Dossier de candidature
à un poste de directeur de recherche DR 2 du CNRS
Concours 07/01

Janvier 2010

Programme de recherche

Modélisation des connaissances pour et par l'accès au contenu textuel

---

Nathalie Aussenac-Gilles
Chargée de recherche au CNRS – section 07
Institut de Recherche en Informatique de Toulouse (UMR 5505)

aussenac@irit.fr

TABLE DES MATIERES

1 RESUME

Domaine de recherche : Mes recherches se situent dans le domaine de l'ingénierie des connaissances (IC), champ de l'informatique qui s'intéresse, depuis la fin des années 80, à la mise au point de logiciels gérant ou manipulant des connaissances (associées à des savoir-faire, des pratiques ou des informations écrites ou structurées) pour assister un utilisateur dans sa tâche. L'ingénierie des connaissances se situe en amont du développement d'applications utilisant les techniques de l'intelligence artificielle et la formalisation logique, mais aussi de tout système assurant un comportement intelligent vis-à-vis de ses utilisateurs en exploitant des connaissances. Les recherches de ce domaine font l'hypothèse que l'identification, le recueil et la structuration des connaissances peuvent s'appuyer sur des modèles conceptuels avant leur formalisation ou opérationnalisation. Ces travaux définissent des techniques, des langages de représentation des connaissances et des logiciels de modélisation. Mais l'IC ne peut être réduite à sa production technique, ni sa finalité à améliorer la qualité de ses modèles ou de ses outils. Les problématiques de l'IC portent aussi sur le statut et la nature des modèles, les méthodes de leur construction puis d'utilisation, enfin leur place dans l'interaction homme-système (autant le couple cogniticien-modèle en cours de construction que le couple utilisateur-système final). Les modèles sont vus comme des traces de connaissances, comme des supports au fonctionnement de logiciels mais aussi à l'interaction entre l'homme et ces systèmes. Souvent présentés comme outils ou résultats d'une analyse, les modèles conceptuels peuvent aller jusqu'à devenir l'instrument de la co-construction de représentations et de connaissances. Leur formalisation est une étape supplémentaire qui leur confère un statut et un intérêt supplémentaire, mais dont l'IC montre qu'elle ne peut être leur seule justification. A ce titre, l'ingénierie des connaissances est fortement pluridisciplinaire, ancrée au cœur des sciences cognitives, ce qui rend parfois difficile d'en identifier les contours et les contributions.

Par ailleurs, parce que les applications, les technologies et les besoins en connaissances évoluent rapidement, le champ d'étude de l'IC se renouvelle régulièrement. Mon parcours reflète un point de vue original sur l'IC, ainsi que ma participation à ces renouvellements scientifiques.

Problématique : Depuis mon recrutement comme chargé de recherches au CNRS en 1991, j'ai développé plusieurs propositions relevant toutes de « démarches ascendantes », qui consistent à construire des modèles conceptuels à partir de traces des connaissances en usage dans activités humaines, les textes ou les discours produits. Je défends également la nécessité d'aborder de manière cohérente toutes les facettes de l'ingénierie des connaissances (méthodes, langages, outils et techniques) et en tenant compte des finalités des modèles, en s'interrogeant sur leur place et leur nature aux différentes étapes de la modélisation. Ma contribution revient de fait à définir des plates-formes associées à des méthodologies intégrant une démarche ascendante, elle privilégie la nature conceptuelle des modèles (et non leur formalisation), leur articulation avec les sources de connaissances et l'interaction dont fait l'objet leur construction. Ces deux choix correspondent à un point de vue original sur le domaine, reconnu au niveau national et international. Ma démarche fait appel à des collaborations interdisciplinaires de manière à approfondir les spécificités des vecteurs de connaissances, comme les activités humaines, les processus cognitifs qui les sous-tendent et leurs résultats, les textes ou la langue, et à des collaborations disciplinaires, de manière à associer les forces de différentes équipes ou collègues pour assurer l'outillage d'un cycle complet de modélisation.

Contributions : En vingt années de recherche, mes principales contributions regroupent donc des méthodes, logiciels et représentations, rassemblés en plates-formes, et sont définies en fonction des sources de connaissances et des modèles envisagés. Je me suis intéressée successivement aux sources de connaissances que sont l'expertise humaine puis l'activité des spécialistes, et ensuite les textes techniques, les terminologies existantes pour mieux exploiter leur complémentarité. Enfin, j'ai cherché à caractériser les modèles à construire en amont de leur opérationnalisation : leur nature conceptuelle ou formelle, leur caractère universel ou régional, leur statut, et ce, en fonction du type d'application dans lequel le modèle doit s'intégrer. J'ai pu étudier une gamme variée d'applications, des systèmes experts jusqu'à la recherche d'information en passant par les systèmes coopératifs. Il en ressort la nécessite d'envisager la modélisation comme le

lieu où doivent se gérer la diversité et la complémentarité (des sources de connaissances, des niveaux d'interprétation, d'abstraction et de formalisation), ainsi que les révisions et évolutions. L'évolution de mes travaux rend compte du glissement qui s'est opéré en 20 ans sur la notion de système à base de connaissances, la capacité de raisonner logiquement étant de moins en moins prioritaire par rapport à celle de fournir à l'utilisateur la bonne information qui va l'aider dans sa tâche ou dans sa décision. Ainsi, plus fondamentalement, ces études contribuent à mieux délimiter le domaine de l'ingénierie des connaissances, à en fonder les méthodes et résultats.

Interdisciplinarité : Ces propositions ne sont pas du seul ressort de l'ingénierie des connaissances et se fondent sur des concepts de la psychologie, de l'ergonomie et de la linguistique, mais aussi de diverses facettes de l'informatique, en particulier la représentation des connaissances et le traitement automatique des langues, pris en compte ou redéfinis dans le champ de l'ingénierie des connaissances grâce aux collaborations que j'ai entretenues avec des chercheurs de ces disciplines. Ma contribution relevant d'une ingénierie, mes résultats comprennent des réalisations comme des modèles de connaissances, des ontologies, et surtout des logiciels, et ma démarche de validation est essentiellement expérimentale. Un des points forts de mes recherches est de s'appuyer sur de nombreux projets en lien avec des entreprises, les expériences en vraie grandeur étant ici la condition indispensable à la crédibilité des résultats. Ce travail est reconnu et a donné lieu à 15 publications dans des revues, 13 chapitres de livres et plus de 50 communications à des conférences avec actes publiés, dont 25 internationales.

Programme de recherche : La dynamique de la double articulation entre textes et connaissances, les textes étant vus comme des sources de construction de modèles comme les ontologies, et en retour, les modèles comme des supports à la fouille de textes ou à la description de leur contenu, est un des sujets de recherche de l'équipe IC3 (Ingénierie des Connaissances, de la Coopération et de la Cognition) de l'IRIT dont je suis responsable depuis septembre 2007.

Mes recherches en cours intègrent les expériences passées pour en tirer des éléments méthodologiques sur les étapes et logiciels utiles à ces deux dimensions : la construction de différents types de modèles terminologiques ou ontologiques à partir de textes et d'activités humaines ; leur utilisation pour la recherche d'information dans des documents, l'analyse de contenus textuels ou l'annotation sémantique. Il s'agit là d'ailleurs de plusieurs des enjeux du web sémantique, et plus largement des difficultés à dépasser pour parvenir à une exploitation fine de l'océan d'informations numériques disponibles sur internet ou dans les entreprises.

