

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

Janvier 2010

Rapport sur les activités antérieures

Méthodes ascendantes pour l'ingénierie des connaissances

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	Résumé	3
2	Curriculum Vitae détaillé	6
2.1	Données personnelles, emplois et titres universitaires.....	6
2.2	Participation à des projets	6
2.2.1	Contrats avec des entreprises (11)	6
2.2.2	Projets régionaux (4)	8
2.2.3	Projets nationaux	8
2.3	Administration et animation de la recherche à l'IRIT	10
2.3.1	Direction de l'équipe IC3 de l'IRIT	10
2.3.2	Collaborations scientifiques dans le laboratoire	13
2.3.3	Animation scientifique au niveau du laboratoire.....	15
2.4	Administration et animation de la recherche hors laboratoire	15
2.4.1	Animation de la communauté nationale	15
2.4.2	Collaborations scientifiques nationales	16
2.4.3	Animation au niveau international.....	19
2.4.4	Collaborations scientifiques au niveau international	19
2.4.5	Séjours à l'étranger.....	19
2.5	Activité d'enseignement	20
2.5.1	Enseignements en 3 ^e cycle et écoles d'ingénieur	20
2.5.2	Ecoles d'été.....	20
2.5.3	Tutoriels à des conférences et séminaires.....	21
2.5.4	Participation à des commissions de spécialistes.....	21
2.6	Activité d'encadrement.....	21
2.6.1	Direction de thèses terminées	21
2.6.2	Direction de thèses en cours	22
2.6.3	Participation à des jurys.....	22
2.6.4	Encadrement d'étudiants et stagiaires	23
2.7	Actions de diffusion de l'information scientifique et technique	23
2.7.1	Participation à des comités de programme	23
2.7.2	Organisation de conférences et journées scientifiques	25
3	Le domaine de recherche	27
3.1	L'ingénierie des connaissances.....	27
3.1.1	L'ingénierie des connaissances en informatique	27
3.1.2	Un domaine de recherche pluridisciplinaire	27
3.1.3	L'ingénierie des connaissances en France	27
3.2	Problématiques actuelles de l'ingénierie des connaissances.....	28
3.2.1	Le Web Sémantique.....	28
3.2.2	Modélisation de connaissances, ontologies et textes	30
4	méthodes ascendantes pour L'ingénierie des connaissances	31
4.1	Modélisation de connaissances expertes.....	31
4.1.1	MACAO, une méthode pour l'acquisition de connaissances expertes	31
4.1.2	MACAO-II, modélisation de connaissances et opérationnalisation	31
4.2	Modélisation à partir de textes	32
4.2.1	Une évolution thématique.....	32
4.2.2	Vers des ressources termino-ontologiques	32
4.2.3	Problématique et partis pris théoriques.....	34
4.2.4	Représentation des connaissances intégrant une composante terminologique	34
4.2.5	Des textes aux applications : méthodes et plate-forme de modélisation	35
4.2.6	Evolution des modèles de connaissances dans le temps.....	36
4.2.7	Outils pour l'extraction des relations conceptuelles	37
4.3	Ontologies et terminologies pour la recherche d'information	39

4.3.1	Historique	39
4.3.2	Ontologies et recherche d'informations générales.....	40
4.3.3	Ontologies pour naviguer dans des collections de documents spécialisés	41
4.3.4	Ontologies et identification statistique de communautés scientifiques	43
5	Références	45

1 RESUME

Domaine de recherche : Mes recherches se situent dans le domaine de l'ingénierie des connaissances, 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 la formalisation ou l'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, leurs méthodes de 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 pluri-disciplinaire, ancrée au coeur 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.

Ma 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 conceptuel 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) 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 inter-disciplinaires de manière à approfondir les spécificités des vecteurs de connaissances, comme les activités humaines, 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 donc étudié une gamme variée d'applications, des systèmes experts jusqu'à la recherche d'information en passant par les systèmes coopératifs. Il en ressort la nécessité 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.

Inter-disciplinarité : 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 14 publications dans des revues, 13 chapitres de livre et près de 50 communications à des conférences avec actes publiés, dont 20 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 inter-disciplinaire. 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 pratiquement une chaîne de traitements et d'analyse de textes adaptés et des aides à la modélisation, objectif du projet de plate-forme DAFOE4App auquel je participe.
- Parmi les connaissances recherchées dans les textes, je me suis intéressé 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 maintenance 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 d'autres types de recherche d'information que la recherche par requête, comme les analyses d'opinions ou les systèmes question-réponse, qui s'appuient de plus en plus sur ce type de modèle.

