004

K. Fukuda, T. Tsunoda, A. Tamura, and T. Takagi, "Information extraction: Identifying protein names from biological papers" in *Proceedings of the Pacific Symposium on Biocomputing '98 (PSB'98)*, Hawaii, January 1998.

Gábor, K., Apidianaki, M., Sagot, B., & de La Clergerie, E. V. Boosting the Coverage of a Semantic Lexicon by Automatically Extracted Event Nominalizations. In *LREC 2012-Eighth International Conference on Language Resources and Evaluation*, 2012.

Grefenstette G., *Explorations in Automatic Thesaurus Discovery*, Kluwer Academic (Pub.), 1994.

Ralph Grishman, Beth Sundheim: Message Understanding Conference - 6: A Brief History. In: *Proceedings of the 16th International Conference on Computational Linguistics (COLING)*, Kopenhagen, 466–471, 1996.

Grishman R., Sterling J., "Generalizing Automaticaly Generated Selectional Patterns", in proceedings of the *16th International Conference on Computational Linguistics (COLING'94)*, 1994.

Harris, Z. "Distributional structure". *Word* **10** (23): 146–162, 1954.

Hindle D., "Noun classification from predicate-argument structure", in proceedings of the *28th Annual Meeting of the Association for Computational Linguistics (ACL'90)*, pp. 268-275, Pittsburgh, 1990.

L. Hirschman, R. Grishman. Grammatically-based automatic word class formation. *Inform. Proc. Manag.,* 11, pp. 39–57, 1975.

Christian Jacquemin. A Symbolic and Surgical Acquisition of terms Through Variation. In: S. Wermter, et al. (eds.) *Connectionist, Statistical and Symbolic Approaches to Learning for Natural Language Processing*, pp. 425-438. Springer Verlag, 1996.

Jin-Dong Kim, Sampo Pyysalo, Tomoko Ohta, Robert Bossy, Ngan Nguyen and Jun'ichi Tsujii. Overview of BioNLP Shared Task 2011. In *Proceedings of BioNLP 2011 Workshop*, pages 1-6. Association for Computational Linguistics, 2011.

Yves Kodratoff, Vassilis Moustakis et Nicolas Graner. Can Machine Learning solve my problem? *Applied Artificial Intelligence: An International Journal*, Volume 8, Issue 1, DOI:10.1080/08839519408945431, pages 1-31, 1994.

Martin Krallinger, Alfonso Valencia and Lynette Hirschman. Linking genes to literature: text mining, information extraction, and retrieval applications for biology. *Genome Biology*, 9(Suppl 2):S8, 2008.

Kripke, Saul. *Naming and Necessity*. Boston: Harvard University Press, 1982.

Van Landeghem S., De Bodt S., Drebert Z.J., Inzé D., Van de Peer Y. The potential of text mining in data integration and network biology for plant research: a case study on Arabidopsis. *Plant Cell*. Mar;25(3):794-807, 2013.

Alain-Pierre Manine, Erick Alphonse, and Philippe Bessières, Information extraction as an ontology population task and its application to genic interactions, *20th IEEE Intl. Conf. Tools with Artificial Intelligence, ICTAI'08*., vol. II, pp. 74-81, 2008.

Chris Manning and Hinrich Schütze, *Foundations of Statistical Natural Language Processing*, MIT Press, 1999.

Marcotte E. M., Xenarios I., and Eisenberg D., "Mining litterature for protein-protein interactions", in *Bioinformatics*, vol. 17 n° 4 2001, pp. 359-363, 2001.

T.M. Mitchell, Generalization as Search, *Artificial Intelligence*, Volume 18, No. 2. Also appears in Readings in Artificial Intelligence, Webber and Nilsson (eds.), Tioga Press, 1981, pp. 517-542, 1982.

Yusuke Miyao and Jun'ichi Tsujii. Feature forest models for probabilistic HPSG parsing. *Computational Linguistics*, 34(1):35–80, 2008.

Alain-Pierre Manine, Erick Alphonse, and Philippe Bessières, Learning ontological rules to extract multiple relations of genic interactions from text. *Int. J. Med. Inform*, 2008.

Muggleton, S.; De Raedt, L. "Inductive Logic Programming: Theory and methods". The *Journal of Logic Programming*. 19-20, 1994.

David Nadeau, and Satoshi Sekine. A survey of named entity recognition and classification. *Linguisticae Investigationes* 30(1):3—26, John Benjamins Publishing Company, January 2007.

Natalya F. Noy, Michael Sintek, Stefan Decker, Monica Crubézy, Ray W. Fergerson, and Mark A. Musen. Creating Semantic Web Contents with Protégé-2000. In: *IEEE Intelligent Systems*, vol. 16, pp. 60--71 2001.

Pazienza, M. T., Pennacchiotti, M. and Zanzotto, F. M. Terminology extraction: An analysis of linguistic and statistical approaches. In S. Sirmakessis (eds.), Knowledge mining: Proceedings of the NEMIS 2004 final conference, p. 255–279, Berlin Heidelberg. Springer, 2005.

Barbara Plank and Alessandro Moschitti, Embedding Semantic Similarity in Tree Kernels for Domain Adaptation of Relation Extraction. *The 51st Annual Meeting of the Association for Computational Linguistics (ACL 2013)*, 2013.

Poittevin L., "REVINOS: An Interactive Revision Tool Based on the Concept of Situation", in Proceedings of the *10th European Knowledge Acquisition Modeling and Management Workshop (EKAW'97)*, Plaza E. et Benjamins R. (Ed.), p. 365-370, Springer Verlag, Espagne, octobre 1997.

Poittevin L., "Representing the situations to help the cooperative revision", in Proceedings of the *11th Knowledge Acquisition Workshop* (KAW'98), Gaines B. (Ed.), Banff, Canada, avril 1998.

Sampo Pyysalo, Tapio Salakoski, Sophie Aubin and Adeline Nazarenko. Lexical adaptation of link grammar to the biomedical sublanguage: a comparative evaluation of three approaches. *BMC Bioinformatics*, 7(Suppl 3):S2 doi:10.1186/1471-2105-7-S3-S2, 2006.

Céline Rouveirol, Jean-Francois Puget: Beyond Inversion of Resolution. International *Machine Learning Conference*, 122-130, 1990.

Sammut, C. A., & Banerji, R. B. Learning Concepts by Asking Questions. In R. S. Michalski Carbonell, J.G. and Mitchell, T.M. (Eds.), *Machine Learning: An Artificial Intelligence Approach*, Vol 2. (pp. 167-192). Los Altos, California: Morgan Kaufmann, 1986.

Helmut Schmid. Probabilistic Part-of-Speech Tagging Using Decision Trees. Proceedings of International Conference on New Methods in Language Processing, Manchester, UK, 1994.

Sclano, F., Velardi, P.: TermExtractor: a Web Application to Learn the Shared Terminology of Emergent Web Communities. In: *Proc. of I-ESA*, Portugal, 2007.

Daniel Sleator and Davy Temperley. "Parsing English with a Link Grammar". *Third International Workshop on Parsing Technologies*, 1993.

S. Staab, R. Studer (eds.). Handbook on Ontologies. International Handbooks on Information Systems, Springer Verlag, 2004

Pontus Stenetorp, Stenetorp, Wiktoria Golik, Thierry Hamon, Donald C. Comeau, Rezarta Islamaj Dogan, Haibin Liu, and John W. Wilbur. BioNLP Shared Task 2013: Supporting Resources. In *Proceedings of BioNLP Shared Task 2013 Workshop*, Association for Computational Linguistics. Sofia, Bulgaria, August 2013.

Yoshimasa Tsuruoka and Jun'ichi Tsujii, Bidirectional Inference with the Easiest-First Strategy for Tagging Sequence Data, *Proceedings of HLT/EMNLP* 2005, pp. 467-474, 2005.

Tecuci G. and Kodratoff Y. Apprenticeship Learning in Imperfect Theory Domains, in *Machine Learning: An Artificial Intelligence Approach,* Vol. 3, Y. Kodratoff and R.S. Michalski (Eds.), San Mateo, CA, Morgan Kaufmann, 1990.

Ono Toshihide, Haretsugu Hishigaki, Akira Tanigami and Toshihisa Takagi. Automated extraction of information on protein-protein interactions from the biological literature. *Bioinformatics* 17, 155-161, 2001.

Torre, F. et Rouveirol C. "Natural Ideal Operators in Inductive Logic Programming", inProc. of the 9th *European Conference on Machine Learning*, Springer Verlag, LNCS 1224, pp 274- 290, Prague, 1997.

Verspoor K, Cohn J, Joslyn C, Mniszewski S, Rechtsteiner A, Rocha LM, Simas T. Protein annotation as term categorization in the gene ontology using word proximity networks. *BMC Bioinformatics*. 2005;6 Suppl 1:S20. May 24, 2005.

James Z. Wang, Zhidian Du, Rapeeporn Payattakool, Philip S. Yu, and Chin-Fu Chen. A New Method to Measure the Semantic Similarity of GO Terms. *Bioinformatics.* 23: 1274-1281, 2007.

W3C OWL Working Group, *OWL 2 Web Ontology Language Document Overview* (Second Edition), W3C Recommendation 11 December 2012.

W3C OWL Working Group, *OWL - Web Ontology Language Overview*, W3C Recommendation Deborah L. McGuinness and Frank van Harmelen eds. February 10, 2004.

J. Z. Wang, Z. Du, R. Payattakool, P. S. Yu, and C.-F. Chen, “A new method to measure the semantic similarity of GO terms.,” *Bioinformatics (Oxford, England)*, vol. 23, no. 10, pp. 1274-81, May 2007.

Wille, R. Restructuring lattice theory: an approach based on hierarchies of concepts. In: Rival, I. (ed.) *Ordered Sets*. p. 445-470. Dordrecht-Boston, Reidel, 1982.

Zhiyong Lu et al., The gene normalization task in BioCreative III, The Third BioCreative - Critical Assessment of Information Extraction in Biology Challenge, *BMC Bioinformatics*, Volume 12, Suppl 8, 2011.

Liste des publications et autres documents ou réalisations

Références aux publications

Dans la partie mémoire de ce document, les références aux publications et autres documents ou réalisation sont sous forme du numéro de la référence, précédé de *s* pour séminaire, *r* pour rapport ou *l* pour logiciel. L'absence de lettre indique une publication d'article. Par exemple, [r129] est la référence du rapport numéro 129.