Les défis relatifs à ces objectifs sont aujourd'hui bien identifiés et constituent un programme de recherche stimulant pour l'IC et pour l'informatique en général dans les années à venir. Un tel programme doit mobiliser des compétences en traitement automatique des langues (statistique et linguistique), en recherche d'information, en représentation des connaissances et en IA, tout en restant dans l'esprit de questionnement de l'IC, pour prendre le recul nécessaire sur la vogue actuelle des ontologies et des techniques du web sémantique. Notre équipe IC3, grâce à ses réflexions sur la place et le statut des modèles, ainsi que ses collaborations dans l'IRIT, avec CLLE-ERSS, et des équipes nationales ou internationales, se situe dans une position privilégiée pour les aborder.

Mon projet de recherche se situe dans cette lignée, insistant sur la nécessité de se focaliser sur des questions sémantiques et sur l'étude de l'articulation entre modèles et contenus documentaires avec un regard interdisciplinaire. Intitulé "Méthodes ascendantes pour l'ingénierie des connaissances : contribution à l'accès au contenu documentaire", il vise d'abord une réflexion théorique sur la sémantique en jeu dans l'analyse de textes et la modélisation des connaissances, et ses conséquences en matière de représentation conjointe de connaissances et d'un lexique. Cette réflexion est indissociable d'un deuxième objectif, visant une contribution pratique, en termes de méthode et d'outils d'analyse et de modélisation. En effet, l'IC se place toujours dans une position d'intégration de techniques, outils et modèles, fournis désormais par le TAL, l'IA et la recherche d'information, mais aussi d'innovation et de création pour parvenir à des solutions utiles, intégrées et accordant toute sa place à l'analyste qui modélise puis à l'utilisateur du système final.

Mise en œuvre du programme : Mes travaux en cours amorcent plusieurs pistes de ce projet de recherche :

- Pour définir *un cadre générique de construction de modèles termino-ontologiques* à partir de textes, la difficulté est de mettre en place une chaîne de traitement et d'analyse de textes adaptée associée à des aides à la modélisation, objectif du projet de plate-forme DAFOE4App auquel je participe.
- Parmi les connaissances recherchées dans les textes, je m'intéresse particulièrement au *repérage de relations sémantiques*. Il s'agit là d'une compétence clé dans l'équipe IC3 que nous allons développer en intégrant d'autres approches (apprentissage automatique et traitement de relations exprimées sur plusieurs phrases) à l'approche par patrons implémentée dans notre logiciel Caméléon. Nous voulons fournir un outil facile à enrichir et qui apporte plus d'aide à l'analyse linguistique puis à la construction et à l'instanciation d'ontologie à composante lexicale.
- Une autre problématique correspond à la *gestion de l'évolution des modèles ontologiques* dans leur contexte d'utilisation. Il s'agit d'assurer la cohérence entre l'ontologie, le vocabulaire et les connaissances d'un domaine, des collections de textes, ainsi que des index ou annotations associant ces ontologies et ces textes ; c'est là un des enjeux des projets Corpus Logicistes et DynamO. Le contexte d'usage des ontologies, et ceci est encore plus criant dans le cas du web, est d'évidence en évolution permanente. Or les ontologies sont souvent considérées comme des représentations stables puisque consensuelles. Nous étudions les évolutions d'ontologies en fonction des évolutions du contexte de leur utilisation. Une de nos études sur ce thème (menée dans le projet Dynamo) s'appuie sur l'interprétation de résultats de fouille de textes à l'aide d'agents adaptatifs.
- Enfin, un dernier axe de mon programme de recherche sera de *diversifier les expériences d'utilisation d'ontologies* à la prise de décision (type d'application coopérative développée dans l'équipe IC3), ainsi d'autres types de recherche d'information que la recherche par requête, comme les analyses d'opinions (Asher *et al.*, 2009) ou les systèmes question-réponse (collaboration avec l'équipe Lilac), qui s'appuient de plus en plus sur ce type de modèle. Dans ces contextes, un des besoins est de rendre compte de points de vue personnels sur un domaine. Nous étudierons l'intérêt de définir des ontologies « individuelles » (Condamines, 2009) (Kim *et al.*, 2001) (Jung, 208).

Ce programme de recherche est étroitement lié à **mon activité d'animation scientifique** et à un travail qui s'appuie sur **un réseau de collaborations**, nationales et de plus en plus *internationales*. Ainsi, en fin de thèse (1989), les équipes françaises de l'IC, domaine encore jeune, étaient dispersées et de faible effectif. Avec des collègues, nous avons eu le souci d'animer la communauté de recherche nationale en fondant le GRACQ, un groupe de travail très actif jusqu'en 1995, rattaché à l'AFIA et au GDR-I3. Le bureau de ce groupe continue de piloter des journées scientifiques, la liste de diffusion info-ic, la conférence IC et gère un site web. A partir de 1998, mes travaux sur la modélisation de connaissances et de terminologies à partir de textes ont été réfléchis au sein du groupe TIA, puis d'un groupe de travail, ASSTICCOT, missionné par le RTP-DOC, deux groupes que j'ai co-animés avec ma collègue linguiste A. Condamines. TIA pilote également la conférence « Terminologie et IA ». Pour compléter les travaux effectués par notre équipe CSC, devenue IC3, à l'IRIT, j'ai également multiplié les collaborations, tout d'abord avec des chercheurs des laboratoires LRI, LIPN, projets Acacia puis Edelweiss de l'INRIA Sophia-Antipolis et IRIN, et depuis plus de 5 ans avec d'autres équipes de l'IRIT (SIG - recherche d'information-, SMAC - multi-agents adaptatifs et LiLAC – sémantique formelle et analyse du discours-). La complémentarité de nos travaux a permis de les fédérer, autour des méthodes MACAO puis TERMINAE, et, depuis 5 ans, via des projets nationaux financés par l'ANR (DAFOE4App, Corpus Logicistes, DynamO et GEONTO). Mon projet contribuera à renforcer la cohérence thématique au sein de l'IRIT, dans l'axe prioritaire « masse de données et calcul », et à renouveler les projets de collaboration au niveau national. En effet, la visibilité des équipes nationales passe par la pérennisation et la diffusion de résultats jusqu'ici ponctuels et localisés aux équipes les produisant. La plate-forme DAFOE peut servir de socle à une telle mise en commun de résultats. Un nouveau projet est à l'étude en ce sens.

L'ensemble a fourni à mes travaux une meilleure visibilité nationale au sein de l'intelligence artificielle, et une reconnaissance internationale dans le domaine. Dès la préparation de ma thèse,

j'ai pris des contacts avec la communauté scientifique internationale à travers des publications et l'organisation de conférences et workshops, privilégiant les congrès spécialisés aux grandes conférences d'IA. Je fais partie du comité éditorial de 2 revues internationales (IJHCS et Applied ontology) et d'une revue nationale (Revue I3), et du comité de pilotage de la conférence EKAW. Ma place est originale car peu de chercheurs français sont visibles et actifs à ce niveau. Depuis peu, j'ai renforcé cette dimension par des séjours courts dans des laboratoires étrangers (LAO à Trento en 2007, KSL à Stanford et universidad de Murcia en 2008 et 2009), que j'envisage de poursuivre et de diversifier (contacts avec le DFKI en Allemagne) en vue de monter un projet international sur l'extraction de relations à partir de textes et sur la représentation des connaissances pour des ontologies à composante terminologique multilingue.

2 PROJET DE RECHERCHE : MODELISATION DES CONNAISSANCES POUR ET PAR L'ACCES AU CONTENU TEXTUEL

2.1 Objectifs généraux

La dynamique de la double articulation entre textes et connaissances, les textes étant vus comme des sources de construction de modèles comme les ontologies, et en retour, les modèles comme des supports à la fouille de textes ou à la description de leur contenu, est un des sujets de recherche de l'équipe IC3[1] (Ingénierie des Connaissances, de la Coopération et de la Cognition) de l'IRIT dont je suis responsable depuis septembre 2007.