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, 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, Corps Logicistes, DynamO et GEONTO).

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), que j'envisage de poursuivre et de diversifier (contacs avec le DFKI en Allemagne) en vue de monter un projet international sur l'extraction de relations à partir de textes.

2 CURRICULUM VITAE DETAILLE

2.1 Données personnelles, emplois et titres universitaires

AUSSENAC-GILLES Nathalie
née le 8 février 1964 à Albi (81)
45 ans, mariée, 3 enfants

IRIT - UPS
118, route de Narbonne
31062 TOULOUSE Cedex 9
Téléphone + 33 (0)5 61 55 82 93
Télécopie + 33 (0)5 61 55 62 58
Courrier électronique : aussenac@irit.fr

<http://www.irit.fr/~Nathalie.Aussenac>

Emplois

Depuis 1991 Chargée de recherche au CNRS, section 07 (Informatique) à l'Institut de Recherches en Informatique de Toulouse (IRIT), UMR 5505 du CNRS (1e classe depuis 1995)

1989 – 1991 Contrat post-doctoral cofinancé par le CNRS et l'entreprise Matra-Marconi Space, laboratoire mixte ARAMIIHS (Toulouse) UMR 155 du CNRS. Méthodes et outils d'analyse de tâche et d'acquisition de connaissances expertes.

1986 – 1989 Docteur-Ingénieur titulaire d'une bourse du CNRS, laboratoire LSI de l'UPS.

Diplômes

- 2005 Habilitation à diriger des Recherches, Université Toulouse 3 intitulée « Méthodes ascendantes pour l'ingénierie des connaissances ». Décembre 2005. Jury :
- | | | |
|---------------------|---|------------------------|
| Catherine Garbay, | DR au CNRS, IMAG (Grenoble) | rapporteur |
| Joost Breuker, | PR, Université d'Amsterdam, SWL | rapporteur |
| Gilles Kassel, | PR, Université d'Amiens, LaRIA (Amiens) | rapporteur |
| Claude Chrisment, | PR, Université Paul Sabatier, IRIT (Toulouse) | président |
| Jean-Luc Soubie, | IR habilité, INRIA, IRIT, | directeur de recherche |
| Pierre Tchounikine, | PR, Université du Mans, LIUM | examinateur |
- 1989 Doctorat en Informatique, Université de Toulouse 3. “ Conception d'une méthodologie et d'un outil d'acquisition des connaissances expertes ”
- 1986 Diplôme d'Études Approfondies en Informatique, Université de Toulouse 3. Option algorithmique numérique.
- 1986 Ingénieur en Informatique de l'ENSEEIH, INPT Toulouse.

Membre d'associations professionnelles

Association de Recherche en Sciences Cognitives (ARCOg) depuis 1990
Association Française d'Intelligence Artificielle (AFIA) depuis 1991
Association pour le Traitement Automatique des Langues (ATALA) depuis 1999
International Association for Ontologies and its Applications (IAOA) depuis 2009

2.2 Participation à des projets

Présentation par type de contrat et par ordre chronologique décroissant.

2.2.1 Contrats avec des entreprises (11)

Projet CNES – R&T (2006-2007) : *Analyse de textes pour anticiper les problèmes liés à l'évolution dans le temps des connaissances et à la pérennité des données*
Responsable : B. Rothenburger (IC3). Financement : Etude de Recherche et Technologie du CNES.

Partenaires : laboratoires de linguistique CLLE-ERSS (Université Toulouse 2) et de Statistiques et de Probabilités (LSP) de l'IMT.

Objectifs : étudier les difficultés d'accès, au cours du temps, à des données scientifiques déposées sur le web, par d'autres communautés scientifiques.

Notre intervention : co-encadrement de la thèse de Nacim Chickhi ; méthodes statistiques exploitant les liens entre sites web et les liens de citations bibliographiques pour identifier automatiquement des communautés scientifiques ; application à des données d'astronomie déposées sur « observatoires virtuels » ; utilisation d'ontologie comme référence pour vérifier l'identification d'une communauté scientifique.

Projet MODE (2004 – 2008) – Financement FEDER : *Mise au point d'un système de diagnostic des calculateurs électroniques de voitures, basé sur une approche multi-modèle et multi-raisonnement.*

Partenaires : ACTIA (PME, Toulouse) et LAAS (laboratoire CNRS, Toulouse), au sein du laboratoire AUTODIAG¹.