I - PUBLICATIONS SCIENTIFIQUES

I.1 - Articles et communications primaires

I.1.1. Dans périodique à comité de lecture (4)

1. Golik W., Bossy R., Ratkovic R., Nédellec C. " Improving term extraction with linguistic analysis in the biomedical domain" in *Proceedings of the 14th International Conference on Intelligent Text Processing and Computational Linguistics (CICLing'13)*. Special Issue of *International Journal of Computational Linguistics and Applications,* (*IJCLA*) ISSN 0976-0962, www.micai.org/rcs. Présentation orale : 24-30 mars, Samos, Grèce, 2013. A paraître.
2. Bossy R., Jourde J., Manine A.-P., Veber P., Alphonse E., van de Guchte M., Bessières P., Nédellec C. "BioNLP Shared Task - The Bacteria Track". BMC Bioinformatics 13(Suppl 11):S3, 26 juin 2012. Impact factor 2012 : 2,75.
3. Nédellec C., "Integration of Machine Learning and Knowledge Acquisition" in *Knowledge Engineering Review*, vol 10(1), p. 77-81, Fox J. (Ed.), Cambridge University Press, mars 1995. Impact factor 2009 : 1.143.
4. Nédellec C., Ferreira J. L., Correia J. et Costa E., "Machine Learning Goes to the Bank", in *Applied Artificial Intelligence - An International Journal*, Special Issue on Machine Learning, vol 8(4), p. 593-615, Kodratoff Y. (Ed.), Taylor & Francis Publishers, octobre 1994. Impact factor 2009 : 0,988.

I.1.2. Invité dans périodique avec comité de lecture (critique) (1)

5. Nédellec C., critique du livre intitulé "Inductive Logic Programming: From Machine Learning to Software Engineering", de Bergadano F. et Gunetti D., MIT Press, in *Knowledge Engineering Review,* Fox J. (Ed.), Cambridge University Press, 1997. Impact factor : 2009 : 1.143.

I.1.3. Rapports diplômants (D.E.A., thèse...) (2)

6. Nédellec C., *APT, apprentissage interactif de règles de résolution de problème en présence de théorie du domaine*, thèse en informatique de l'Université Paris-Sud. 300 pages. Rapporteurs : J. F. Perrot, (LIP6, Université Paris VI, Paris) et F. Bergadano, (Université de Turin).
7. Nédellec C., *Généralisation interactive à petit pas pour contrôler la surgénéralisation*, Mémoire de DEA IARFA (Intelligence artificielle, reconnaissance des formes et applications), 96 pages, Université Paris 6, septembre 1990.

I.1.4. Communications dans congrès (63)

I.1.4.a Colloques internationaux avec comité de lecture (texte intégral) (11)

8. Golik W., Dameron O., Bugeon J., Fatet A., Hue I., Hurtaud C., Reichstadt M., Salaün M.-C., Vernet J., Joret L., Papazian F., Nédellec C. et Le Bail P.-Y. "ATOL: the multi-species livestock trait ontology" in proceedings of *The 6th Metadata and Semantics Research Conference (MTSR 2012),* pages 289-300. Springer Verlag Communications in Computer and Information Science Serie. Cadiz, Espagne, 28 au 30 novembre 2012. DOI: 10.1007/978-3-642-35233-1_28
9. Nédellec C., Golik W., Aubin S., Bossy R., "Building Large Lexicalized Ontologies from Text: a Use Case in Indexing Biotechnology Patents ", *International Conference on Knowledge Engineering and Knowledge Management (EKAW 2010),* Springer 2010 Lecture Notes in Computer Science ISBN 978-3-642-16437-8. Pages 514-523, Lisbonne, Portugal, 11 au 15 octobre, 2010.
10. Galibert O., Quintard L., Rosset S., Zweigenbaum P., Nédellec C., Aubin S., Gillard L., Raysz J.-P., Pois D., Tannier X., Deléger L., Laurent D., « Named and specific entity detection in varied data: The Quæro Named Entity baseline evaluation », *The seventh international conference on Language Resources and Evaluation (LREC-2010*), European Language Resources Association (ELRA) publisher, pages 3453-3458, Malte, 19-21 mai 2010.

11. Aubin S., Nazarenko A. and Nédellec C. "Adapting a General Parser to a Sublanguage", *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP'05)*. Pages 89-93. Borovets, Bulgarie, 2005.
12. Nédellec C., Ould Abdel Vetah M., et Bessières P., "Sentence Filtering for Information Extraction in Genomics, a Classification Problem". In Proceedings of the 5th *European conference on Practical Knowledge Discovery in Databases*, PKDD'2001, Lecture notes in computer science. Heidelberg : Springer - Verlag pages 326-338, Freiburg, septembre 2001.
13. Faure D. et Nédellec C.,"Knowledge Acquisition of Predicate-Argument Structures from technichal Texts using Machine Learning" in Proceedings of *Current Developments in Knowledge Acquisition: EKAW-99*, p. 329-334, Fensel D. et Studer R. (Ed.), Springer Verlag, Karlsruhe, Allemagne, avril 1999.
14. Nédellec C., "Cooperation with a Learning Apprentice", in Proceedings of the *First International Conference on Applied Ergonomics (ICAE'96),* Özok A. F. (Ed.), p. 497-503, USA Publishing, Istanbul, mai 1996.
15. Feldman R. et Nédellec C., "A Framework for Specifying Explicit Bias for Revision of Approximate Knowledge Bases", in Proceedings of the 7th *Knowledge Acquisition Workshop, KAW-94*, p. 15.1-19, Gaines B. et Musen M. (Eds.), Banff, janvier 1994.
16. Nédellec C. et Causse K., "Knowledge Refinement with KA and ML Methods", in Proceedings of *5th Current Developments in Knowledge Acquisition: EKAW-92*, p. 171-190, Wetter T. et al. (Eds.), Springer-Verlag, mai 1992.
17. Nédellec C., "How to specialize by Theory Refinement", in Proceedings of the *10th European Conference on Artificial Intelligence (ECAI-92)*, p. 474-478, Neuman B. (Ed.), John Wiley & sons (Pub.), Vienne, août 1992.
18. Nédellec C., "A Smallest Generalization Step Strategy" in Proceedings of the *Eight International Workshop on Machine Learning (IWML 91)*, p. 529-533, Evanston (USA), Birnbaum L. A. et Collins G. C. (Eds.), Morgan Kaufman, juin 1991.

1.1.4.b Workshops internationaux avec comité de lecture (texte intégral) (19)

19. Hue I , Bugeon J Dameron O, Fatet A, Hurtaud C, Joret L, Meunier-Salaün MC, Nédellec C, Reichstadt M, Vernet J, Le Bail PY. ATOL AND EOL ONTOLOGIES, STEPS TOWARDS EMBRYONIC PHENOTYPES SHARED WORLDWIDE?, 4th Mammalian Embryo Genomics meeting, Québec, octobre 2013. Soumis.
20. Bossy R., Bessières P., Nédellec C. BioNLP Shared Task 2013 – An overview of the Genic Regulation Network Task. In *Proceedings of the BioNLP 2013 Workshop*, Association for Computational Linguistics, pages 23-30. Sofia, Bulgaria, 2013.
21. Bossy R., Golik W., Ratkovic Z., Bessières P., Nédellec C.. BioNLP shared Task 2013 – An Overview of the Bacteria Biotope Task. In *Proceedings of the BioNLP 2013 Workshop*, Association for Computational Linguistics, pages 74-82. Sofia, Bulgaria, 2013.
22. Nédellec C., Bossy R., Kim J.-D., Kim J.-j., Ohta T., Pyysalo S., Zweigenbaum P. Overview of BioNLP Shared Task 2013. In *Proceedings of the BioNLP 2013 Workshop*, Association for Computational Linguistics, pages 1-7. Sofia, Bulgaria, 2013.
23. Papazian F., Bossy R. et Nédellec C., "AlvisAE: a collaborative Web text annotation editor for knowledge acquisition", *The 6th Linguistic Annotation Workshop (The LAW VI*), LAW VI '12 Proceedings of the Sixth Linguistic Annotation Workshop. Stroudsburg (USA) : Association for Computational Linguistics 2012. Pages 149-152, Jeju, Corée, juillet 2012.
24. Ratkovic Z., Golik W., Warnier P., Veber P., Nédellec C., "BioNLP 2011 Task Bacteria Biotope – The Alvis system", *BioNLP workshop associé à ACL,* The 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies, Workshop. Madison (USA) : Omnipress. Pages 102-111. Portland, Etats-Unis, 2011.

25. Bossy R., Jourde J., Bessières p., van de Guchte M., Nédellec C., « BioNLP shared Tasks 2011 - Bacteria Biotope », *BioNLP workshop associé à ACL,* The 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies, Workshop. Madison (USA) : Omnipress. Pages 56-64. Portland, Etats-Unis, 2011.
26. Warnier P., Nédellec C., "Sentence Filtering for BioNLP: Searching for Renaming Acts" *BioNLP workshop* associé à ACL, The 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies, Workshop. Madison (USA) : Omnipress. Pages 121-129 Portland, Etats-Unis, 2011.
27. R. Bossy, A. Kotoujansky, S. Aubin and C. Nédellec, "Close Integration of ML and NLP Tools in BioAlvis for Semantic Search in Bacteriology" , Proceedings of the *Workshop on Semantic Web Applications and Tools for Life Sciences*, A. Burger, A. Paschke, P. Romano, A. Splendiani (Eds), CEUR-WS, Vol-435, 15 pages, Edinbourg, Royaume Uni, 28 novembre, 2008. http://www.swat4ls.org/2008/
28. Bessières P., Bossy R., Mannine A.-P., Alphonse E., Nédellec C., "Getting the Unknown from the Known in Bacteria, and the Role of Text Mining", In Proceedings of the *Data and text mining in integrative biology workshop*, associé à la conférence ECML/PKDD, M. Hilario et C. Nédellec (Eds), pages 67-72, Berlin, Allemagne, septembre 2006.
29. Nédellec C., Bessières P., Bossy R., Kotoujansky A., Manine A.-P., "Annotation Guidelines for Machine Learning-Based Named Entity Recognition in Microbiology", In Proceedings of the *Data and text mining in integrative biology workshop*, associé à ECML/PKDD, M. Hilario et C. Nédellec (Eds), pages 40-54, Berlin, Allemagne, septembre 2006.
30. Nazarenko A., Nédellec C., Alphonse E., Aubin S., Hamon T., Manine A.-P., "Semantic Annotation in the Alvis Project", in *Proceedings of the International Workshop on Intelligent Information Access*, W. Buntine et H. Tirri (eds), 4 pages, Helsinki, Finlande, juillet 2006. http://cosco.hiit.fi/search/IIIA2006/ http://videolectures.net/iiia06_nedellec_saap/
31. Alphonse E., Aubin S., Bessières P., Bisson G., Hamon T., Lagarrigue S., Manine A.-P., Nazarenko A., Nédellec C., Ould Abdel Vetah M., Poibeau T., Weissenbacher D., Event-based information extraction for the biomedical domain: the Caderige project. In *Proceedings of International Joint workshop on* Natural *Language Processing in Biomedicine and its Applications (NLPBA/BioNLP (associé à COLING'04)*, Ruch P., Collier N. and Nazarenko A. (eds.), New York (USA) : ACM - Association for Computing Machinery, pages 43-49. Genève, Suisse, 2004.
32. Bisson G., Nédellec C. et Canamero D. "Designing clustering methods for ontology building - The Mo'K workbench" in Proceedings of the *workshop on Ontology Learning, workshop of the European Conference on Artificial Intelligence (ECAI-2000),* Staab S. et al (Eds)., pages 13-19, Berlin, août 2000.
33. Nédellec C., "Corpus-based learning of semantic relations by the ILP system, Asium" in proceedings of *Learning Language in Logic workshop*, co-located with ICML'99, pages 28-39, Cussens J. (Ed.), Bled, Slovénie, juin 1999.
34. Faure D. et Nédellec C.,"ASIUM: Learning subcategorization frames and restrictions of selection." in Proceedings of the *Text Mining* workshop, 10th European Conference on Machine Learning (ECML 98), Kodratoff Y. (Ed.), 11 pages, Chemnitz, Allemagne, avril 1998.
35. Faure D. et Nédellec C., "A Corpus-based Conceptual Clustering Method for Verb Frames and Ontology Acquisition " in Proceedings of *Adapting lexical and corpus resources to sublanguages and applications,* workshop of the 1st *International Conference on Language resources and Evaluation (LREC)*, p. 1-8, Velardi P. (Ed.), Grenade, Espagne, mai 1998.
36. Nédellec C., "APT, a cooperative ML System", in Proceedings of the *AAAI'96 Spring Symposium on Knowledge Acquisition and Learning by Demonstration* , p. 97-103, Cypher A., Gil Y., Pazzani M. (Eds.), AAAI report, Stanford, mars 1996.
37. Nédellec C. et Rouveirol C., "Hypothesis Selection Biases for Incremental Learning", in

Proceedings of the *AAAI Spring Symposium on Training Issues in Incremental Learning*, A. Cornuejols (Ed.), AAAI Press, Menlo Park, pages 109-118, mars 1993.