Intitulé "Modélisation de connaissances pour et par l'accès au contenu textuel", mon programme de recherche se situe naturellement dans le prolongement de ces travaux et problématiques. Il se focalisera d'une part sur des questions théoriques : étude de l'articulation entre modèles et contenus documentaires avec un regard interdisciplinaire, sémantique en jeu dans l'analyse de textes et la modélisation des connaissances, et ses conséquences en matière de représentation conjointe de connaissances et d'un lexique. Un deuxième objectif vise une contribution pratique, en termes de méthode et d'outils d'analyse et de modélisation, selon deux dimensions : la construction de différents types de modèles terminologiques ou ontologiques à partir de textes et des activités humaines ; leur utilisation pour la recherche d'information dans des documents, l'analyse de contenus textuels ou l'annotation sémantique.

Les techniques, outils et modèles à intégrer sont ceux fournis par le TAL, l'IA et la recherche d'information, mais aussi des résultats d'innovation et de création de l'équipe pour parvenir à des solutions utiles, intégrées et accordant toute sa place à l'analyste qui modélise puis à l'utilisateur du système final. Ce programme rejoint plusieurs des enjeux du web sémantique, et plus largement les difficultés à dépasser pour parvenir à une exploitation fine de l'océan d'informations disponibles sur internet ou dans les entreprises.

2.2 Motivations

2.2.1 Une demande croissante de descriptions sémantiques

Les défis relatifs à ces objectifs sont aujourd'hui bien identifiés et constituent un programme de recherche stimulant pour l'IC et pour l'informatique dans les années à venir. Un tel programme doit mobiliser des compétences en traitement automatique des langues (statistique et linguistique), en recherche d'information, en représentation des connaissances et en IA, tout en restant dans l'esprit de questionnement de l'IC, pour prendre le recul nécessaire sur la vogue actuelle des ontologies et des techniques du web sémantique.

De plus, une forte demande vient des entreprises mais aussi de la société pour disposer de services et outils pour accéder au contenu informationnel des textes, mais aussi de modèles conceptuels de domaine comme les ontologies. Cette demande émane en particulier du projet du Web Sémantique, et des équipes de recherche d'information qui explorent de nouvelles perspectives d'amélioration des réponses aux utilisateurs grâce à l'utilisation d'ontologies.

Il s'agit là d'une troisième motivation à mon projet de recherche : dépasser les limites du projet du web sémantique pour répondre aux besoins en souplesse, flexibilité et adaptabilité des modèles qu'expriment les utilisateurs et au caractère dynamique du contexte d'utilisation de ces modèles, dépasser la vague actuelle d'utilisation systématique d'ontologies pour parvenir à une réflexion critique et raisonnée visant une gamme de modèles adaptés à chaque type d'application.

[1] http://www.irit.fr/-Equipe-IC3-

2.2.2 Le programme de recherche de l'équipe IC3

Notre équipe IC3, grâce à ses réflexions sur la place et le statut des modèles, ainsi que ses collaborations dans l'IRIT, avec CLLE-ERSS, et des équipes nationales ou internationales, se situe dans une position privilégiée pour les aborder. La dynamique d'IC3, ses compétences et ses projets, les collaborations actuelles et les nombreuses collaborations à développer, tant au niveau de l'IRIT qu'au niveau international, offrent des perspectives très riches pour approfondir les méthodes et outils de modélisation des connaissances et d'accès au contenu textuel que nous développons. Il semble opportun de bâtir un projet cohérent pour toute l'équipe et de viser des résultats plus visibles et surtout plus faciles à diffuser et mutualiser avec des partenaires, en matière d'ontologies et accès au contenu textuel.

Le cœur des recherches de l'équipe IC3 concerne l'étude théorique et technologique de systèmes socio-techniques au sein de situations de travail coopératives et collectives. L'équipe se distingue par la mise en œuvre de solutions à base de connaissances (représentées explicitement ou non) dans les solutions techniques développées, et par la volonté que ces solutions répondent à la fois à des théories cognitives et à des solutions pragmatiques opérationnelles sur le terrain. De plus, l'équipe a défini et évalué de manière complémentaire des solutions éprouvées, robustes et représentationnelles pour des situations nominales et, des solutions basées sur des simulations, des études en environnement virtuel et simulé, l'étude des phénomènes de percolation, pour organiser des situations dégradées et faciliter la gestion de crise.

Dans le cadre du prochain plan quadriennal[2], l'équipe a l'ambition d'aborder ces deux types de situations et de développer des solutions (théoriques et techniques) permettant de gérer la rupture ou la continuité entre situations nominales et situation dégradées jusqu'à la gestion de crise lorsque les situations le nécessitent. Son projet comporte donc un volet théorique et un plus technique :

- Définir un cadre théorique permettant d'aborder la continuité entre approches représentationnelles et approches constructivistes
- Mettre en œuvre des solutions, modèles et méthodes relevant du cadre théorique choisi pour étudier :
    - L'interaction sociale en milieu virtuel
    - Les technologies d'aide à la prise de décision collective, et le passage de situations nominales à des situations de crise
    - L'ingénierie des ontologies dynamiques et individuelles, et leur apport à la description et à l'interrogation du contenu de documents textuels.

2.3 Modèles de domaines et gestion documentaire

Mon projet de recherche consiste à poursuivre la définition d'une approche d'ingénierie des connaissances faisant appel au TAL, à la linguistique et la terminologie, au service de la modélisation des connaissances pour et par l'accès au contenu textuel dans des domaines spécialisés. Cette approche s'intéresse au modèle conceptuel non seulement du point de vue de son utilisation finale mais aussi de sa construction et de sa capacité à être relié à des éléments linguistiques. Ce modèle doit donc inclure soit des éléments lexicaux (ou terminologiques) soit les outils qui permettent de mettre en relations des éléments de modèle et des fragments de texte. Le modèle est vu comme un moyen de rendre explicites des connaissances et comme un support pour une construction collective du sens par les acteurs concernés par son utilisation.

Dans la continuité de mon expérience de recherche, ce projet d'ingénierie intègre l'étude conjointe de trois facettes : représentation des connaissances, outils et techniques d'aide, méthodologie.

A plus long terme, l'objectif sera de donner les moyens de définir des modèles conceptuels très flexibles, évolutifs, s'adaptant à leur contexte d'utilisation, et de les associer facilement à des textes, qui puissent être intégrés comme des représentations opératoires dans des logiciels d'aide,

[2] http://www.irit.fr/-Projet-de-recherche-2011-2014-

individuelle ou collective, à la réalisation de tâches, à la prise de décision ou d'accès à l'information.

2.3.1 Modèles comme support à la co-construction du sens

Dans la continuité des motivations et choix des recherches menées jusqu'ici (en particulier du projet Dynamo), la notion de modèle conceptuel est considérée non pas comme une structure statique, mais comme une représentation pouvant être remise en question et évoluer. La construction d'un modèle conceptuel ne se définit pas comme un processus linéaire mais plutôt comme un cycle au cours duquel le modèle doit être adapté pour prendre en compte un nouveau contexte d'utilisation, de nouvelles connaissances ou terminologies, de nouveaux textes, etc. Ainsi, le modèle est vu comme le résultat d'une série d'explicitations, d'interprétations, d'ajustements et de décisions de représentation.