Notre intervention : contribuer à définir et développer un système de recherche d'information, basé sur une ontologie et sur une annotation sémantique, qui simplifie la saisie des descriptions de pannes et oriente rapidement l'utilisateur vers une fiche de diagnostic adaptée lorsqu'elle existe, ou si non, vers un mode de résolution plus pertinent ; encadrement de la thèse CIFRE d'Axel Reymonet (2004-2008).

Projet VERRE (2002) - Contrat Saint-Gobain Recherche : *Méthode de construction d'ontologie à partir de textes techniques sur la fabrication de la fibre de verre.*

Objectifs : évaluer la faisabilité pour l'entreprise de construire une ontologie utilisée par un système de classement de documents (routage) pour la veille technologique.

Notre intervention : définir une méthode de construction d'ontologie à partir de textes pour s'appuyer sur des logiciels d'extraction de termes (Syntex) et de relations sémantiques (Yakwa et Caméléon) développés à l'IRIT et l'ERSS, ainsi que sur le logiciel de modélisation Terminae (du LIPN) ; construire un noyau d'ontologie.

Projet Caméléon (1996 – 1999) – contrat avec le CEN du CEA, Cadarache - *Evolution de la méthode REX : apports de systèmes de traitement du langage naturel et d'analyse de textes à la construction de modèles conceptuels dans la démarche REX de gestion des connaissances acquises par retour d'expérience.*

Partenaires : Centre d'Essais Nucléaires du CEA et Euriware (Paris). Encadrement de la thèse CIFRE de P. Séguéla (1997-2001) réalisée au CEN du CEA de Cadarache.

Projets Mougli et Hyperplan (1995 et 1997) - Contrats d'Etude et Recherche avec la DER de EDF-GDF et GIS « sciences de la cognition » du CNRS : *Modélisation de la tâche pour l'intégrer dans des systèmes de consultation de documents techniques.*

1997 : MOUGLIS, analyse du besoin en matière de consultation électronique de document, contribution à la modélisation de la tâche de différents types d'utilisateurs. Collaboration avec le groupe SOAD, projet financé par le GIS 'Sciences de la cognition' : définition d'une base de connaissances terminologiques dans le domaine du génie logiciel scientifique, contribution à la définition d'un outil de gestion de bases de connaissances terminologiques.

1995 : HYPERPLAN, hypertexte de consultation documentaire prenant en compte le contexte de réalisation de la tâche de l'utilisateur. Responsable scientifique du projet.

Projet SADE (1993) – contrat d'Etude et Recherche avec la DAP d'EDF-GDF : *Acquisition et modélisation de connaissances juridiques pour la gestion de prêts financiers*

Objectifs : définir un système à base de connaissances pour l'aide à la gestion de dossiers de prêts financiers. Innovations : utilisation d'un extracteur de termes, LEXTER, pour faciliter la modélisation du domaine ; comparaison des méthodes CommonKADS et MACAOII pour la modélisation conceptuelle.

Notre intervention : responsable scientifique du projet, encadrement d'un stagiaire.

¹ <http://www.irit.fr/-Laboratoire-AUTODIAG->

MACAOII (1992-1995) : *Modélisation d'une démarche coopérative d'acquisition et de modélisation des connaissances.*

Partenaires : C. Reynaud et F. Tort (LRI), I. Delouis et J.P. Krivine (EDF-DER).

Objectif : intégrer les approches de modélisation des connaissances des 3 partenaires.

Responsable du développement de l'environnement de modélisation.

Projet SAMIE (1990-1992) – contrat avec MMS (Matra Marconi Space France) Toulouse : *Système d'Aide à la Maintenance Informatique Externe*

Projet du laboratoire ARAMIIHS. Objectifs : évaluer la faisabilité et l'apport d'un système à base de connaissances pour l'aide à la maintenance informatique. Responsable du projet et réalisation de l'acquisition et de la modélisation des connaissances.

Projet SADIA4 (1993) – contrat avec MMS, Toulouse : *Système d'Aide au Diagnostic pour l'Intégration de la case à équipements d'Ariane 4*

Projet du laboratoire ARAMIIHS. Contribution à l'acquisition des connaissances à l'aide de MACAO et encadrement du projet côté chercheur.

ARAMIIHS (1988-1994) - UMR 115 du CNRS : *Action de Recherche et Application Matra-Irit sur les Interactions Homme-Système.*

Chercheur détaché participant aux activités en Intelligence Artificielle des ingénieurs de MMS (Matra Marconi Space).

2.2.2 Projets régionaux (4)

Projet OntoTextes « Ontologies et textes » - Financement BQR-UPS (2007) : *Modélisation de connaissances pour la gestion documentaire.*

Partenaires : 3 équipes de l'IRIT : SIG-EVI, SMAC et IC3.