38. Nédellec C., "Knowledge Refinement Using Knowledge Acquisition and Machine Learning Methods", in Working Notes of the *AAAI Spring Symposium on Cognitive Aspects of Knowledge Acquisition,* Gaines B. (Ed.), Stanford University, mars 1992.

1.1.4.c Articles invités dans les actes de workshops internationaux (9)

39. Nédellec C., Bossy R., Kotoujansky A., "Alvis Semantic Search Engine, Adaptation to Biology », Abstract et Conférence invitée à *Machine Learning for Natural Language Processing workshop*., Université d'Amsterdam, mai 2007.
40. Nédellec C. "Learning Language in Logic - Genic Interaction Extraction Challenge" in *Proceedings of the Learning Language in Logic* (LLL05) workshop joint to ICML'05. Cussens J. and Nédellec C. (eds). Pages 31-37, Bonn, août 2005.
41. Nédellec C., Machine Learning and Natural Language Processing for Textual Information Annotation in Genomics", 12 pages. At *Dagstuhl Seminar on Machine Learning for the Semantic Web*, février 2005.
42. Nédellec C., "Machine Learning Applied to Information Extraction in specific domains. An Example: gene interaction extraction from bibliography in genomics.", communication invitée. In *Proceedings of the 2nd ECML/PKDD'2002 Workshop on Semantic Web Mining*, Berendt B. et al. (Eds), p. 1-7, Helsinki, Finlande, août 2002. http://km.aifb.kit.edu/ws/semwebmine2002/.
43. Nédellec C., "Knowledge Extraction from Text, a Machine Learning Approach", communication invitée. In Proceedings of the *Third International Conference on Human-System Learning, CAPS'3, Learning WWW*, 7 pages, Europia Production (Pub.), Paris, France, décembre 2000.
44. Nédellec C., "Modeling Machine Learning and Comprehensibility" in Proceedings of the *IJCAI'95 workshop on Machine Learning and Comprehensibility*, C. Nédellec et Y. Kodratoff (eds), p. 15-18, Montréal, août 1995.
45. Nédellec C., *Integration of Machine Learning and Knowledge Acquisition*, in Actes du workshop *Integration of Machine Learning and Knowledge Acquisition* associé à la 11th European Conference on Artificial Intelligence (ECAI-94), C. Nédellec et Y. Kodratoff (eds), 6 pages, août 1994.
46. Nédellec C. et Rouveirol C., "Shift of Bias for Incremental Learning", in Proceedings of the *MLnet workshop on MultiStrategy Learning,* Saitta L. (Ed.), 3 pages, Blanès, Espagne, septembre 1993.
47. Nédellec C. et Canamero D. "Learning and Problem-Solving with APT", in Proceedings of the *MLnet workshop on Problem-Solving and Machine Learning*, Van Someren M. (Ed.), 3 pages, Blanès Espagne, septembre 1993.

1.1.4.d Workshops internationaux avec comité de lecture (abstract et poster ou démonstration) (2)

48. Salaun, M.-C., Bugeon, J., Dameron, O., Fatet, A., Hue, I., Hurtaud, C., Nédellec, C., Reichstadt, M., Vernet, J., Reecy, J., Park, C., Le Bail, P.-Y. ATOL: an ontology for livestock. In : Book of abstracts of the 63rd Annual Meeting of the European Federation of Animal Science, Bratislava (Slovaquie).Wageningen (NLD) : Wageningen Academic Publishers (EAAP Book of Abstracts, 18), page 299, août 2012.
49. Aubin S., Bessières P., Bossy R., Gillard L., Jourde J., Papazian F., Veber P., Nédellec C., "BioAlvis II, NLP-based semantic mining of literature on molecular biology of bacteria", p. 41, Poster et abstract. *Workshop BioCreative II.5*, B. Bañeres *et al*. (eds), CNIO Madrid, 7 au 9 octobre 2009.http://www.biocreative.org/media/store/files/2009/Proceedings.pdf

1.1.4.e Colloques nationaux avec comité de lecture (15)

50. Golik W., Warnier P., Nédellec C. "Corpus-based extension of termino-ontology by linguistic analysis: a use case in biomedical event extraction. " *Ontology and Lexicon: new insights. Actes*

du workshop TIA 2011 : 9th International Conference on Terminology and Artificial Intelligence, M. Slodzian et al., (eds), pages 37-39, Paris, novembre 2011.

51. Alphonse E., Aubin S., Bessières P., Bisson G., Hamon T., Lagarrigue S., Nazarenko A., Nédellec C., Ould Abdel Vetah M., Poibeau T., Weissenbacher D., "Extraction d'information appliquée au domaine biomédical - apprentissage et traitement automatique de la langue", in actes de la *Conférence International de Fouille de texte, CIFT-04*, Antoni M.-H. et Yvon F. (Eds), pages 7-19, La Rochelle, juin 2004.
52. Manine A. P., Nédellec C. et Bessières P. (2004). "Application of Machine Learning to knowledge extraction from bibliographical abstracts in functional genomics", *Journées post-génomique de la Doua, JPGD*, 2004.
53. Abdel Vetah M. O., Nédellec C., Bessières P., Caropreso F., Manine A.-P., Matwin S., "Sentence Categorization in Genomics Bibliography: a Naïve Bayes Approach", in actes de la *Journée informatique et transcriptome*, associée aux journées Post-Génomique de la Doua (JPGD), J.- F. Boulicaut et O. Gandrillon (Eds), 11 pages, Lyon, avril 2003.
54. Ould Abdel Vetah M., Nédellec C. et Bessières P., "Application de la classification supervisée au filtrage de phrase mentionnant des interactions géniques dans MedLine", *Journées Ouvertes Biologie Mathématique (JOBIM'2002),* J. Nicolas & C. Thèrmes (éds), p. 337-341, Saint-Malo, juin 2002.
55. Bessières P., Nazarenko N. et Nédellec C., "Apport de l'apprentissage à l'extraction d'information : le problème de l'identification d'interactions géniques.", in *Actes du Colloque International sur le Document Electronique, Méthodes, Démarches et Techniques Cognitives, CIDE'2001*, Toulouse, p. 165-183, octobre 2001.
56. Bisson G. et Nédellec C., "Aide à la conception de méthodes de classification pour la construction d'ontologies : l'atelier Mo'K" in *Actes des Journées Francophones d'Extraction et de Gestion des Connaissances (EGC'2001),* Briand H. (Ed.), Hermès (Pub.), p. 213-225, Nantes, janvier 2001.
57. Nédellec C., Ould Abdel Vetah M., Bessières P., Brun C. et Jacq B., "Reconnaître des fragments de phrase pertinents pour l'extraction d'information dans les textes de génomiques, un problème de classification". In *Actes de la Conférence Francophone d'Apprentissage (CAP)*, Bisson G. (Ed.), 12 pages, Grenoble, juin 2001.
58. Faure D., Nédellec C. et Rouveirol C., "Acquisition de connaissances sémantiques par des méthodes d'apprentissage, le système ASIUM" in Actes des 13èmes *Journées Françaises d'Apprentissage (JFA'98),* Nicolas J. (Ed.), 12 pages, Lens, mai 1998.
59. Faure D., Nédellec C. et Rouveirol C., "Apprentissage automatique de schémas de sous-catégorisation et de restrictions sémantiques à partir de textes" in Actes de la *5ème conférence annuelle sur le traitement automatique des langues naturelles (TALN'98),* pages 233-235, Zweigenbaum P. (Ed.), Paris, France, juin 1998.
60. Nédellec C., "Configurer différents opérateurs pour contrôler l'exploration d'une base de données relationnelle" in Actes des *12èmes Journées Francophones d'Apprentissage Automatique (JFA-97),* Soldano S. (Ed.), 15 pages, Roscoff, mai 1997.
61. Nédellec C., "Expliquer pour valider les systèmes d'apprentissage automatique" in *Actes des Journées Acquisition des Connaissances du PRC-GDR-IA (JAC-93),* Reynaud C. (Ed.), 13 pages, Saint-Raphaël, mars 1993.
62. Nédellec C., "Un Système d'Apprentissage s'explique", in *Actes de la Journée "Explication et Coopération Homme-Machine, vers la co-construction d'explications" (JEEX),* Karsenty L. et Brézillon P. (Eds), 13 pages, Paris, Rapport CNAM n°105, juin 1993.
63. Rouveirol C. et Nédellec C., "Les Biais dans la Révision Interactive de Théories" in Actes des *7èmes Journées Apprentissage du PRC-GDR-IA (JFA-93),* de Sainte Marie C. (Ed.), 16 pages, Saint-Raphaël, mars 1993.
64. Nédellec C., "Une stratégie de généralisation à petits pas pour traiter la surgénéralisation" in Actes

des *5èmes Journées Françaises de l'Apprentissage (JFA-91),* p. 128-144, Quinqueton J. (Ed.), pages 128-144, Sète, mai 1991.

1.1.4.f Article invité dans les actes d'ateliers nationaux (1)

65. Nédellec C. et Nazarenko A., "Application de l'apprentissage à la recherche et à l'extraction d'information - Un exemple, le projet *Caderige* : identification d'interactions géniques.". Communication invitée. 12 pages. In Actes de la *Journée thématique Exploration de données issues d'Internet*, organisé par Bennani Y., Janvier E., Kanawati R. et Salotti S., LIPN-CNRS, Institut Galilée, Université Paris 13, Villetaneuse, mars 2001.