2.3.2 Les besoins en matière d'accès au contenu documentaire

L'analyse des textes sur le web suppose de gérer la profusion de documents électroniques qui règne sur le web ou au sein des entreprises. Afin de renouveler les approches que nous avons développées jusque là, nous prévoyons de confronter et exploiter la complémentarité d'approches dont nous avons étudié en partie les atouts :

- web sémantique / web social : au-delà du réflexe premier de formaliser les folksonomies, l'idée serait d'adapter au processus d'évolution des ontologies, la démarche participative favorisant la négociation et l'interprétation du sens des concepts, entre autres au moyen d'exemples d'utilisation. L'atelier IC2.0[3] organisé à IC 2008 a permis un premier débat sur ces questions, qui sont traitées dans de nombreux ateliers internationaux.
- modèles formels et précis / thésaurus et vocabulaires : au niveau des méthodes de construction, nos expériences nombreuses nous permettront de préciser au sein de DAFOE 1) quand et comment exploiter certaines ressources et sources de connaissances (expertise, textes, modèles existants) pour amorcer ou compléter un modèle en cours de construction, 2) quel type de modèle prévoir pour un type d'application donné, en particulier définir son degré de formalisation.
- analyses linguistiques / analyses statistiques : un travail a été amorcé pour caractériser les genres textuels adaptés à certaines méthodes d'extraction de termes ou de relations. Une étude plus approfondie devra être poursuivie pour spécifier comment combiner certaines approches linguistiques et statistiques de manière à compenser les points faibles de chacune, ou bien pour trouver des techniques adaptées à la nature des textes à analyser.
- niveaux dans les analyses linguistiques : plusieurs plates-formes facilitent l'analyse linguistique de textes et leur annotation à différents niveaux, de la morphologie à la sémantique voir aux relations discursives (Plates-formes Gate, Alvis, LinguaStream, AnnoDis[4] …) ; ces différents niveaux sont de plus en plus utilisés en extraction d'information (Nédellec *et al.*, 2005), mais, à cause de la relative complexité de leur manipulation, ils sont sous-exploités pour la construction de modèles.
- approches par domaine spécialisé / approches langue générale : jusque là, nous avons fait le choix d'étudier des modèles portant sur des domaines spécialisés, et des collections de textes spécialisés, qu'elles soient volumineuses ou pas. Au-delà d'applications générales, l'utilisation de ressources générales peut être pertinente dans des domaines spécialisés pour situer les connaissances spécifiques (Baziz, 2005).

Ce projet passe par l'identification et éventuellement la maîtrise des éléments de complexité. Parmi les verrous attendus, certains sont propres à la langue : multilinguisme, polysémie, ambiguïté, anaphores, variabilité ; d'autres découlent de la nature des corpus et des documents qui les composent : variabilité, format, genre, fiabilité des sources ; enfin certaines difficultés sont liées à la

[3] http://apassant.net/home/2008/05/ic/
[4] http://w3.erss.univ-tlse2.fr/annodis

nature des modèles à construire : adéquation type de modèle / type d'application ; aides à l'extraction de tous les composants d'une ontologie, y compris les axiomes et règles ; degré de formalisation souhaitable.

2.3.3 Des modèles aux textes, des textes aux modèles

Ce projet étudie l'articulation entre textes et modèles de connaissances selon deux directions : des textes vers les modèles (construction de modèles à partir de textes) et des modèles vers les textes (extraction ou recherche d'information), car la prise en compte de ces deux facettes d'un même axe donne un éclairage complémentaire et évite de plaquer des solutions techniques a priori. De fait, une grande partie des problématiques et des questions fondamentales s'avère commune, avec des réponses parfois différentes selon qu'il s'agisse de construire des modèles à partir de textes ou d'utiliser ces modèles pour extraire des informations ou caractériser le contenu de textes.

Des textes vers les modèles : textes comme une des sources de connaissances pour la construction de modèles

Objectifs liés aux techniques et outils d'analyse automatique des textes

- Diversifier et intégrer les approches existantes en matière d'extraction de relations sémantiques : étendre les analyses à des relations autres que binaires, à des relations exprimées sur plusieurs phrases (en particulier en biologie) (Kamel, 2008), parvenir à une sémantique plus précise (collaborations avec Lilac, le LOA et le LIUPPA sur les relations spatiales et temporelles) (Aurnague *et al.,* 2007) (Sallabery *et al.,* 2007) et plus formelle des relations (graphes conceptuels) ; intégrer ces techniques dans des modules logiciels associés à la plate-forme DAFOE (Aussenac-Gilles et Hernandez, 2009) ;
- Repérer et extraire des connaissances plus complexes pour construire des ontologies dites « lourdes » : règles (collaboration avec le LIPN), axiomes et de contraintes (Hitzler *et al.,* 2007)

Objectifs méthodologiques

- Assurer une meilleure complémentarité des textes avec d'autres sources de connaissances : dans le cadre de DAFOE, un travail est en cours sur la complémentarité entre réutilisation de terminologies, d'ontologies ou de thésaurus et utilisation de textes ; au-delà, nous voulons mieux définir comment réutiliser des éléments d'ontologies (concepts ou relations entre concepts) ; une des pistes est d'utiliser des logiciels de recherche comme WATSON[5], OntoSearch[6] ou OntoSelect[7] ou le logiciel de réutilisation de relations SCARLET (Sabou *et al.,* 2008) ;
- Définir le processus de construction d'ontologie comme un processus cyclique et incrémental, en particulier en intégrant des connaissances tirées de nouveaux textes ; il s'agit de reprendre les résultats qui seront établis sur l'évolution d'ontologies dans les projets Dynamo et Arkeotek en cours pour assurer une cohérence entre l'utilisation des outils de TAL, l'analyse des connaissances ainsi obtenues et l'enrichissement / évolution d'ontologies ;
- Intégrer dans une méthode les moyens de définir puis de prendre en compte des critères de « qualité » de l'ontologie pour guider et stabiliser le processus ; ce travail a également débuté dans le cadre du projet Dynamo (thèse de A. Tissaoui) ;

Des modèles vers les textes : retrouver des éléments de connaissances dans des énoncés et des textes dans un contexte de recherche d'information

Processus d'annotation sémantique

[5] http://kmi-web05.open.ac.uk/WatsonWUI/
[6] http://www.ontosearch.org/
[7] http://olp.dfki.de/ontoselect/

- dépasser la vue naïve actuellement retenue dans IC3, basée sur la seule présence de termes ou de patrons pour associer des concepts à des textes ; notre expérience en matière d'utilisation de patrons et d'expressions régulières pour « fouiller » des textes se rapproche des techniques d'extraction d'information. L'idée est d'exploiter ce type de patron, et donc de les associer à l'ontologie.
- Dans la même idée, une autre piste est de tenir compte des différents niveaux et types d'annotation pour suggérer différentes manières d'associer termes et concepts, en fonction des besoins d'utilisation de l'ontologie ;

Objectifs liés aux ontologies utilisées dans des applications de recherche d'information

- Un premier type d'application que nous souhaitons étudier sont les systèmes de questions-réponses à partir de textes, et en particulier le rôle des ontologies pour faciliter le repérage d'opinions : une collaboration a commencé fin 2009 avec F. Benamara de l'équipe Lilac de l'IRIT (stage co-encadré) ;
- Une autre classe d'applications concerne la recherche d'informations précises dans des collections documentaires,
  - processus d'interrogation et de requête sur textes annotés : Les graphes conceptuels s'avèrent un mode de représentation efficace pour assurer l'interrogation d'ontologies ou de textes annotés ; ce formalisme est au cœur du travail d'O. Haemmerlé dans notre équipe, et a été expérimenté dans le projet ANR WebContent),
  - étude de la complémentarité entre informations linguistiques et distance sémantique entre concepts pour associer requêtes et documents ; en effet, les méthodes actuelles (qui s'appuient sur les noms des concepts) sont assez simplistes pour « reconnaître » la connaissance présente dans chaque phrase du texte ; une collaboration sur ce sujet est en cours avec l'équipe SIG-EVI de l'IRIT et le laboratoire Lalic (Paris 4) dans le cadre du projet ANR Dynamo.
- innover dans la manière d'exploiter les collections documentaires annotées : support à la confrontation de fragments de textes, à la recherche de similarités et de différences, liés à des points de vue (résultats scientifiques (Arkeotek)) ou à des évolutions dans le temps (CNES, collaboration avec CLLE-ERSS)