Objectifs : fédérer les travaux des 3 équipes en matière de construction et de maintenance d'ontologies à partir de textes et de thésaurus, ainsi que de leur utilisation pour l'annotation de documents et la navigation au sein de collections documentaires. Notre contribution : gestion du projet et mise en place du projet ANR Dynamo.

Projet IndexWeb (2000 – 2002) - Projet Région Midi-Pyrénées : *Introduction d'analyses linguistiques et de connaissances ontologiques pour améliorer les résultats du logiciel IndexWeb d'indexation de pages Web.*

Partenaires : Synapse-Développement (PME qui a développé IndexWeb), laboratoire ERSS (Université Toulouse Le-Mirail), ARIST Midi-Pyrénées.

Notre intervention : constitution de deux terminologies pour les entreprises CLS et CEDOM à partir des textes de leurs sites web et de documents de communication. Démonstration de l'intérêt de cette étude pour une meilleure structuration du site, pour guider le choix de mots clés et améliorer le référencement des pages.

Projet DDE (1997-1998) - Projet Région Midi-Pyrénées : *Recueil et représentation informatique des terminologies des différents partenaires collaborant pour surveiller la circulation dans l'agglomération toulousaine.*

Partenaires : DDE Haute Garonne, Mairie de Toulouse et SEMVAT : Contrat avec la DDE géré par l'ERSS.

Notre contribution : fourniture et suivi de l'utilisation du logiciel Géditerm pour représenter les résultats d'analyse de textes dans une base de connaissances terminologiques ; évaluation ergonomique du logiciel.

Maintenance des SBC (1992-1994) - Contrat Région Midi-Pyrénées 930-0244. Co-responsable du projet avec J.-L. Soubie.

2.2.3 Projets nationaux

Projet FRANTIC – ANR-07-CORP- (2008-2010) : *Formalisation du thésaurus PACTOL pour l'indexation conceptuelle*

Responsable : B. Lequeux, MAE Nanterre.

Objectifs : Diffusion d'une version formalisée, sous forme d'ontologie, du thésaurus PACTOL pour l'indexation d'articles scientifiques en sciences humaines (archéologie) dans les archives de revue.org.

Notre contribution : choix de la représentation des connaissances (évaluation des standards OWL et SKOS), processus de transformation du thésaurus en ontologie.

Projet DynamO – ANR-07-TLOG-004 (2008-2010) : *Dynamic Ontologies*

Responsable : J. Thomas (ACTIA) - Partenaires : ACTIA, ARTAL , LaLIC (Univ. Paris 4), IRIT éq. IC3, SIG-EVI et SMAC, Préhistoire et Technologie (Paris X).

Objectifs : définir une méthode et des outils pour améliorer la recherche d'information et la satisfaction des utilisateurs en prenant en compte l'évolution des collections documentaires consultées. La recherche d'information s'appuie sur une ontologie, dont on assure la maintenance en fonction de l'évolution du corpus, des connaissances, de la terminologie du domaine, ou des besoins des utilisateurs.

Notre intervention : définition du processus de maintenance cohérente d'ontologie et d'annotations à partir de textes selon deux approches : à l'aide de traitement automatique des langues et à l'aide d'un système multi-agents ; définition d'une représentation de l'ontologie adaptée pour intégrer des termes.

Projet « Corpus Logicistes » – ANR-07-CORP-006 (2008-2010) : *accès sémantique au contenu de corpus logicistes en sciences humaines pour favoriser les échanges scientifiques*

Responsable : V. Roux (« préhistoire et technologies » , MAE Nanterre). Partenaires : laboratoires « préhistoire et technologies » de Paris X, Editions Epistèmes.

Objectifs du projet : (1) développer des méthodes et outils pour constituer des corpus dits "logicistes", composés de documents structurés en règles d'interprétation ; (2) constituer, sur ce modèle, des corpus en tracéologie ; (3) développer un outil d'annotation automatique basé sur une ontologie pour interroger les corpus logicistes essentiellement sur les règles d'interprétation ainsi que sur les données.

Notre intervention : construire l'ontologie, définir un outil d'annotation sémantique et proposer un système de recherche d'information dans les collections du site Arkeotek.

Projet GEONTO – ANR-07-MDCO-005 (2008 – 2010) : *Constitution, alignement, comparaison et exploitation d'ontologies géographiques hétérogènes*

Responsable : C. Reynaud, LRI - Partenaires : LRI eq. IASI/gemo (Univ. Paris Sud), IRIT eq. IC3 (Toulouse 3), LIUPPA eq. DESI (Univ. de Pau) et CoGIT (IGN).