1.1.4.g Ateliers nationaux avec comité de lecture (abstract et poster ou démonstration) (6)

66. M. C. Meunier-Salaün, J. Bugeon, O. Dameron, A. Fatet, I. Hue, C. Hurtaud, L. Joret, C. Nédellec, M. Reichstadt, J. Vernet, PY Le Bail., *Les ontologies ATOL et /EOL: des outils en appui aux nouveaux challenges en production porcine : phénotypage et élevage de précision*, Journées de la Recherche Porcine (JRP), à paraître, 4 et 5 février 2014.
67. Bossy R., Papazian F. et Nédellec C. "AlvisAE, un éditeur d'annotation sémantique structurée guidée par une ontologie évolutive. " *Actes du workshop IC 2013 In_Ovive, Intégration de sources/masses de données hétérogènes et Ontologies, dans le domaine des sciences du VIVant et de l'Environnement*, Plateforme AFIA, J. Dibie-Barthélémy et al., (eds), 1 page. Lille, juillet 2013.
68. Aubin S., Golik, Papazian F., Nédellec C., « TyDI : un outil d'assistance à la construction de terminologie structurée », poster et démonstration à *TIA'09, 8ème Conférence Internationale Terminologie et Intelligence Artificielle*, 2 pages, Toulouse, 18-20 octobre, 2009.
69. Nédellec C. et Ould Abdel Vetah M.,"Modélisation des interactions géniques à partir de textes.", 1 page. In actes des *Journée Post-Génomique de la Doua (JPGD)*, Lyon, avril 2001.
70. Bessières P., Bisson G., Nazarenko A., Nédellec C., Ould Abdel Vetah M. et Poibeau T, "Ontology learning for information extraction in genomics bibliography - the Caderige project". 1 page. In actes des *Journées IMPG (Action Informatique, Mathématique et Physique pour la Génomique) Ontologie et Extraction d'Information en Génomique*, Grenoble, mai 2001.
71. Nédellec C. et Ould Abdel Vetah M., Beissières P., Brun C. et Jacq B., "Text filtering for information extraction in genomics, a classification problem". 1 page. In actes des *journées IMPG (Action Informatique, Mathématique et Physique pour la Génomique) Ontologie et Extraction d'Information en Génomique*, Grenoble, mai 2001.

I.1.5. Autres supports.

I.1.5.a Chapitre d'ouvrage (4)

72. Nédellec C., "Corpus-based learning of semantic relations by the ILP system, Asium" in *Learning Language in Logic*, Cussens J. and Dzeroski S. (Eds.), Springer Verlag, Lecture Notes in Computer Science Volume 1925, pages 259-278, septembre 2000.
73. Nédellec, C. and Tiberghien, A., "Teaching the Energy Concept with the Machine Learning System APT." In D. Kayser & S. Vosniadou (Eds.) *Modelling Changes in Understanding. Case Studies in Physical Reasoning*. Oxford : Pergamon in association with the European Association for Research on Learning (EARLI) Elsevier Science. pp.192-222, 1999.
74. Nédellec C., Rouveirol C., Adé H., Bergadano F. et Tausend B.,"Declarative Bias in ILP", in *Advances in Inductive Logic Programming*, p. 82-103, De Raedt L. (Ed.), IOS Press, 1996.
75. Saitta, L., Neri, F., Bajo, M. T., Cañas, J., Chaiklin, S., Esposito, F., Kayser, D., Nédellec, C., Sabah, G., Tiberghien, A., Vergnaud, G., Vosniadou, S. Knowledge representation changes in humans and machines. In P. Reimann, et H. Spada (Eds.), *Learning in humans and machines: Towards and interdisciplinary learning science*, pages. 109–127. New York: Pergamon, 1995.

I.1.5.b Communications invitées orales (50)

Communications invitées dans des congrès et colloques internationaux (8)

76. “Extraction of gene interaction networks from texts : the Caderige Project", présentation avec G. Bisson de l'INRIA, *XRCE - INRIA workshop on Information Extraction in Genomics*, Grenoble, juin 2000.
77. "Knowledge extraction from text, a machine learning approach", *CAPS'3, 3rd International Conference on Human-System Learning*, , décembre 2000.
78. “Machine Learning applied to Information Extraction in specific domains - an example, gene interaction extraction from bibliography in genomics”, *Semantic Web Mining*, 2nd Workshop at ECML/PKDD, Helsinki, 20 août 2002.
79. “Towards the automatization of knowledge extraction from text in genomics, strength and limits”, NEMIS 2003 Launch Conference, 1st International Workshop on Text Mining & its Applications, Patras, Grèce, 5 avril 2003.
80. “Ontology Learning and Populating in Genomics”, *Ontology population*, ECAI-2004 Workshop, 2004
81. "ML and NLP for Textual Information Annotation in Genomics", at *Dagstuhl Seminar on Machine Learning for the Semantic Web*, février 2005.
82. "Machine Learning for semantic annotation of textual data in genomics", at *15th Amsterdam Colloquium*, décembre 2005.
83. "Alvis Semantic Search Engine, Adaptation to Biology ”, présentation Nédellec C., Bossy R., Kotoujansky A., *Machine Learning for Natural Language Processing workshop*, Université d’Amsterdam, mai 2007, Pays-Bas.

Communications invitées dans des congrès et colloques nationaux (5)

84. “Une Application bancaire de l’Apprentissage Symbolique”, *Quatrièmes Journées Francophones sur l’Induction Symbolique et Numérique*, Université Paris-Sud, France, mars 1994.
85. “Le système d'apprentissage APT et la modélisation cognitives de phénomènes physiques”, *journées de l'Institut des Sciences Cognitives et de la Communication*, Université Paris-Sud, janvier 1996.
86. "L'apprentissage (semi)-automatique de connaissances à partir de corpus – avec des ontologies, vers des ontologies.", *Atelier OntoTexte associé à TIA* (Rencontres Terminologie et Intelligence Artificielle), Sophia-Antipolis, France, 10 octobre, 2007.
87. "Apprentissage automatique pour l’analyse de contenu textuel en biologie", Journée *Apprentissage Artificiel: la carte, le territoire et l'horizon*, en l'honneur de Yves Kodratoff, INA-PG, Paris, 23 mai, 2008.
88. "Ontologies et recherche d’informations dans le domaine scientifique", Journée *Comunication technique de plein champ : recherche et innovation*, organisée par Patricia Minacori, Université Paris Diderot, Paris, 26 Mars 2011.

Communications invitées à des séminaires internationaux (8)

89. “Cooperative Machine Learning”, au *Laboratoire d'Informatique de l'Université de Turin*, Italie, Juillet 1994.
90. “Acquisition of semantic knowledge from specification texts by ML methods, the ASIUM system" au *Laboratoire d'Informatique de l'Université de Turin*, Italie, novembre 1997.
91. « *Some* accomplishments and challenges in *text-mining* for biology, Visit of *Greifswald Transcriptomics – Functional Genomics group*, Jouy-en-Josas, 26 January 2009.
92. Bessières P., Nédellec C. and Bibliome group « Bibliome for genome annotation », visit to *Mikrobiologie Institut für Mikrobiologie und Genetik*, University of Göttingen, 10 décembre 2009.
93. « Bibliome, text content analysis », Présentation à la délégation de l’université de Novosibirsk, MIG, 5 octobre 2009.

94. Claire Nédellec et Bibliome group « Text Mining at Bibliome », DSMZ (Deutsche Sammlung von Mikroorganismen und Zellkulturen GmbH), 4 octobre 2011, Braunschweig, Allemagne.
95. "Semantic search of Animal Journal ", *Management Board of Animal Journal* (CUP, INRA, BSAS, EAAP), Edimbourg, 19 mai 2011.
96. Claire Nédellec, « Semantic search with the ATOL ontology » Présentation à la délégation du Leibnitz Institute for Farm Animal Biology (FBN Dummerstorf), séminaire ATOL, Rennes, 5 mars 2013.

Communications invitées à des séminaires académiques nationaux (6)

97. "Acquisition de connaissances sémantiques à partir de textes de spécifications par apprentissage automatique, le système ASIUM", *Xerox Research Center*, Grenoble, janvier 1998.
98. "Apprentissage et traitement sémantique du langage naturel", *LIPN*, Villetaneuse, mai 1998.
99. "Acquisition d'ontologies à partir de textes de spécialité, application aux brevets", *LORIA*, Nancy, octobre 1998.
100. "Application de l'apprentissage automatique à l'acquisition de connaissances à partir de corpus", *LIMSI*, Orsay, décembre 1998.
101. "Annotation sémantique de documents en biologie", INIST, Nancy, octobre 2006.
102. "Acquisition de connaissances linguistiques par apprentissage automatique – application à la recherche documentaire", *Séminaire du LIG*, Université Joseph Fourier, Grenoble, 22 janvier 2010.

Séminaires invités à des journées thématiques (4)

103. "Application de l'apprentissage à la recherche et à l'extraction d'information - Un exemple, le projet *Caderige* : identification d'interactions géniques.", présentation avec A. Nazarenko du LIPN, *Journée thèmatique Exploration de données issues d'Internet*, organisé par Y. Bennani, E. Janvier, R. Kanawati et S. Salotti, LIPN-CNRS, Institut Galilée, Université Paris 13, Villetaneuse, 2 mars 2001.
104. Nédellec C. et Ould Abdel Vetah M., Bessières P., Brun C. et Jacq B., "Text filtering for information extraction in genomics, a classification problem". *Journées IMPG Ontologie et Extraction d'Information en Génomique*, Grenoble, mai 2001.
105. Bessières P., Bisson G., Nazarenko A., Nédellec C., Ould Abdel Vetah M. et Poibeau T, "Ontology learning for information extraction in genomics bibliography - the Caderige project", *Journées IMPG Ontologie et Extraction d'Information en Génomique*, Grenoble, mai 2001.
106. Nédellec C. et Ould Abdel Vetah M.,"Modélisation des interactions géniques à partir de textes.", *Journée Post-Génomique de la Doua* (JPGD), Lyon, avril 2001.

Séminaires invités à un groupe de recherche national (5)

107. "Explications et Apprentissage Symbolique", *Groupe PRC-IA Explication*, Paris, avril 1994.
108. "Apprentissage par PLI dans les Bases de Données Relationnelles", *Groupe AFCET-AFIA, GTRA*, Nantes, novembre 1997.
109. "Acquisition des Connaissances et Apprentissage", *Journée Acquisition des Connaissances et Apprentissage pour l'interprétation des Signaux et des Images* organisée par l'Action PRC 10.2, Paris, novembre 1997.
110. "Apprentissage de schémas de sous-catégorisation et d'ontologies", Journée d'étude Extraction d'Information (T. Poibeau), séminaire de *l'Association pour le Traitement Automatique des Langues* (ATALA), Paris, mars 1999.
111. "Extraction d'information à partir de notices sur les interactions géniques", *Séminaire IMPG, région Ouest*, IRISA, Rennes, avril 2001.

Séminaires INRA (11)

112. "Des technologies sémantiques pour la construction et l'exploration de corpus documentaires", journées DEPE (Direction de l'Expertise, de la Prospectives et des Etudes), 28 juin 2013.

113. "Extraction d'information à partir de texte", journées ModalTub, Réseau *Elucider le devenir de l'aliment dans le Tube Digestif*, 21 mars 2013.

114. Claire Nédellec, Maarten van de Guchte, Philippe Bessières, "Réseau OntoBiotope", journées Animation *MEM Days*, 31 janvier 2012.

115. "Activités INRA sur l'extraction d'information sur les biotopes", journées Réseau OntoBiotope, 31 novembre 2012.