Problématiques communes

Objectifs liés aux techniques et outils d'analyse automatique des textes

- *Prendre en compte deux niveaux supplémentaires dans l'analyse des textes* : l'analyse syntaxique des dépendances syntaxiques (résultats de Syntex, Cordial ou Alpage) et l'analyse du discours (Perry-Woodley *et al.,* 2009) ; et des éléments relatifs à la mise en forme matérielle pour en tirer des informations rhétoriques ou sémantiques, exploitées ensuite comme indices pour repérer des noms de concepts ou des relations sémantiques (travaux de J. Virbel sur les titres et des énumérations (Eyrolles *et al.,* 2008)) ;
- *Prise en compte de la variabilité* au sein des textes, des genres textuels mais aussi des objectifs d'utilisation de l'ontologie, des annotations ou des connaissances extraites : pour la construction d'ontologie, l'idée serait de permettre une démarche plus paramétrable et plus coopérative, qui propose une aide pertinente à l'utilisateur à chaque étape de processus en fonction des sources de connaissances exploitées et de l'objectif de modélisation ; pour l'analyse et la caractérisation du contenu de textes à l'aide d'une ontologie, il s'agit de prendre en compte la nature des annotations en fonction des contextes (web sémantique ou web social) et non pas seulement leur différence de forme (représentation, langage choisi)

Représentation des connaissances

- Le modèle retenu pour les ressources termino-ontologiques permet d'associer des éléments linguistiques (des termes et des informations associées) aux concepts d'une ontologie. Cette

solution s'avère insuffisante pour retrouver des variantes de formes et surtout pour gérer la combinaison de termes qui peut renvoyer ou non à de nouveaux concepts. Elle présente également un manque pour ce qui est de l'équivalent linguistique des relations. Une première étude en cours fait ressortir une ambiguïté dans le statut des termes selon que l'on se place dans la perspective de construire une ontologie (termes désignant des concepts) ou bien d'annotation sémantique (terme indice de la présence/mention d'un concept dans un texte). La confrontation des deux perspectives est éclairante. Cette étude doit être poussée plus loin pour répondre aux besoins du multilinguisme, de l'annotation automatique ou encore pour gérer les particularités de langues autres que l'anglais ou le français (Buitelaar *et al.*, 2009). Pour cela, nous avons entrepris des collaborations nationales (linguistes de CLLE-ERSS, informaticiens du LIPN et INSERM au sein de DAFOE4App et Dynamo), et internationales avec les équipes concernées par la représentation des termes dans les ontologies, en participant au workshop OntoLex, en accueillant Th. Declerck du DFKI, en collaborant avec P. Buitelaar du DERI et P. Cimiano (Univ. d'Amsterdam). Ces équipes envisagent de monter un projet européen pour contribuer à la définition d'un standard pour la représentation d'ontologies à composantes terminologies.

Dynamique et évolution des connaissances

- L'étude conjointe de la dynamique d'ontologies et de collections documentaires annotées est à la fois un enjeu clé pour les applications du web sémantique, et un contexte tout à fait prometteur pour étudier comment rendre compte ou traiter informatiquement l'articulation entre langue et connaissance, entre textes et ontologies. Ce travail est amorcé avec le projet Dynamo, qui permet de dessiner en perspective des questions à approfondir :
  - dépasser la vision d'une ontologie statique pour tenir compte de la dynamique des connaissances dans les logiciels et méthodes permettant de les gérer (on retrouve les questions méthodologiques) ;
  - mieux assurer cohérence des modèles (ontologies, RTO), des collections documentaires et des annotations de documents faisant appel à ces modèles
  - capacité des modèles de type ressources termino-ontologiques à permettre de déceler des changements et des évolutions dans les connaissances ou la terminologie d'un domaine à partir de collections documentaires ; capacité de ces modèles à en rendre compte (on retrouve ici les questions relatives à la représentation des connaissances.

Pérennisation et diffusion des logiciels

- Cet objectif est propre au domaine de l'IC où les logiciels et outils méthodologiques doivent être évalués en vraie grandeur pour valider des innovations, et commun à toute recherche en informatique, du fait d'une certaine pénurie d'ingénieurs pour accompagner le travail des chercheurs. Cet objectif part d'un constat partagé par plusieurs équipes françaises, qui ont du mal à faire vivre et à diffuser leurs logiciels et plus généralement leurs résultats. Notre objectif est de définir des logiciels modulaires et adaptables (par exemple sous forme de plug-in associés à des plates-formes comme DAFOE ou WebContent), et de se concerter avec d'autres équipes pour valoriser la complémentarité de nos travaux. Des contacts ont été pris dans cet objectif avec les partenaires des projets auxquels nous participons mais aussi avec l'équipe INRIA Edelweiss. La plus-value de ce travail devrait être avant tout de mieux répondre à la variabilité des formes et de la sémantique au sein des textes et entre genres textuels, de mieux gérer les différents types d'ontologies à construire en fonction des applications, d'assurer une meilleure ergonomie et adéquation de ces logiciels aux usages prévus , et enfin de les situer au sein d'un processus coopératif (et non passif comme le sont les plates-formes actuelles) d'aide à l'ontographe.

2.3.4 Premières étapes du projet de recherche

Le domaine de l'ingénierie des connaissances (et actuellement des ontologies) est particulièrement actif, il s'intéresse à des applications et technologies en évolution. Dans ce cadre, fixer les détails d'un programme de recherche à trop long terme n'a pas de sens. Néanmoins, à long

terme, l'idée directrice de mon projet est de s'intéresser à l'articulation langue/ textes/ modèles de connaissances dans des domaines spécialisés selon un point de vue d'ingénierie des connaissances, et cela en déclinant cette problématique en fonction des applications, technologies et types de modèles pertinents. Ce projet est incrémental et consiste à aborder progressivement différentes applications, terrains et types de corpus, de capitaliser des pratiques, éléments méthodologiques et logiciels, pour déboucher sur une ou plusieurs propositions les intégrant. Deux tensions le soutiennent donc : un effort de découverte et mise en œuvre de nouvelles techniques et de nouveaux logiciels, de diversification et de problématisation de cette variation ; en même temps, un souci d'intégration et de synthèse de ces solutions au sein de démarches unifiées, et surtout une volonté d'assoir théoriquement ces recherches.

A moyen terme, les deux premières phases de ce travail seraient les suivantes :

Phase 1

- pousser plus loin le travail sur l'extraction de relations selon les différentes directions envisagées ; rendre ces résultats disponibles sous forme de différents modules complémentaires de la nouvelle plate-forme DAFOE ; monter un projet s'appuyant sur les collaborations nationales et internationales (projet ANR ou international)
- poursuivre les études liées aux projets ANR en cours, reliées aux problématiques suivantes : évolution des connaissances et des terminologies et de l'identification d'évolutions en corpus ; gestion conjointe des ontologies et d'annotations sémantiques ; repérage de relations spatiales ou temporelles ; représentation d'ontologies et d'éléments linguistiques associés ;
- étudier les méthodes et techniques requises pour construire des ontologies utilisées des divers types de systèmes de recherche d'information : exploitation des modèles pour l'annotation sémantique pour la recherche d'opinion, la confrontation scientifique, etc. Proposer un langage pour formuler des critères de qualité d'un modèle par rapport à une collection, et la qualité des annotations produites (toujours en collaboration avec des chercheurs en RI)
- amorcer plusieurs réflexions pour intégrer les différentes propositions présentes dans IC3
    - définir la construction du modèle comme un processus de décision collaborative, travail avec le groupe « systèmes coopératifs » de l'équipe IC3
    - étudier la capacité des ontologies à rendre compte de points de vue et ce que pourraient être des ontologies « personnelles ».