Objectifs du projet : assurer l'interopérabilité de données géographiques hétérogènes à l'aide d'une ontologie résultant de deux approches complémentaires : (1) l'alignement d'ontologies construites à partir des schémas de bases de données géographiques hétérogènes ; (2) son enrichissement par l'analyse de textes complémentaires.

Notre intervention (lot 1) : construction d'ontologies par analyse automatique du texte de spécifications des BD et de documents géographiques, extraction des relations sémantiques.

Projet DAFOE4App – ANR-06-TLOG-010 (2007-2009) : *Differential and formal ontology editor*

Responsable : J. Charlet (INSERM) - Partenaires : INSERM UMR_S 872 équipe 20 (Paris 6/Paris 5), ENST/GET (Paris), IRIT (Toulouse 3), LIPN (Paris 13), LISI (Poitiers), Mondeca (Paris), Supélec (Saclay), UTC (Compiègne).

Objectif du projet : développer une méthode associée à une plateforme technique pour concevoir des ontologies, de la modélisation à la formalisation, en privilégiant l'analyse de textes par des outils de traitement automatique des langues.

Notre intervention : analyse des besoins, spécification du modèle de données pour la représentation d'éléments terminologiques associés à une ontologie, spécification des outils d'extraction et de représentation de relations sémantiques.

Projet DYNAMO (2004) – pré-projet du programme TCAN du CNRS : *Système multi-agents pour la construction et la maintenance d'ontologies à partir d'analyse de textes.*

Partenaire : équipe SMAC de l'IRIT. Notre intervention : Spécifications préliminaire d'un système multi-agents procédant à une analyse syntaxique partielle. Co-encadrement du Mastère de K. Ottens.

Projet ArkeoTek (2003-2005) – projet de programme « société de l'information » STI du CNRS : Ontologie pour un recherche sémantique d'information au sein de documents logicistes.

Partenaires : éditions Epistèmes et laboratoire d'archéologie « Préhistoire et techniques » de la MAE à Nanterre (V. Roux). Notre intervention : construire une ontologie à partir de documents structurés et développer un prototype d'annotation et de recherche documentaire utilisant cette ontologie.

Projet Th(IC)² (1999–2000) - financement DGLF : Constitution d'un thésaurus de l'ingénierie des connaissances à partir de l'analyse de textes à l'aide de logiciels d'extraction de terminologies et de relations sémantiques.

Projet mené au sein du groupe TIA, partenaires : ERSS (D. Bourigault), DIAM AP-HP (J. Charlet) et LIPN (S. Szulman). Notre intervention : constitution du corpus, construction d'une ontologie des outils de l'ingénierie des connaissances à l'aide de Terminae et Géditerm. Extraire des relations conceptuelles du corpus avec Caméléon. Définition d'un protocole de validation de termes via Internet.

Maintenance des SBC (1992-1994) - Contrat MEN 92-727 Gis « Sciences de la Cognition ».
Co-responsable du projet avec J.-L. Soubie. Partenaire : Univ. D'Amsterdam, dépat. SWI (J. Breuker)

2.3 Administration et animation de la recherche à l'IRIT

2.3.1 Direction de l'équipe IC3 de l'IRIT

Depuis septembre 2007, je suis responsable de l'équipe Ingénierie de la Connaissance, de la Cognition et de la Coopération (IC3) de l'IRIT (<http://www.irit.fr/-Equipe-IC3->).

Composition de l'équipe et fonctionnement

Les membres de l'équipe sont localisés sur 2 sites : l'université Paul Sabatier et l'université Toulouse le Mirail (UTM). L'équipe comporte 13 permanents (dont 11 ont une activité de recherche et 2 sont associés) :

- Nathalie Aussenac-Gilles (responsable équipe), Chargée de Recherche, CNRS
- Guy Camilleri, Maître de Conférences, UPS
- Corinne Chabaud, Maître de Conférences, UTM (permanent associé)
- Pierre-Jean Charrel, Professeur, UTM
- Catherine Comparot, Maître de Conférences, UTM
- Ollivier Haemmerlé, Professeur, UTM
- Nathalie Hernandez, Maître de Conférences, UTM
- Jean-Michel Inglebert, Maître de Conférences, UTM (permanent associé)
- Mouna Kamel, Maître de Conférences, Univ. Perpignan
- Bernard Pavard, Directeur de Recherche, CNRS
- Bernard Rothenburger, Ingénieur de Recherche, INRIA
- Jean-Luc Soubie, Ingénieur de Recherche, INRIA
- Pascale Zaraté, Maître de Conférences, INPT