116. Claire Nédellec, Maarten van de Guchte, Philippe Bessières, "Réseau OntoBiotope", journées Animation du Metaprogramme MEM Days, 31 janvier 2012.

117. « Extraction de régulations géniques à partir d'articles", Séminaire du réseau MIA Inférence de réseau, Paris, 9 février 2012.

118. « Gestion des connaissances - Ontologies", Séminaire stratégique CEPIA (Département Caractérisation et Elaboration des Produits issus de l'agriculture), Paris, 17 mars 2010. https://intranet.inra.fr/cepia/vie_collective/seminaires_du_departement/seminaire_cepia_2010.

119. "Modélisation de connaissance termino-ontologique application à la recherche documentaire", séminaire Modélisation du département Phase, Physiologie Animale, Ecully, 9 mars 2010.

120. « Conception d'ontologie et recherche sémantique dans le domaine scientifique", Séminaire du groupe de travail ATO-Phase (Département Physiologie Animale), Rennes, 3 novembre 2010.

121. "Acquisition automatique d'ontologie pour l'annotation sémantique de texte en biologie", séminaire de l'unité INRA, MetaRisk, février 2006.

122. "Recherche sémantique d'information textuelle", Claire Nédellec, Pascale Sébillot, Philippe Bessières, Vincent Claveau, Journées INRA-INRIA, 30 mai 2007.

Séminaires publics de projet transversal (2)

123. Claire Nédellec, Robert Bossy, Alain Kotoujansky, Annick Lacombe, Philippe Bessières. « Alvis Semantic Search Engine – Adaptation to Patent Search ». Revue annuelle du projet Epipagri, septembre 2007.

124. *Bibliome content analysis technology for system biology* Micalis (Microbiology IdF) project. Jouy-en-Josas, 16 février 2009.

I.1.5.c Rapports de recherche (5)

125. Dialekti Valsamou, Bertrand Dubreucq, Loïc Lepiniec, Claire Nédellec,. *Annotation Guidelines. Campaign on Scientific Literature on* Arabidopsis thaliana. 2013.

126. C. Nédellec, A. Nazarenko, *Ontologies and Information Extraction* Rapport LIPN, hal-00098068, 2005.

127. C. Nédellec, C. Rouveirol. *Specifications of the HAIKU system*, Rapport de Recherche n° 928 de l'Université Paris-Sud, septembre 1994.

128. C. Nédellec, C. Rouveirol. *Biases in interactive theory revision*. Rapport de Recherche n° 797 de l'Université Paris-Sud, décembre 1992.

129. C. Nédellec, *Smallest generalization step strategy to deal with over generalization*, Rapport de recherche n° 626, Université de Paris-Sud, 1990.

I.1.5.d Rapports de projet institutionnel de recherche

Rapports de recherche du projet MEM OntoBiotope

130. Robert Bossy et Claire Nédellec, *Consignes d'annotation des biotopes des bactéries et archaea*. Projet OntoBiotope, tâche Bacteria Biotope de BioNLP-ST'13. novembre 2012.

Rapports de recherche du projet FSOV Sam Blé

131. Claire Nédellec, Robert Bossy, Marion Ranoux, Pierre Sourdille, Dialekti Valsamou, *Consignes d'annotation d'articles sur la sélection du blé par marqueurs génétiques*. Projet FSOV Sam Blé. Avril 2013.

Principaux rapports de recherche du projet Quaero (145)
(75 delivrables contractuels et 60 rapports d'activités)

132. Coordinateur et co-auteur. Claire Nédellec et al., *WP3 Natural Language Processing progress report - period 1,* CD.CTC.3.1_NaturalLanguageProc_V1_0, oct 2008.

133. Coordinateur des 12 livrables CD.CTC.3.30 à CD.CTC.3.43 des deux tâches *Ontology Acquisition* et *Semantic Annotation*.

134. Co-auteur des livrables 12 livrables ID.CTC.3.27 à ID.CTC.3.42 des deux tâches *Terminology acquisition* et *Named entity recognition*.

135. Rôle : Co-auteur. *Periodic Corpora and Activity Report on Task 3.1: Terminology Acquisition*, six Contractual Deliverables CD.CORP.3.7, CD.CORP.3.31, CD.CORP.3.55, CD.CORP.3.79, CD.CORP.3.103, CD.CORP.3.115.

136. Rôle : Co-auteur. Six contractual deliverables *Periodic Corpora and Activity Report on Task 3.2-1: Named Entity Recognition in Scientific and Technical Information,* 6 Contractual Deliverables CD.CORP.3.8, CD.CORP.3.32, CD.CORP.3.56, CD.CORP.3.80, CD.CORP.3.104, CD.CORP.3.116.

137. Rôle : Co-auteur. Six Contractual Deliverable, *Periodic Corpora and Activity Report on Task 3.3: Ontology Acquisition* CD.CORP.3.11, CD.CORP .3.35, CD.CORP .3.59, CD.CORP .3.83, CD.CORP .3.107, CD.CORP.3.119.

138. Rôle : Co-auteur. *Periodic Corpora and Activity Report on Task 3.4: Semantic annotation*, six Contractual Deliverable CD.CORP.3.12, CD.CORP.3.36, CD.CORP.3.60, CD.CORP.3.84, CD.CORP.3.108, CD.CORP.3.120.

139. Rôle : Co-auteur. Quaero participant to TIAE project, *activity report and description prototypes*, six contractual deliverables CD.TIAE.6.1 à 6.

140. Rôle : Co-auteur. Claire Nédellec et l'équipe Bibliome, *Evaluation of the integration of content analysis technologies*, six contractual deliverables CD.TIAE.7.1à 7.6.

141. Rôle : Co-auteur. Claire Nédellec et l'équipe Bibliome, *Demonstrators*, 4 contractual deliverables CD.TIAE.8.3 à 8.6.

142. Claire Nédellec, 5 livrables Rapport de propriété industrielle de l'Institut National de la Recherche Agronomique. Rapport contractuel du projet Oseo Quaero.

143. Claire Nédellec, *Sources for bacteria ecology description*, INRA internal report, nov. 2008.

144. Claire Nédellec, E*xtraction patterns for biotope description*, V1, INRA internal report, nov. 2008.

145. Claire Nédellec, *Ontology biotope*, V1, INRA internal report, Nov. 2008.

146. Claire Nédellec, *Inferences for biotope description*, V1, INRA internal report, nov. 2008.

147. Claire Nédellec, Report on the context of named entities of ambiguous types in the Biology corpus, rapport de recherche du projet Quaero, juillet 2009.

Principaux rapports de recherche du projet Epipagri (2)

148. Rôle : Auteur de la section WP1, *Epipagri IP information system*, Final Activity Report, juillet 2008

149. Claire Nédellec, Robert Bossy, Alain Kotoujansky, Annik Lacombe, Adrien Canteloube, *Evaluation de trois extracteurs de termes appliqués à un corpus de brevet*, internal report, Epipagri project, 3 mai 2007.

Principaux rapports de recherche du projet Alvis (10)

150. Rôle : Co-auteur. *Report on Tests*, deliverable D8.3 Projet IST-FP6 Alvis, mars 2007.

151. Rôle : Coordinateur. *Semantic class learning and syntactic resources tuning*, deliverable D6.4 Projet IST-FP6 Alvis, mars 2007.

152. Rôle : Coordinateur. Prototype and documents for learning and integration of named entities and terminology, deliverable D6.3 Projet IST-FP6 Alvis, mai 2006.

153. Rôle : Co-auteur. *Final report on NLP analysis and normalization*, deliverable D5.3 Projet IST-FP6 Alvis, mai 2006.

154. Rôle : Coordinateur. Prototype and documents for learning and integration of named entities and terminology, deliverable D6.3 Projet IST-FP6 Alvis, mai 2006.

155. Rôle : Co-auteur. *Report on theory and software of normalization options for IR*, deliverable D5.2, Projet IST-FP6 Alvis, décembre 2005.

156. Rôle : Co-auteur. *Report on test plans*, deliverable D8.2 Projet IST-FP6 Alvis, août 2005.

157. Rôle : Co-auteur. Report on method and language for the production of augmented document representations D5.1, Projet IST-FP6 Alvis, décembre 2004.

158. Rôle : Coordinateur. Requirements for integration of WP6 results into WP5 normalization and representation tasks and into WP9 query refinement task, deliverable D6.2, Projet IST-FP6 Alvis, décembre 2004.

159. Rôle : Coordinateur. Requirements and training data for learning and testing in a biological domain, deliverable D6.1, Projet IST-FP6 Alvis, juin 2004.

Principaux rapports de recherche du projet ExtraPloDocs (9)

http://www-lipn.univ-paris13.fr/RCLN/Extra/ExtraPloDocs/rapports.html

160. Rôle : Coordinateur. *Etude de l'existant et des besoins*, rapport D2.1 et D2.2 du projet RNTL ExtraPloDocs**,** février 2003.

161. Rôle : Co-auteur. *Architecture et spécifications fonctionnelles*, rapport D2.3 du projet RNTL ExtraPloDocs**,** juin 2003.

162. Rôle : Coordinateur. *Acquisition d'une ontologie*, rapport D3.2 du projet RNTL ExtraPloDocs**,** janvier 2005.

163. Rôle : Coordinateur. *Choix, évaluation et adaptation d'un analyseur syntaxique*, rapport D3.3a du projet RNTL ExtraPloDocs**,** décembre 2003.

164. Rôle : Coordinateur. *Constitution des données d'apprentissage et de test pour le filtrage de document* Rapport D5.1 du projet RNTL ExtraPloDocs**,** septembre 2004.

165. Rôle : Coordinateur. *Rapport d'évaluation des méthodes d'apprentissage pour la sélection de fragments pertinents*, Rapport D5.2 du projet RNTL ExtraPloDocs**,** septembre 2004.

166. Rôle : Coordinateur. *Constitution et annotation du corpus de test pour l'apprentissage de règle d'extraction de connaissances*, Rapport D6.1 du projet RNTL ExtraPloDocs**,** septembre 2004.

167. Rôle : Coordinateur. *Méthodes pour l'acquisition automatique de règles d'extraction*, Rapport D6.2 du projet RNTL ExtraPloDocs**,** décembre 2005.

168. Rôle : Coordinateur. *Evaluation des règles d'extraction de connaissance*, Rapport D6.3 du projet RNTL ExtraPloDocs**,** décembre 2005.