Phase 2

- intégrer les résultats de nos différents projets en cours (GEONTO, Arkeotek**,** DYNAMO et DAFOE4App), et de projets sur l'annotation au niveau discours (comme ANNODIS) en matière de représentation des connaissances ; de prise en compte de l'évolution ; de mise en forme de patrons exploitant ou produisant différents niveaux d'annotation ; d'extraction de relations sémantiques et de repérage de concepts à partir d'éléments linguistiques ;
- proposer des évolutions méthodologiques de manière à intégrer les contributions retenues au sein de l'équipe IC3 et suite à des collaborations avec des linguistes :
    - sur la dimension collaborative dans la construction et mise à jour de modèles conceptuels et d'ontologies ;
    - vers des ontologies « individuelles », la prise en compte de points de vue personnels.

2.4 Contributions envisagées

2.4.1 Extraction de relations à partir de textes

Vers une conception adaptative des systèmes de traitement automatique

La mise au point de patrons réutilisables pour Caméléon a soulevé la question de la variabilité des résultats d'un même traitement appliqué à différents genres de textes. La plupart du temps, les recherches en matière de traitement automatique, lorsqu'elles présentent des analyses sur corpus, indiquent des résultats obtenus sur seulement UN corpus ou un ensemble indistinct de textes. Or il est désormais possible de faire entrer en jeu les nuances requises par la diversité des textes. Cette problématique émerge dans des travaux récents en traitement automatique des langues (influence des corpus sur les performances de stratégies de rattachement prépositionnel), dans le domaine de la recherche d'informations (variété de l'efficacité des requêtes), en linguistique de corpus (mise au point d'outils de typologie textuelle). Notre étude apporte ici une pièce supplémentaire en faveur d'une conception diversifiée et adaptative des traitements automatiques. Cette analyse a été développée pour le cas du repérage des relations sémantiques dans un article écrit avec Anne Condamines (2009).

Le programme de recherche envisagé pour l'extraction de relations se dessine clairement suite à un séminaire d'équipe, à la venue à l'IRIT de T. Declerck et à un exposé de notre à la réunion annuelle du laboratoire ILIKS (Aussenac-Gilles, Kamel et Hernandez, 2008).

Travaux en cours

GEONTO : expérimentation de GATE (plate-forme ouverte, règles pour exprimer les patrons, éditeur d'ontologie associé comme un des plug-in), identification automatique de relations et de concepts, prise en compte de la structure de documents, prise en compte de la mise en forme (énumérations, polices)

TAT-CG : problématisation de la recherche de relations n-aires ; pistes pour rechercher des relations exprimées sur plusieurs phrases (étude linguistique de la résolution des anaphores à l'aide de connaissances ; utilisation de graphes conceptuels)

DAFOE4app : spécification de la part des traitements prise en charge par le module d'extraction de relations et celle intégrée à la plate-forme de modélisation ; définition d'écrans et des étapes de l'ensemble du processus ; développement d'un module (plug-in) d'extraction de relations sémantiques à l'aide de patrons, basé sur le système Caméléon-III.

Vers une plate forme de recherche de relations sémantiques

objectifs : disposer d'une plate-forme publique et facile à faire évoluer

- Logiciel qui facilite les collaborations, la capitalisation de patrons, la réutilisation ;
- disposer d'une plate-forme permettant à la fois de rechercher des relations entre concepts pour construire une ontologie et des relations entre instances pour la peupler ou pour réaliser une annotation sémantique
- disposer de ressources (patrons et typologie de relations) pour le français

Intégrer des capacités *d'apprentissage* pour le repérage des patrons :

- guider l'apprentissage automatique de nouveaux patrons, module d'apprentissage réalisé en 2008 par un étudiant M2R
- intégrer les résultats du séjour post-doctoral de M. Chagnoux : guider la mise en forme de patrons à partir de l'observation de phrases … (articles EKAW 2008 et OLP 2008)

Favoriser la *réutilisation*

- pouvoir intégrer des patrons mis au point par d'autres équipes

- réutiliser des relations venant d'autres ontologies (articles EKAW 2008 et OLP 2008) à l'aide de logiciels de recherche d'ontologies sur le web comme Watson et SCARLET (Sabou et al., 2008)) ou Ontosearch développé au DFKI P. Buitelaar
- s'appuyer sur une plate-forme ouverte d'analyse du langage, comme GATE[8] (dont nous avons l'expérience et qui a servi à défini un outil de ce type, ANNICK) ou LinguaeStream[9], développée à Caen, ou ALVIS[10], développée dans un projet européen auquel participe le LIPN

étendre la notion de relation

- relations autres que binaires : projet TAT-GC (travaux de M. Kamel (IC3) et GEONTO (Van Tien Nguyen au LIUPPA)
- repérage des axiomes, des contraintes, reprendre l'approche proposée par Volker, Cimiano, 2008
- vers une sémantique plus précise des relations spatiales et temporelles : collaboration avec Laure Vieu et P. Muller, collaboration avec le LIUPPA pour la description d'itinéraires
- relations entre concepts (pour enrichir une ontologie) et relations entre instances de concepts (pour peupler une ontologie ou définir une annotation sémantique)

étendre la notion de patron

- extraction de relations sur plusieurs phrases : travaux de M. Kamel, du LIuPPA et projet de collaboration en cours avec l'ERSS
- prise en compte de la mise en forme matérielle et de la structure des textes : amorcé dans GEONTO
- mieux exploiter les différents niveaux d'annotation des textes : ceux qui seront définis dans le projet ANR ANNODIS où collaborent l'IRIT et CLLE-ERSS

2.4.2 Vers une plate-forme de construction de RTO

Travail en cours : le projet DAFOE4App

Au plan expérimental, il s'agit de développer, en collaboration avec plusieurs équipes de recherche françaises, une plate-forme de modélisation permettant à la communauté de l'IC de mutualiser, d'expérimenter et de valoriser ses outils et les points de vue originaux qu'elle défend. Le projet DAFOE4App[11] vise cet objectif, en se focalisant sur l'acquisition de connaissances à partir de texte à laide de logiciels de TAL. Mais il ne sera que partiellement atteint au terme du projet. Deux difficultés se présentent que nous avions sous-estimées :

- **la représentation des informations linguistiques associées à une ontologie :** La question fondamentale est de savoir jusqu'où l'ontologie, objet livré à la fin du processus, s'enrichit d'informations linguistiques. La frontière est difficile à identifier entre des informations linguistiques gérées provisoirement au cours du processus de modélisation, et donc inutiles à conserver, et les éléments linguistiques pertinents pour documenter l'ontologie, ceux pertinents pour en justifier la structure et le contenu, enfin ceux qui faciliteraient l'association entre des éléments d'ontologies et des textes. Une fois la réponse apportée à cette question, la représentation structurée ou même formelle peut être rapidement définie.
- **l'articulation entre analyse et fouille de texte, d'une part, et, d'autre part, analyse d'autres sources de connaissances** au sein du processus de modélisation des connaissances : la complémentarité entre sources de connaissances est encore vue comme

[8] http://gate.ac.uk/
[9] http://www.linguastream.org/
[10] http://linux.softpedia.com/progDownload/Alvis-NLPPlatform-Annotation-Download-44801.html
[11] http://dafoe4app.fr/

une juxtaposition de tâches monolithiques dont seuls les résultats doivent être intégrés au sein de l'ontologie ; au contraire, les pratiques actuelles de réutilisation d'ontologies ou de thésaurus, d'analyse de textes, et d'intervention d'expertise humaine doivent être intégrées dans Dafoe et devraient éviter de fragmenter la contribution de chacune des sources de connaissances. Cependant, cette complémentarité est très dépendante de l'application ou du corpus, et ne peut être figée a priori.