L'équipe comporte 14 doctorants et contractuels (10 doctorants, 2 post-doctorants, 2 contractuels) et collabore de façon continue avec 2 chercheurs associés :

- Abdelkader Adla, Doctorant
- Hassan Ait-Haddou, PostDoc
- Tarek Ben Mena, Doctorant
- Hakim Bendjenna, Doctorant
- Nacim Fateh Chikhi, Doctorant

- Sandrine Darcy, Contractuelle
- Marion Laignelet, PostDoc
- Colin Lalouette, Doctorant
- Arnaud Martin, Doctorant
- Axel Reymonet, Chercheur associé
- Pascal Salembier, Chercheur associé
- Zied Sellami, Doctorant (co-encadrement SMAC)
- Noria Taghezout, Doctorant
- Anis Tissaoui, Doctorant
- Ismael Valente, M2R
- Philippe Viguié, Doctorant
- Jacques Virbel, Chercheur associé

L'équipe a accueilli courant 2009 un chercheur invité, T. Declerck du DFKI (Allemagne).

Afin d'entretenir une vie d'équipe et de favoriser les échanges d'information, et en continuité avec le fonctionnement mis en place par Jean-Luc Soubie, l'équipe se retrouve régulièrement :

- réunions d'information bi-hebdomadaires de toute l'équipe sur un des deux sites, et lettre d'information diffusée par email ;
- réunions régulières d'un bureau d'animation de l'équipe composé de B. Pavard, O. Haemmerlé, P. Zaraté et moi-même ;
- séminaire annuel faisant le point de travaux de chacun et en vue de redéfinir une problématique de recherche commune, tenant compte des évolutions de chacun ;
- réunions de travail scientifiques et régulières des différents sous-groupes.

Historique

- Depuis janvier 2007 : je coordonne le groupe « modélisation de connaissances et textes » de l'équipe IC3 (2 PR, 4 CR, 1 IR, 1CR, 4 thésards)
- La constitution actuelle de l'équipe date de janvier 2007 et résulte de l'intégration de 2 groupes courant 2006 et 2007, alors que le responsable était Jean-Luc Soubie
 - en janvier 2007, ont été intégrées dans l'équipe 2 professeurs et 3 maîtres de conférences d'équipe ISYCOM du laboratoire GRIMM de l'université Toulouse le Mirail.
 - Fin 2006, l'équipe GRIC de B. Pavard a rejoint l'équipe CSC pour former IC3.
- Auparavant, en janvier 2006, M. Kamel (MCF à l'Univ. de Perpignan) et B. Rothenburger ont rejoint l'équipe CSC
- De 2002 à fin 2006: coordination des travaux de Bernard Rothenburger (IR INRIA à l'IRIT) avec les miens.

Thématiques de recherche de l'équipe

L'équipe IC3 (Ingénierie de la Connaissance, de la cognition et de la coopération) a comme axe central l'ingénierie des modèles de connaissances et/pour des systèmes coopératifs. Ses travaux s'orientent vers la conception d'outils afin de pouvoir utiliser, acquérir des connaissances et assurer une représentation adéquate pour l'utilisateur. De plus ces outils doivent être capables de gérer la dynamique de la connaissance. La prise en compte du contexte de l'utilisateur est un premier élément clé de cette dynamique. De ce fait, l'équipe IC3 mène des études sur la coopération (homme-machine et homme-homme) afin d'appréhender un contexte multi-acteur, ce qui nous conduit au développement de systèmes coopératifs. L'étude du contexte d'utilisation de ces systèmes passe aussi par l'étude des processus cognitifs de l'utilisateur. L'équipe IC3 a développé un savoir faire interdisciplinaire faisant appel à l'ergonomie, la représentation des connaissances, l'aide à la décision. Tous ces travaux sont appliqués dans un premier temps au domaine de l'aide à la décision coopérative, à la gestion des risques, à la gestion des connaissances ainsi qu'à la recherche d'information.

Cette équipe regroupe trois composantes : ingénierie de la connaissance, ingénierie de la coopération, ingénierie de la cognition.