Projet bilateral Incremental Revision of Relational Models in Robotics and Vision (1)

169. LRI / Bar Ilan University, *Activity report*, septembre 1993. *(Rôle : co-rédacteur)*.

Projet ILP 1 et 2 (3)

170. ILP Declarative Bias Workpackage *Declarative Bias*, Deliverable DB 3. Projet ILP, août

1995. *(Rôle : coordination et synthèse)*

171. LRI, Deliverable LRI2. *Revision with small example sets*, Projet ILP 2, Janvier 1998. *(Rôle : rédacteur)*

172. LRI, Deliverable LRI4. Acquisition of semantic knowledge from specification texts by ML methods, the ASIUM system, Projet ILP 2, janvier 1998. (Rôle : co-rédacteur)

Principaux rapports de recherche du projet MLT (5)

173. Coimbra, ISoft et LRI, *Treatment of Hypertension*, Tech. Report, Projet Esprit MLT, mai 1991. *(Rôle : co-rédacteur)*

174. Coimbra, ISoft et LRI, *Case Study on Loan Analysis*, Tech. Report, Projet Esprit MLT, septembre 1992. *(Rôle : co-rédacteur)*

175. ISoft, LRI et Coimbra, *User's Guide for APT*, Deliverable 4.2, Projet Esprit MLT, mars 1993. *(Rôle : rédacteur)*

176. ISoft, LRI et Coimbra, *Final Discussion on APT and KBG,* Deliverable 4.4, Projet Esprit MLT, Mars 1993. *(Rôle : co-rédacteur)*

177. WP4 MLT partners, "The Learning Tools" in Final Report on Esprit Project Machine Learning Tollbox n° 2154, juin 1993. (Rôle : coordination et synthèse)

I.2 - Synthèses scientifiques

I.2.1. Dans périodique à comité de lecture. (article invité). (1)

178. Nédellec C., "Bibliographical Information Extraction in Genomics", article invité dans *IEEE Intelligent Systems: Trends & Controversies - Mining Information for Functional Genomics*, N. Shadbolt (Ed.), 17 (3), p. 76-80, mai-juin, 2002. Impact factor 2009 : 3.144.

I.2.3. Chapitre d'ouvrage. (5)

Chapitre d'ouvrage international (4)

179. Nédellec C., Nazarenko A. et Bossy R., "Information Extraction", *Ontology Handbook.*, S. Staab, R. Studer (eds.), Springer Verlag, Berlin (DEU) : Springer Science - Business Media Deutschland GmbH (International Handbooks on Information Systems), 2nde édition révisée, pages 663-686, 2009.

180. Nédellec C. et Nazarenko A., " Ontology and Information Extraction: A Necessary Symbiosis", *Ontology Learning from Text: Methods, Evaluation and Applications.* p. 155-170, Volume 123 Frontiers in Artificial Intelligence and Application, P. Buitelaar, P. Cimiano, B. Magnini (eds.), IOS Press, 2005.

181. Nédellec C., Machine Learning for Information Extraction in Genomics - State of the Art and Perspectives", *Text Mining and its Applications: Results of the NEMIS Launch Conference Series: Studies in Fuzziness and Soft Computing,* p. 99-118, S. Sirmakessis (Ed.), Springer Verlag, 2004.

182. Kayser D., Vosniadou S., Nédellec C., Tiberghien A. et Zucker J. D., Introduction. *Knowledge Representation Changes in Humans and Machines*, Kayser D. et Vosniadou S. (Eds.), Elsevier Science, 2000.

Chapitre d'ouvrage national (1)

183. Nédellec C., "APT, un Système d'Apprentissage Coopératif", in *Acquisition et Ingénierie des Connaissances, tendances actuelles*, p. 307-327, Reynaud C., Laublet P. et Aussenac N. (Eds.), Cépaduès, 1995.

I.2.7. Autres supports.

I.2.7.a Communication invitée (7)

Communications invitées à des séminaires internationaux (3)

184. "Application of Machine Learning to Knowledge Acquisition from Texts", *Colloquium serie,*

Université de Karlsruhe, novembre 2000.

185. "Apprentissage automatique de connaissance à partir de textes pour l'extraction d'information", séminaire de l'EPFL (*Ecole Polytechnique Fédérale de Lausanne*), mars 2001.

186. « Apprentissage de connaissances pour l'extraction d'information textuelle en génomique" (Machine Learning for Information Extraction in Genomics), CUI, Université de Genève, décembre 2002.

Communications invitées à des séminaires académiques nationaux (2)

187. "Apprentissage automatique de connaissances à partir de documents textuels", *Journée du LRI*, Villemartin, juin 2000.

188. "Apprentissage de connaissances et extraction d'information à partir de textes", *Séminaire du GREYC*, Université de Caen, décembre 2000.

Séminaire invité à un groupe de recherche national (2)

189. "Classification conceptuelle pour la formation de classes sémantiques", *A3CTE*, Pitié-Salpétrière, novembre 1999.

190. "Extraction d'information appliquée au domaine biomédical – apprentissage et traitement automatique de la langue ». Groupe *Statistiques des Séquences Biologiques* (SSB), janvier 2004

II - DOCUMENTS À VOCATION DE TRANSFERT ou relatifs à l'animation de la recherche.

II.1 - Travaux personnels *(58)*

II.1.2. Dans périodique sans comité de lecture. (5)

Dans périodique sans comité de lecture international (1)

191. Article : Nédellec, C., " Automatic metadata production ". *Content Mining Service, Newsletter technic*. 2.0 2006.

Dans périodique sans comité de lecture national (4)

192. Modéliser la fouille de données, dossier *Mécanismes du vivant: La modélisation omniprésente*, INRA, http://www.inra.fr/Chercheurs-etudiants/Mecanismes-du-vivant/Tous-les-dossiers/La-modelisation-omnipresente/Modeliser-la-fouille-de-donnees, 2013.

193. Article : Nédellec, C., " Activité de MIG en IA et bioinformatique", *Bulletin AFIA* (*Association Française pour l'Intelligence Artificielle*) n° 63, pages 25-26, 2007.

194. Article : Nédellec, C., " MIG-INRA, Activité en IA - le projet Caderige ", *Bulletin AFIA*, 2003.

195. Article : Nédellec, C., " Le projet ExtraPloDocs, EXTRAction de connaissance pour l'EXPLOitation de la DOCumentation Scientifique ", *Bulletin AFIA*, 2003.

196. Article : Nédellec, C., " Activité du LRI en Acquisition des Connaissances", *Bulletin AFIA*, 1992.

II.1.3. Chapitre d'ouvrage. (4)

Encyclopédie (2)

197. Coste F., Nédellec C ., Schiex T. Vert J.-P., "IA et Bioinformatique", *Encyclopédie de l'intelligence artificielle.*, Marquis P. et al. (eds.), Cépaduès, à paraître en 2013.

198. Nédellec C. " Apprentissage automatique", *Encyclopédie de l'Informatique et des systèmes d'information*, J. Akoka et I. Wattiau (eds), p. 1243-1253, Vuibert, 2006.

Etat de l'art (2)

199. Robardet C., Nédellec C., Boulicaut J.-F. Fouille de données et analyse du transcriptome. *Chapitre 4 de Informatique pour l'analyse du transcriptome*, pages 101-141, Hermès Lavoisier, traité IC2, 2004.

200. Traduction du chapitre "LEAP, a Learning Apprentice for the VLSI Design" de Mitchell T.M., Mahadevan S. et Steinberg L. I. paru dans *Machine Learning III, An Artificial Intelligence Approach*, Kodratoff Y. et Michalski R. S. (Eds.), Morgan Kaufman, 1990, dans *L'apprentissage Automatique : Une Approche Intelligence Artificielle*, p. 239-269, Kodratoff Y., Michalski R. S., Carbonell J. G., et Mitchell T. M. (Eds.), Cépaduès Edition, 1993.

II.1.6. Communication invitée et tutoriels (19)

Séminaire invité à un groupe de recherche national (1)

201. « Le projet Quaero », réunion du réseau *IST* (Informatique scientifique et technique) des organismes de recherche nationaux, 27 avril 2006.

Séminaires INRA (3)

202. "Annotation sémantique de documents scientifique" séminaire *Texte et Connaissance, méthodes, besoin, perspectives* Jouy-en-Josas, 3-4 juillet 2008. http://genome.jouy.inra.fr/bibliome/TC/).

203. "Modélisation de connaissance termino-ontologique application à la recherche documentaire", *séminaire modélisation du département Phase*, Physiologie Animale, Ecully, 9 mars 2010.

204. “Gestion des connaissances - Ontologies", *Séminaire stratégique CEPIA (Département Caractérisation et Elaboration des Produits issus de l'agriculture*), Paris, 17 mars 2010. https://intranet.inra.fr/cepia/vie_collective/seminaires_du_departement/seminaire_cepia_2010

Séminaires public de projet transversal (3)

205. Présentation publique : "Demonstration of the EPIPAGRI database and website", *Epipagri Network introduction meeting*, 22 avril 2008.

206. Claire Nédellec, Robert Bossy, Frédéric Papazian, Sarhane Safraoui, Sophie Aubin, Sandra Makuntima "Epipagri.org, Presentation of the final version of the Web site", Epipagri Final Review, Barcelone – septembre 2008.

207. Présentation publique : Claire Nédellec, R. Bossy, B. Teyssendier, W. Golik, F. Papazian, P. Veber, A. Kammoun « Recherche d’information dans les brevets du domaine de *l’ARP VegA* », *Assemblée Générale de l’ARP (Atelier de Recherche Prospective) VegA*, 12 janvier 2010. http://www.inra.fr/arpvega/travaux_de_l_arp_vega/assemblees_generales/3eme_assemblee_generale.

Tutoriels destinés à des utilisateurs de la recherche (professionnels, partenaires institutionnels...) (13)

Tutoriel international (2)

208. Organisation de la *training session text-mining* (1 semaine juin 2009) de l’action Marie Curie Transys. http://genome.jouy.inra.fr/bibliome/Transys/ 3 cours.

209. Tutoriel : "Machine Learning and Natural Language Processing for Information Extraction in Genomics", workshop *text and literature mining* of the *Industry Programme of the European Bioinformatics Institute* (EBI), 19-23 janvier 2003.

Tutoriel national (5)

210. C. Nédellec et A. Girard. Conférence invitée "Traitement automatique du langage dans la gestion des données", Journées Formation Frédocs de Renatis, (réseau national des professionnels de l’information scientifique du CNRS), Les données de la recherche : gestion et valorisation, Aussois, 7-10 octobre 2013.

211. C. Nédellec et E. Benoit, Conférence invitée "Recherche d’information spécialisées, l’enjeu de l’automatisation", *Rencontres des professionnels de l'IST,* Nancy, le 20 juin 2007. http://avoir.inist.fr/article.php3?id_article=285&var_recherche=NEDELLEC

212. "Text Mining : Extraction de connaissances à partir de textes ”, aux *journées francophones d'extraction et de gestion des connaissances*, (EGC), Montpellier, janvier 2002.

213. "Apprentissage de connaissances sémantiques spécialisées" *TCAN-CNRS tutorial serie, "Langue, connaissances, information"*, 22 au 27 mai 2005.

http://www.dr4.cnrs.fr/tcan/tcan/activites/2005/ecole/TCAN_descriptionduprojet.html

Tutoriel INRA (7)

214. "Thesaurus et recherche sémantique d’information", 7 documentalistes du projet TriPhase (dépt Phase, INRA), INRA Rennes, le 4 avril 2013.

215. "Annotation sémantique avec AlvisAE pour la sélection du blé par marqueur", 13 sélectionneurs du projet FSOV Sam Blé, cours et TP sur PC, INRA Jouy-en-Josas, le 16 avril 2013.

216. "TyDI, acquisition de terminologie à partir de corpus", cours et TP sur PC, 7 documentaliste du projet TriPhase, (dépt Phase, INRA), INRA Jouy-en-Josas, le 27 mai 2013. Durée 5 h.