Associer connaissances et formulations linguistiques

Le projet DAFOE4App a fait un choix adapté des travaux de A. Reymonet, le projet Dynamo reprend également cette réflexion dans un contexte différent. Il sera intéressant de voir si les conclusions différent, et si cela est dû à des perspectives d'utilisation différentes des ontologies. Il semblerait que la difficulté réelle soit bien là : une réponse unique semble difficile à établir et il serait plus pertinent de permettre de paramétrer la nature des informations linguistiques à associer à l'ontologie en fonction des objectifs de cette ontologie.

Ce besoin de souplesse et d'adaptabilité des modèles s'impose également dès que l'on imagine l'utilisation faite des éléments linguistiques associés aux concepts et aux relations. Plus on prévoit de stocker des informations figées, des données comme des termes, plus il faudrait sophistiques l'outillage qui permet d'en retrouver les différentes variantes dans les textes, lorsqu'un adjectif ou un adverbe se glisse à l'intérieur d'un terme par exemple. A l'inverse, si on stocke de petits automates ou des patrons (MA *et al.*, 2009), la distance entre les formes linguistiques et les représentations peut être plus grande, on gagne en souplesse et en efficacité, on évite de multiplier les formes stockées (ces formes seront calculées). Reste à savoir comment enregistrer ces automates ou patrons, sans dénaturer ou alourdir le modèle et tout en gérant simplement le lien entre ces informations et le modèle.

Mes travaux en cours et les perspectives actuelles permettront d'approfondir à la fois une vue statique et figée sur cette articulation entre représentation des connaissances/ éléments linguistiques, et une plus dynamique, liée aux traitements faits sur les modèles.

Concernant les aspects « statiques » liés à la nature des modèles de données au sein des RTO, il semble utile de :

- Poursuivre les évaluations des solutions actuelles par des utilisations dans les différents projets en cours où des ontologies avec une composante lexicale sont sur le point d'être utilisées ;

- Prendre en compte de réflexions sur l'articulation entre l'ontologie formelle et une composante terminologique menés par A. Gangemi en lien avec le LOA. Ce travail trouvera sa place dans le laboratoire ILIKS.

Concernant les aspects plus « dynamiques » consistant à associer des capacités de traitement linguistique et non des résultats aux modèles, les réflexions en cours orientent le travail dans deux directions :

- un besoin de standardisation dans la définition de patrons de fouille de texte : en effet, les patrons (tels qu'ils sont définis pour l'extraction d'information ou la recherche de relations sémantiques) permettraient de prendre en compte aussi bien que la variation terminologique, la variation dans l'expression des relations ;
- la nécessité d'une méta-modélisation permettant d'ajuster les structure de représentation des informations (conceptuelles et linguistiques) au sein d'une RTO, et ce à différents niveaux de formalisation.

2.4.3 Utilisation des ontologies pour l'accès au contenu documentaire

Rappelons que ce qui nous intéresse dans l'utilisation des ontologies pour l'accès au contenu documentaire, c'est ce que ces applications fixent comme contraintes sur les modèles qui leur sont les plus pertinents. Le fait de décrire un contenu documentaire à l'aide de concepts suppose de pouvoir établir la validité de l'ontologie qui les définit par rapport au domaine et au corpus de

documents, de savoir associer des fragments de ce document à des sous-ensembles d'ontologie (graphes partiels d'instances par exemple).

Les résultats attendus en la matière concernent donc la représentation des connaissances dans les ontologies, et surtout la manière de gérer le lien entre ontologies et documents, entre ontologies et éléments linguistiques. Nous avons développé plus haut les résultats attendus sur ce point.

En particulier, nous étudierons l'annotation des documents structurés, et l'exploitation de la structure (implicite ou explicite) liée à la rhétorique du discours qu'ils contiennent pour repérer concepts, instances de concepts ou relations sémantiques.

Une autre perspective est de diversifier les exploitations faites des annotations, au-delà de l'interrogation par requête. Dans le cadre du projet Arkeotek, une consultation « intelligente » rapprochant des passages de documents ou des documents proches est envisagée, le but étant d'aider les chercheurs d'un domaine à faire des recoupements inattendus, à naviguer dans la collection selon les concepts du domaine et ce qu'il en est dit, comme dans le cas de recherche de nouveauté en extraction d'information.

Un autre type de résultat attendu concerne notre participation à l'axe « masse de données et calcul » de l'IRIT. Cette perspective oriente nos travaux vers le traitement de gros volumes de données tirées du web.

2.4.4 Dynamique des modèles et des terminologies

Nous avons déjà développé une autre problématique dégagée de nos travaux précédents : celle de la maintenance des modèles en cohérence avec le vocabulaire et les connaissances du domaine, avec des collections de textes, ainsi que des index ou annotations utilisant ces ontologies et ces textes ; c'est là un des enjeux des projets Corpus Logicistes et DynamO. Le contexte d'usage des ontologies, et ceci est encore plus criant dans le cas du web, est d'évidence en évolution permanente. Or les ontologies sont souvent considérées comme des représentations stables puisque consensuelles. Nous étudions les évolutions d'ontologies en fonction des évolutions du contexte de leur utilisation. Une de nos études sur ce thème (menée dans le projet Dynamo) s'appuie sur de la fouille de textes à l'aide d'agents adaptatifs.

Comme nous l'avons souligné plus haut, les résultats attendus et en cours de réalisation concerne la définition d'un processus d'évolution d'ontologie qui permette à l'ontographe de maîtriser le modèle construit, de l'enrichir en fonction des besoins d'utilisation.

Du point de vue de l'analyse de la langue pour l'étude de la dynamique des connaissances, nous pensons reprendre des travaux sur le repérage de nouveauté (thèse de Marion Laignelet et thèse d'Aurélie Picton). Pour juger de l'adéquation modèle / corpus, des critères de couverture d'un corpus par une ontologie (et les moyens de les mesurer) ainsi que des critères de qualité d'annotations, qui peuvent conduire à remettre en question l'ontologie ou sa composante linguistique, sont à l'étude pour des ontologies dynamiques (projet Dynamo).

Grâce au projet terminé mené avec le CNES, nous disposons d'outils statistiques pour analyser de gros volumes de documents, identifier des communautés thématiques associées à des domaines, et des termes clés représentatifs de ces domaines. Dans des domaines qui évoluent, nous étudierons la complémentarité entre ces approches statistiques (text mining) et la modélisation à l'aide d'ontologies pour déterminer les frontières de domaines spécialisés puis les évolutions des frontières de ces domaines ou de leur « contenu ».

2.5 Conclusion

Ma position actuelle au sein de l'IRIT ainsi que l'expérience que j'ai acquise au cours de mon activité de chercheur CNRS me permettent donc d'établir un projet de recherche en ingénierie des connaissances qui soit original, sur un sujet peu représenté au sein du CNRS, bien que de nombreuses équipes françaises et internationales s'y intéressent. Ce projet est à la fois personnel et ancré dans celui de l'équipe IC3 que je dirige. Il se situe aujourd'hui au cœur des enjeux de l'évolution du web mais aussi de l'accès au contenu informationnel des documents numériques. De

par mon expérience de collaboration avec des linguistes, des ergonomes mais aussi avec des informaticiens en recherche d'information, et mon intérêt pour les sciences cognitives, j'envisage de porter ce projet en complément avec ces domaines de recherche, et non comme le seul problème de l'ingénierie des connaissances. Grâce à mes contacts au niveau national et international, j'ai le souci de donner à ce projet une cohérence avec les résultats nationaux et une qualité de niveau international.