Un premier volet des recherches vise à déterminer la nature et la structure des modèles de connaissances pour des logiciels permettant d'accéder à des données ou à des informations textuelles. On étudie également l'ingénierie de ces modèles, les méthodes et logiciels assurant leur construction, en particulier à partir de textes, ainsi que leur dynamique et leur maintenance. Ces sont là les travaux du groupe « Ontologie et textes » que j'ai animé en 2006-2007. La plupart des modèles étudiés sont des ontologies, des ressources termino-ontologiques ou des modèles conceptuels représentés par des graphes conceptuels. Ces travaux s'étendent à l'utilisation et l'intégration des modèles dans des applications de recherche d'information (web sémantique ou collections de documents dans des domaines spécialisés), de gestion des connaissances (pérennisation de données scientifiques) ou de résolution de problème (diagnostic). En s'intéressant à l'analyse et à la caractérisation du contenu de documents pour leur annotation sémantique, ils interrogent l'articulation langue /connaissances /représentations conceptuelles. De plus, la diversité des connaissances et des acteurs impliqués dans l'élaboration et l'usage de ces modèles est prise en compte dans des approches centrées sur les points de vue.

Dans le deuxième volet, les recherches de l'équipe IC3 se préoccupent de l'ingénierie de la coopération homme/homme ou homme/machine dans des situations complexes - complexité d'un problème à résoudre, ou introduite par la relation qui s'établit entre les acteurs au sein d'un système, ou par le caractère critique de la situation elle-même, lorsque les procédures ne permettent pas d'y apporter des solutions, par le caractère distribué de leur réalisation et par la multiplicité des acteurs impliqués -.

Les systèmes coopératifs utilisent des modèles de connaissances de résolution de problème et des modèles de coopération tels que les définit l'ingénierie des connaissances. L'articulation de ces travaux avec les ontologies se dessine petit à petit : l'étude théorique des processus d'interaction homme-système a montré la nécessité de formuler une ontologie de l'interaction ; les ontologies sont vues comme des modèles de connaissances individuelles et permettant de confronter des groupes d'utilisateurs, ou partagées dans des systèmes d'aide à la décision collective et distribuée ; enfin, elles peuvent servir à annoter et partager des ressources communes dans le cadre d'un travail collaboratif.

Les choix de méthodologie scientifique font également l'unité d'IC3. Il s'agit d'abord d'une volonté forte d'aborder les problèmes sous l'angle des sciences cognitives et selon une approche pluridisciplinaire. Du point de vue théorique, le projet que nous avons élaboré dans le cadre du nouveau quadriennal du laboratoire (2011-2014) s'appuie sur les deux courants dont se reconnaissent les chercheurs de l'équipe (théorie de la complexité et dynamique des connaissances d'une part, théorie des représentations et cognitivisme d'autre part). Notre objectif est d'identifier la complémentarité de ces courants en fonction de la criticité et de la complexité des situations d'interaction (taille du collectif impliqué dans une activité, nature des tâches réalisées, caractère critique ou nominal des situations dans lesquelles une assistance informatique est prévue, etc.). Nous souhaitons étudier ensemble des terrains applicatifs nous permettant de situer la frontière entre les deux types d'approches et la pertinence de leur « cohabitation » pour définir des aides au travail mieux adaptées ?

Il s'agit aussi de situer les contributions théoriques et logicielles dans une démarche d'ingénierie, d'apporter des solutions formelles, méthodologiques ou technologiques novatrices à des problèmes spécifiques. Pour l'équipe IC3, il est important que nos recherches soient validées par des applications, la validation expérimentale faisant partie de notre démarche scientifique. Aussi, nous avons besoin de contribuer à des projets, et ceux-ci requièrent parfois de coûteux développements. Ces deux points nous différencient fondamentalement de recherches qui abordent des problèmes théoriques d'intelligence artificielle, validés par des preuves formelles.

L'ingénierie des connaissances au sein du thème 3 de l'IRIT

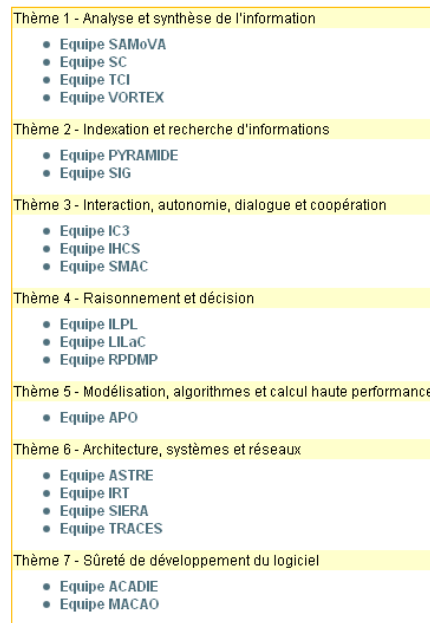


Figure 1 : organigramme de l'IRIT

L'équipe IC3 est rattachée au thème 3 de l'IRIT « Interaction, Autonomie, Dialogue et Coopération » (cf. l'organigramme présenté en figure 1). Les recherches au sein du thème 3 portent sur la **modélisation et la conception de systèmes** dans lesquels l'interaction, le dialogue et la coopération entre agents sont prépondérants, que ces agents soient de natures différentes (humain/artificiel) ou de même nature. Ces recherches sont abordées de manière pluridisciplinaire et s'organisent en trois domaines :