217. Tutoriel : "TyDI, acquisition de terminologie à partir de corpus", *ATOL*, (dépt Phase, INRA), Rennes, le 12 février 2012.

218. Tutoriel : "Fouille de texte", *PEPI INRA Gestion de Données* Les-Loges-en-Josas, 6) mai 2011.

219. Tutoriel : "Texte et Connaissance - Quelques éléments méthodologiques ", séminaire *Texte et Connaissance, méthodes, besoin, perspectives* . Public : IST, chercheur et direction scientifique INRA. Jouy-en-Josas, 3-4 juillet 2008. http://genome.jouy.inra.fr/bibliome/TC/).

220. Tutoriel : Nédellec, C., "Des outils d'analyse sémantique pour la recherche d'information", *formation INRA aux techniques d'analyse de l'information textuelle*. Public : documentalistes de l'INRA. Paris, le 20 mars 2007.

II.1.7. Rapports écrits (expertises). (6)

221. Rôle : co-auteur. *Synthèse des contributions des sites de la région I'île de France Sud-Ouest - Science et Société* par le Comité Local d'Organisation des Etats Généraux de la Région l'île de France–Sud-Ouest. (14 pages).

222. Rôle : co-auteur. *Synthèse des contributions des sites de la région I'île de France Sud-Ouest - Evaluation* par le Comité Local d'Organisation des Etats Généraux de la Région l'île de France–Sud-Ouest. (11 pages).

223. Rôle : co-auteur. *Synthèse des contributions des sites de la région I'île de France Sud-Ouest – Organisation et financement de la recherche* par le Comité Local d'Organisation des Etats Généraux de la Région l'île de France–Sud-Ouest. (13 pages).

224. Rôle : co-auteur. *Synthèse des contributions des sites de la région I'île de France Sud-Ouest – Statut des personnels* par le Comité Local d'Organisation des Etats Généraux de la Région l'île de France–Sud-Ouest. (19 pages).

225. Nédellec, C., Contribution au rapport *d'autoévaluation de l'INRA pour l'AERES*, " Recherche en Sciences et Technologies de l'Information Scientifique et Technique et de la Communication ". 7 pages. 2008. Voir p. 43 du document final : *Rapport d'auto-évaluation 2009*.

226. Bernard Teyssendier de la Serve, Claire Nédellec, Philippe Lenée, Franck Leguerhier, Pour une initiative française en vue du développement d'un système d'information sur les brevets dans les agrobiotechnologies (SIB AgroBiotech), 17 pages, 18 septembre 2008.

II.1.8. Créations informatiques ou audiovisuelles. (24)

II.1.8.a Services public en ligne (6)

Coordination de la recherche, du développement et de la maintenance des services en ligne

227. A. Kotoujansky, R. Bossy, C. Nédellec. **BioAlvis**. Recherche sémantique documentaire en microbiologie (400 000 références PubMed), depuis 2007. http://bibliome.jouy.inra.fr/webalvis

228. F. Papazian, S. Safraoui, R. Bossy, C. Nédellec. **Epipagri**. Recherche sémantique de brevet en agrobiotechnologie, en 2008. https://www.epipagri.org/

229. R. Bossy et C. Nédellec. **VegA'lvis**. Recherche sémantique de brevet dans le domaine de l'exploitation de la biomasse, en 2010. http://bibliome.jouy.inra.fr/A01H

230. P. Bui-Quang, C. Nédellec. **GIS HP2E**. Systèmes de production de grande culture à hautes performances économiques et environnementales (GC-HP2E), en 2010. http://bibliome.jouy.inra.fr/gisdemo

231. R. Bossy et C. Nédellec. **AnimalIR**, moteur de recherche bibliographique de la revue Animal basé sur l'ontologie ATOL et sa lexicalisation. http://bibliome.jouy.inra.fr/test/alvisir/Animal/

232. R. Bossy et C. Nédellec. **SamBlé.** Sélection du blé assistée par marqueur. http://bibliome.jouy.inra.fr/test/alvisir/FSOV.

233. R. Bossy et C. Nédellec. **Biotope**. Moteur de recherche bibliographique sur les biotopes bactériens. http://bibliome.jouy.inra.fr/test/alvisir/webalvis/biotopes-demo/

II.1.8.b Corpus, lexiques, dictionnaires et thesaurus (9)

Coordination et contribution à la conception de,

Implantation des cultures

234. P. Bui Quang, W. Golik et C Nédellec. ***Terminologie et l'ontologie GIS HP2E*** sur l'implantation des cultures, pour l'indexation de documents (1 959 termes et 408 concepts) (service en ligne). (2010-2012). Diffusion à des fins de R & D.

Biotopes bactériens

235. C. Nédellec, R. Bossy, V. Loux. ***Corpus Biotope*** pour l'extraction d'information de biotope microbien. Un corpus de milliers de pages web de vulgarisation (titre origine et noms des espèces) à partir des sites du NCBI, GOLD, JGI, EBI and Genoscope. (2009)

236. Equipe Bibliome. **Corpus Biotope de** 200 documents annotés manuellement avec des entités et relations par l'équipe Bibliome. Les entités sont référencés dans les catégories de l'ontologie OntoBiotope. Il est distribué dans BioNLP'11 et 13 Shared Tasks.

C. Nédellec, Sources for bacteria ecology description, INRA internal report, Nov. 2008.

C. Nédellec, Extraction patterns for biotope description, V1, INRA internal report, Nov. 2008.

C. Nédellec, Inferences for biotope description, V1, INRA internal report, Nov. 2008.

237. C. Nédellec. **OntoBiotope** : une ontologie des bactéries, biotopes et phénotypes. La partie habitat contient 2 120 concepts et synonymes. Elle est publiquement distribuée au format OBO dans le cadre du challenge BioNLP-ST'13 pour la tâche Bacteria Biotope à http://bibliome.jouy.inra.fr/MEM-OntoBiotope/OntoBiotope_BioNLP-ST13.obo

Marqueurs génétiques du blé

238. C. Nédellec, T. Marcel. ***Terminologie et ontologie sur les phénotypes et facteurs environnementaux du blé*** pour l'extraction d'information à partir d'articles (388 concepts structurés).

239. D. Valsamou, C. Nédellec. **Wheat marker corpus** Corpus de 5 000 articles complets sur les marqueurs génétiques du blé (service SamBlé).

240. P. Sourdille, M. Ranoux, D. Valsamou, C. Nédellec, et les partenaires du projet FSOV Sam Blé. **Wheat marker corpus annoté** : 120 articles annotés avec les entités et relations n-aires décrivant les marqueurs génétiques, leur type, leurs relations avec les gènes, phénotypes et variétés. (2011).
Base de donnée de marqueurs génétiques du blé avec interface d'interrogation. (2013).

Réseaux de régulation des bactéries

241. C. Nédellec. ***Quaero_t3.2_gene*** (juin 2009) : un ensemble de 422 références bibliographiques de la base MedLine centrés sur la transcription chez *Bacillus subtilis* et dont les noms de gène et de protéines sont annotés manuellement au format XML. Diffusion à des fins de R & D.

241. J. Jourde, K. Fort, P. Bessières, C. Nédellec. ***Renaming corpus*** (2010) : un ensemble de 1 644 références PubMed contenant 450 renommages de gènes annotés manuellement. Corpus de la tâche Renaming de BioNLP-ST'11.

242. C. Nédellec et P. Bessières. T***erminologie et Ontologie*** (1 500 concepts) en ***microbiologie*** pour l'indexation de références PubMed (service BioAlvis). (2007-2008). Diffusion à des fins de R & D.

242. P. Bessières, E. Alphonse, A.-P. Manine, S. Aubin, C. Nédellec. **Corpus de la compétition LLL** (*Learning in Logic*). 271 exemples d'interactions géniques dans 80 phrases avec les analyses syntaxiques curées manuellement. 2005.
Complété en 2011 : corpus de la compétition BI (*Bacteria Interaction*) de BioNLP-ST'11 puis GRN (*Genic Regulation Network*) de BioNLP-ST'13.

243. A. Lacombe, F. Le Guerhier et C. Nédellec. T***erminologie agrobiotech*** pour l'indexation de brevets (10 000 termes environ) (service Epipagri). (2007). Diffusion à des fins de R & D.

Propriété intellectuelle et biomasse

244. B. Teyssendier, W. Golik et C. Nédellec. T***erminologie et ontologie A01H*** pour l'indexation de brevets (30 296 termes dans 24 906 ensembles de synonymes et 2 802 concepts structurés en 11

niveaux, plus 22 104 concepts isolés). (Service Veg'Alvis). (2010). Diffusion à des fins de R & D.

II.1.8.c Logiciels (9)

Coordination et contribution à la conception de,

245. F. Papazian, R. Bossy, C. Nédellec. **AlvisNER*ML*** (2008-) est un composant logiciel qui apprend des règles de reconnaissance d'entités nommées et les applique. Le composant d'apprentissage automatique de règle de reconnaissance d'entité nommée dans des documents en langage naturel inclut une fonction de désambiguïsation endogène et une fonction de reconnaissance de nouveaux noms basée sur le contexte et des critères typographiques. Distribution Quaero.

246. S. Aubin, F. Papazian, W. Golik, R. Bossy, P. Veber, C. Nédellec. Le logiciel ***TyDI*** (2008-) est une application web documentée dédiée à la construction collaborative de terminologie et d'ontologie qui permet la gestion, l'affichage et la qualification de termes et de relations de synonymie et d'hyperonymie entre termes. Il assiste la conception d'ontologie hiérarchique à partir de la terminologie. Il est écrit en Java et basée sur la plate-forme logicielle *NetBeans Platform* de Sun Microsystems et le système de gestion de bases de données relationnelles PostgreSQL. Il est utilisé par une vingtaine de personnes INRA et non INRA. Distribution Quaero.

247. R. Bossy, C. Nédellec. ***Alvis NLP/ML*** (2008-) est une chaîne logicielle documentée, implémentée en Java, de traitement pour l'annotation de documents textuels, intégrant des outils de traitement automatique des langues naturelles pour la segmentation en mots/phrases, la reconnaissance d'entités nommées, l'analyse de termes, l'analyse syntaxique et le typage sémantique. Il inclut plusieurs composants pour l'acquisition (semi)-automatique de ces ressources, fondées sur des techniques d'apprentissage automatique : ***AlvisNER, BioYateA, Asium***. La chaîne est facilement extensible par ajout de nouveaux composants. Il est utilisé en interne, en cours de transfert interne à l'INRA.

248. R. Bossy, L. Patey, F. Papazian, P. Veber, M. Taylor, W. Buntine, C. Nédellec. ***AlvisIR*** (2007-). Logiciel documenté pour l'indexation de documents et la recherche sémantique en ligne de document. Il inclut un composant d'indexation appelé Zebra et une interface Web en PHP 5. Il permet la recherche de document à partir de requête interprétées en CQL en incluant l'interprétation de concepts et la synonymie de termes, ou à travers la navigation dans une ontologie. Sa conception générique permet d'instancier une nouvelle version dédiée en quelques heures. Cinq moteurs sont utilisés en 2013.