3 REFERENCES

ASHER N., BENAMARA F., MATHIEU Y., Appraisal of Opinion Expressions in Discourse. Dans : *Lingvisticæ Investigationes*, John Benjamins Publishing Company, Amsterdam, Vol. 32:2, 2009 (à paraître).

AURNAGUE M.,HICKMANN M., VIEU L., *The Categorization of Spatial Entities in Language and Cognition*, John Benjamins Publishing Company, Vol. 20, Human Cognitive Processing, 2007.

AUSSENAC-GILLES N., CHAGNOUX M., HERNANDEZ N., « An Interactive Pattern-Based Approach for Extracting Non-Taxonomic Relations from Texts», *Pacific Graphics, Patras, Greece, 22/07/2008*, P. Buitelaar, P. Cimiano, G. Paliouras, M. Spilliopoulou (Eds.), *proceedings of the ECCAI workshop OntoLex08 - From Text to Knowledge: The Lexicon/Ontology Interface*, p. 1-6, 2008.

AUSSENAC-GILLES, KAMEL, HERNANDEZ N., Towards a platform for supervised relation extraction from text. *Interdisciplinary Laboratory on Interactive Knowledge Systems (ILIKS) 2008 annual meeting, Toulouse (F), 01/12/2008-02/12/2008*, Laure Vieu (Eds.).

AUSSENAC-GILLES, N. and HERNANDEZ N.,. Du li**n**guistique au conceptuel : identification de relations conceptuelles à partir de textes, in : *Atelier « Relations sémantiques » associé à TIA2009*, N. Grabar et S. Deprès (eds.), 2009.

BAZIZ M., BOUGHANEM M., AUSSENAC-GILLES N., CHRISMENT C., Semantic Cores for Representing Documents in IR, *Proceedings of SAC-IAR'05, the ACM SAC Track on Information Access and Retrieval*. Santa Fe (NM, USA), 1011 – 1017. 2005.

BAZIZ M., BOUGHANEM M., AUSSENAC-GILLES N., A Conceptual Indexing Approach based on Document Content Representation, P*roceedings of COLIS 2005 (5th International Conference onConception of Libraries and Information Science)* - Context: nature, impact and role . Univ. Of Strahclyde, Glasgow (UK), July 2005. F. Crestani and I. Ruthven (Eds.): LNCS 3507. Berlin : Springer-Verlag . 171-186. 2005.

P. BUITELAAR, P. CIMIANO, P. HAASE, AND M. SINTEK. Towards linguistically grounded ontologies. *In Proc. of ESWC, 2009.*

BUITELAAR P., CIMIANO P., MAGNINI B., Ontology Learning From Text: Methods, Evaluation and Applications, IOS Press, 2005.

CIMIANO P., *Ontology Learning and Population from Text. Algorithms, Evaluation and Applications*. Berlin: Springer. 2007.

CONDAMINES A., Comment peut-on construire une ontologie personnelle à partir de textes ? considéérations linguistiques, *actes de la conférence TIA 2009*, Toulouse (F). 2009.

DECLERCK T., LENDVAI P., Towards a model for improving interoperability of conceptual, terminological and linguistic resources, In Proc. Of ICGL 2010, Hong-Kong. (to appear)

EYROLLE H., VIRBEL J., LEMARIÉ J., Effect of incomplete correspondance between document titles and their text on users representations: A cognitive and linguistic analysis based on 25 technical documents. Dans : *Applied Ergonomics*, Elsevier Science, Vol. 39, p. 241-246, février 2008.

JUNG J.J., Ontology-based context synchronization for ad hoc social collaborations. *Knowledge-Based Systems*, vol.21, issue 7. p. 573-580. 2008.

KAMEL M., Une proposition pour l'extraction de relations non prédicatives. *EGC - Atelier Extraction et Gestion Parallèles Distribuées des Connaissances, Sophia-Antipolis, 29/01/2008-01/02/2008*, Cépaduès, p. 215-216, 2008.

KAMEL M., PERRET E., Extraction d'Information pour le ciblage des gènes impliqués dans les maladies génétiques. *Journées Ouvertes Biologie, Informatique et Mathématiques (JOBIM 2007), Marseille, 10-12/07/2007*, Cépaduès, 2007.

KIM S., HALL W. & A. KEANE A., Using document structures for personal ontologies and user modelling. In, *Proceedings of User Modeling 2001: 8th International Conference, UM 2001, London (UK)*, Springer (Lecture Notes in Computer Science 2109), 240-242, 2001. http://eprints.soton.ac.uk/21884/

MA Y., AUDIBERT L., NAZARENKO A., Ontologies étendues pour l'annotation sémantique. Actes d'IC 2009: 205-216. 2009

MAEDCHE A., *Ontology learning for the Semantic Web*, volume 665. Kluwer Academic Publisher, 2002.

NÉDELLEC C., NAZARENKO A., Ontology and Information Extraction: A Necessary Symbiosis, *Ontology Learning from Text: Methods, Evaluation and Applications*. Volume 123 Frontiers in Artificial Intelligence and Application, P. Buitelaar, P. Cimiano, B. Magnini (eds.), IOS Press, 2005.

PERY-WOODLEY M.-P., ASHER N., ENJALBERT P., BENAMARA F., BRAS M., FABRE C., FERRARI S., HO-DAC L.-M., LE DRAOULEC A., MATHET Y., MULLER P., PREVOT L., REBEYROLLE J., TANGUY L., VERGEZ-COURET M., VIEU L., WIDLÖCHER A. (2009) ANNODIS : une approche outillée de l'annotation de structures discursives (poster), TALN'09, 24-26 june 2009, Senlis, France.

REYMONET A., THOMAS J., AUSSENAC-GILLES N., Modélisation de Ressources Termino-Ontologiques en OWL, *18èmes journées francophones d'Ingénierie des Connaissances (IC 2007). Grenoble (France), 4 au 6 Juillet 2007*. F. Trichet (Ed.), Cépaduès Editions, 169-180, 2007a (Prix AFIA meilleur article de la conférence).

REYMONET A., THOMAS J., AUSSENAC-GILLES N., Modelling Ontological and Terminological Resources in OWL-DL. *OntoLex07 - From Text to Knowledge: The Lexicon/Ontology Interface - Workshop at ISWC07 – 6th International Semantic Web Conference, Busan (South Korea), 11/11/2007*, P. Buitelaar, K.-S. Choi , A. Gangemi, C.-R. Huang (Eds.), http://olp.dfki.de/OntoLex07/, 2007b.

ROTHENBURGER B., Du mode de prise en compte ontologique et terminologique de l'évolution des connaissances dans les domaines techniques. Revue *Information - Interaction - Intelligence*, Cépaduès Editions, Numéro spécial *: Des documents aux connaissances : évolution et maintenance dans les textes, les terminologies et les ontologies*, Vol. Hors-série, p. 9-29, 2006.

SABOU M., D'ACQUIN M., MOTTA E., SCARLET: SemantiC RelAtion DiscoveRy by Harvesting OnLinE OnTologies, in The Semantic Web: Research and Applications Proceedings of ISWC 2008, Springer : Berlin / Heidelberg, LNCS Vol. 5021, 854-858. 2008.

SALLABERRY C., BAZIZ M., LESBEGUERIES J., GAIO M., Towards an IE and IR System Dealing with Spatial Information in Digital Libraries - Evaluation Case Study. ICEIS (5) 190-197, 2007.

STAAB S., MAEDCHE A., Ontology Learning for the Semantic Web, *IEEE Intelligent Systems*, Special Issue on the Semantic Web, 16(2), p. 72-79, 2001.

VÖLKER J., HITZLER P., CIMIANO P.: Acquisition of OWL DL Axioms from Lexical Resources. *ESWC 2007*: 670-685. 2007.