249. F. Papazian, R. Bossy, P. Veber, G. Bisson, C. Nédellec. **AlvisAE** (2010-). Visualisation et édition d'annotations sémantiques collectives de documents manuelles graphiques, d'entités (surlignage) et de relations (lignes) textuelles basé sur la technologie GWT utilisable avec un navigateur Web. Les annotations sont visualisables dans un format tabulé. Les formats d'entrée et de sortie sont les formats standards du web sémantique. Par rapport la version v0.5 inclut des fonctionnalités puissantes de révision d'ontologie en temps réel et collective, gérée par TyDI. Plus de 20 utilisateurs en cours.

250. W. Golik, S. Aubin, C. Nédellec. **BioYateA** (2010-) est une version enrichie du logiciel YaTeA d'extraction terminologique. Il inclut des fichiers de filtrage (processing et post-processing) de termes spécifiques à la biologie. Il inclut également le traitement des adjectifs verbaux (ex. *NN (NNS) of X-ing*) et des groupes prépositionnels en *at* (ex. *stability of the microbe at high temperature*) et *to* (ex. *resistance of children to bacteria)*. Il est publiquement distribué sous forme de module CPAN.

251. D. Valsamou, C. Nédellec. **AlvisCrawler** (2011-). Aspiration d'articles scientifiques sur le Web et transformation de format de documents. En Perl. Données en mémoire. Documents et métadonnées sous forme de fichiers. Autonome ou composant d'AlvisNLP.

252. J. Jourde, S. Aubin, A. Lemaçon, T. Lacroix, C. Nédellec, P. Bessières. **CoCitation** (2007-). Visualisation d'annotations sémantiques dédiées à la biologie. En Java. Client Web (GWT) –

serveur. Données en PostgreSQL. WebService pour AlvisNLP. Intégré à la plateforme de l'unité MIG, IGO.

II.2. Travaux encadrés ou coordonnés par l'auteur (33)

II.2.1. Edition d'un ouvrage collectif. (11)

Acte de conférence (1)

253. Nédellec C., Rouveirol C., éditrices des *Proceedings of the Tenth European Conference on Machine Learning*, (ECML'98), Springer Verlag, avril 1998.

Acte de workshops (7)

254. Hilario M. et Nédellec C. éditrices des Proceedings of the "*Data and text mining for integrative biology*", associé à ECML/PKDD, European Conference on Machine Learning, Berlin, Allemagne, 18 septembre 2006. http://cui.unige.ch/~hilario/ecml-pkdd06-biows/

255. Cussens J. et Nédellec C., Proceedings of the *Learning Language in Logic* ICML joint workshop and the *Challenge task, Extracting Relations from Biomedical Texts*, Bonn, Allemagne, août 2005. http://www.cs.york.ac.uk/aig/lll/lll05/

256. Staab S., Mädche A., Nédellec C. et Hovy E., IJCAI Workshop Notes of the *Ontology Learning*, workshop of the 17th International Joint Conference on Artificial Intelligence (IJCAI), Vancouver, août 2001. http://ol2001.aifb.uni-karlsruhe.de/

257. Cardie C., Daelemans W., Nédellec C. et Tjong Kim Sang E., Proceedings of the *Fourth Conference on Computational Natural Language Learning and of the Second Learning Language in Logic Workshop*, Omni Press (Pub.), Lisbonne, septembre 2000. http://www-ai.ijs.si/~ilpnet2/events/lll-00-report.html

258. Staab S., Mädche A., Nédellec C. et Wiemer-Hastings P., ECAI Workshop Notes of the *Ontology Learning*, workshop of the 14th European Conference on Artificial Intelligence (ECAI), Berlin, août 2000. http://ol2000.aifb.uni-karlsruhe.de/

259. Kodratoff Y. et Nédellec C., *Machine Learning and Comprehensibility*, actes du workshop de 14th International Joint Conference on Artificial Intelligence, (IJCAI-95), Montréal, août 1995.

260. Kodratoff Y. et Nédellec C., *Integration of Machine Learning and Knowledge Acquisition*, Actes du workshop of the 11th European Conference on Artificial Intelligence (ECAI-94), août 1994.

Acte d'ateliers nationaux (4)

261. Buche P., Dibie-Barthélemy J., Nédellec C., Neveu P. et Soler S. Actes de l' Atelier INtégration de sources/masses de données hétérogènes et Ontologies, dans le domaine des sciences du VIVant et de l'Environnement. IN-OVIVE IC 2013, plate-forme IA 2 juillet 2013, Lille, France.

261. Nédellec C. et Zweigenbaum P., *Actes de la journée Apprentissage, connaissances sémantiques et textes*, atelier associé à la plate-forme AFIA'2003, Laval, juin 2003.

262. Gallinari P., Jacquemin C. et Nédellec C., Actes de la journée, *Apprentissage Statistique et Symbolique pour la Recherche d'Information et la Fouille de Textes*, plate-forme AFIA'99 Ecole Polytechnique, Palaiseau, juin 1999.

263. Nazarenko A. et Nédellec C., Actes de la journée *A3CTE, Applications, Apprentissage et Acquisition de Connaissances à partir de Textes Electroniques*, plate-forme AFIA'2001, Grenoble, juin 2001. (Voir http://www-leibniz.imag.fr/AFIA).

II.2.2. Mémoires de stages. (3)

Stages de Master et DEA depuis 2001. (3)

264. Sophie Aubin**,** Choix et l'intégration d'un analyseur syntaxique pour l'apprentissage de connaissances sémantiques**,** Master en ingénierie multilingue de l'INALCO, 2002.

265. Zihyu Qian**,** *Terminologie et filtrage de document*, DEA d'informatique I3 (Information, Interaction, Intelligence), de l'Université Paris 11 à Orsay, 2002.

266. Mathieu Roche *Apprentissage d'ontologie en support technique logiciel*, DEA d'informatique I3 (Information, Interaction, Intelligence), de l'Université Paris 11 à Orsay, 2001.

II.2.3. Autres. (17)

II.2.3.a Thèses soutenues (4)

267. Poittevin Luc, *Un outil générique de conception et de révision coopérative de Bases de Connaissances s'appuyant sur la notion de situation*, Thèse en informatique de l'Université Paris-Sud, soutenue le 11 septembre 1998. Rapporteurs, M. C. Haton (LORIA) et J. Breuker (Univ. Amsterdam).

268. Faure David, Conception de méthodes d'apprentissage symbolique et automatique pour l'acquisition de cadres de sous-catégorisation de verbes et de connaissances sémantiques à partir de textes : le système Asium., Thèse en informatique de l'Université Paris-Sud, soutenue le 20 décembre 2000. Rapporteurs, A. Nazarenko (LIPN) et A. Napoli (LORIA).

269. Abdel Vetah, Mohammed Ould (2005). *Apprentissage automatique appliqué à l'extraction d'information à partir de textes biologiques.* Thèse d'informatique de l'Université d'Orsay, France. Soutenue le 21 septembre 2005, France. Rapporteurs, Pascale Sébillot (INRIA) et Stan Matwin (Univ. Ottawa).

270. Manine Alain-Pierre (2006). *Application de l'apprentissage à l'extraction de connaissances à partir de notices bibliographiques en génomique.* Thèse d'informatique de l'Université d'Orsay, France. Soutenue le 25 juillet 2006. Rapporteurs, James Cussens (Univ York), Patrice Enjalbert (Univ. Caen).

II.2.3.b Mémoire d'apprentissage de Master (1)

271. Audrey Lemaçon, *Développement d'un outil d'aide à l'exploration bibliographique,* Mémoire du Master Sciences Technologies Santé - Mention Biologie Santé Bioinformatique. UFR des Sciences et des Techniques. Centre de Formation par Apprentissage. Rapport de mission professionnelle. 21 juin 2010.

II.2.3.c Articles (12)

Thomas Jamet

272. Ferreira J.L., Correia J., Jamet T., Costa E., "An Application of Machine Learning in the Domain of Loan Analysis", in Proceedings of the *European Conference on Machine Learning (ECML-93)*, p. 414-419, Brazdil P. (Ed.), Springer-Verlag, avril 1993.

Luc Poittevin

273. Poittevin L., "Représentation des Connaissances par Nodules de Situation", in Actes du *2ème Colloque des Jeunes Chercheurs en Science Cognitive*, p. 354-357, Cognito Ergo Sum, In Cognito & ARC, Giens, juin 1996.

274. Poittevin L., "Comprendre pour réviser : comment la situation peut-elle s'expliquer ?", in Actes des *3èmes Journées Explication*, p. 31-243, Joab M. et Giboin A. (Eds.), Juan Les Pins, 1996.

275. Poittevin L., "REVINOS : Un outil de révision interactive s'appuyant sur la notion de situation", in Actes des *Journées Ingénierie des Connaissances (IC)*, Zacklad M. (Ed.), 1997.

276. Poittevin L., "REVINOS: An Interactive Revision Tool Based on the Concept of Situation", in Proceedings of the *10th European Knowledge Acquisition Modeling and Management Workshop (EKAW'97)*, Plaza E. et Benjamins R. (Ed.), p. 365-370, Springer Verlag, Espagne, octobre 1997.

277. Poittevin L., "Representing the situations to help the cooperative revision", in Proceedings of the *11th Knowledge Acquisition Workshop* (*KAW'98*), Gaines B. (Ed.), Banff, Canada, avril 1998.

David Faure

278. Faure D., "Acquisition de cadres de sous-catégorisation de verbes et d'ontologies par une méthode de classification", in Actes des *4èmes Rencontres Nationales des Jeunes Chercheurs en Intelligence Artificielle (RJCIA)*, Marquis P. (Ed.), Toulouse, septembre 1998.

279. Faure D., "ASIUM: quelques expérimentations préliminaires" in Actes de la *Journée Apprentissage Statistique et Symbolique pour la Recherche d'Information et la Fouille de Textes*, Gallinari P., Jacquemin C. et Nédellec C. (Eds), juin 1999.

Sophie Aubin

280. Sophie Aubin. "Evaluation comparative de deux analyseurs produisant des relations syntaxiques", *Actes de la 10ème conférence annuelle sur le Traitement Automatique des Langues Naturelles*. Pages 67-76. Vol 2. 2003.

281. Sophie Aubin et Marc Barbier. "Extraction d'information à partir de documents contractuels", Actes de l'atelier *Acquisition, apprentissage et exploitation de connaissances sémantiques pour l'accès au contenu textuel*, plate-forme AFIA. Pages 17-28. 2003.

282. Sophie Aubin, *Challenge LLL Syntactic Analysis Guidelines*, INRA report, 21st February 2005.

Equipe Bibliome

283. Ratkovic Z., Golik W., Warnier P., "Event extraction of bacteria biotopes: a knowledge-intensive NLP-based approach". *BMC Bioinformatics* 2012, 13(Suppl 11):S8, 26 June 2012.

284. Julien Jourde, Alain-Pierre Manine, Philippe Veber, Karën Fort, Robert Bossy, Erick Alphonse, Philippe Bessières, BioNLP Shared Task 2011 – Bacteria Gene Interactions and Renaming, *proceedings of the BioNLP workshop*, Association for Computational Linguistics, Portland, 2011.