- **interactions homme-système** avec notamment : l'ingénierie des systèmes interactifs, une approche formelle, la multi-modalité et l'interaction dégradée ;
- **texte** : étude de l'architecture textuelle et construction d'ontologies à partir de textes, c'est dans ce domaine que se situent mes travaux ;
- **systèmes coopératifs** tel que : les systèmes multi-agents, le dialogue formel, des méthodologies de conception et la simulation de systèmes coopératifs.

Une caractéristique de ce thème est le souci de validation et d'évaluation en utilisant soit la méthode expérimentale, soit la simulation, soit la réalisation d'applications en collaboration avec le tissu économique. Ces recherches se traduisent aussi par des coopérations entre équipes du thème. En matière d'ingénierie des connaissances et surtout des ontologies, IC3 collabore avec SMAC dans le cadre du projet DynamO (thèse de K. Ottens soutenue en 2007, thèse de Z. Sellami débutée en sept. 2008). D'autres collaborations concernent les systèmes coopératifs (projet Dialogue commun avec DIAMANT-LIHS, méthodes de conception de systèmes coopératifs, etc.).

2.3.2 Collaborations scientifiques dans le laboratoire

Je collabore directement avec d'autres équipes de l'IRIT sur les 3 thématiques de recherche détaillées ci-dessous. D'autres collaborations ont lieu entre les équipes IC3, LiLaC et Diamant par la participation de G. Camilleri du groupe Dialogue entre 2005 et 2008, entre les équipes IC3 et Vortex par la participation de B. Pavard à des projets en réalité virtuelle, et entre les équipes IC3 et IHCS sur des questions d'ergonomie de l'interaction homme-système.

Ontologies pour la recherche (sémantique) d'informations

Equipe SIG-EVI (J. Mothe) : annotation sémantique dans des contextes dynamiques et spécialisés. Co-encadrement de la thèse de N. Hernandez (2002-2005) ; collaboration au sein du projet « ontologies et textes » financé par l'UPS ; collaboration au sein du projet Dynamo depuis janvier 2008.

Equipe SIG-RFI (M. Boughanem) : apport des ontologies à la recherche d'informations générales et à la contextualisation des réponses. Co-encadrement du DEA puis de la thèse de M. Baziz (2001-2005) ; participation commune au lot consacré à l'annotation sémantique de dossiers patients à l'aide d'une ontologie dans le projet IAPA.

Projet « OntoTextes », avec les équipes SMAC et SIG-EVI (2007) : modélisation des connaissances pour la gestion documentaire. Montage et responsabilité du projet, qui a fédéré les forces de l'IRIT en matière d'annotation et construction d'ontologies à partir de textes.

Equipe SMAC (M.-P. Gleizes et P. Glize) : utilisation d'un système multi-agents adaptatifs pour la construction et la maintenance d'ontologies à partir de textes. Co-encadrement de la thèse de K. Ottens (2004-2007) ; mise en forme du projet Dynamo soumis à l'ANR – TechLog en 2006 et 2007, accepté en 2007 (détails en 4.2). Co-encadrement de la thèse de Z. Sellami depuis sept. 2008 : approche multi-agents adaptatifs pour l'enrichissement d'une ontologie par des relations (hiérarchiques et non hiérarchiques) entre concepts identifiées par analyse de textes.

Traitement Automatique des Langues et modélisation sémantique

Collaboration avec IHCS-Diamant (M. Mojahid) : prise en compte de la mise en forme matérielle et de la structure des documents dans les patrons de recherche de relations sémantiques. Expérimentation dans le projet GEONTO, où les textes comportent des éléments de structuration explicite (XML) et de mise en forme matérielle.

Collaboration avec LiLAC (F. Bénomara) : ontologies pour la recherche d'expressions d'opinions dans des textes spécialisés, et pour définir des systèmes de question-réponse ciblés. Mise au point d'une ontologie d'opinions, identification de marqueurs en langue d'éléments d'opinion présents dans l'ontologie, association d'une ontologie du domaine à l'ontologie d'opinion. Définition d'un sujet de M2R commun, présentation d'un article commun à la conférence IC 2006. Co-encadrement d'une stagiaire M2R en 2009-2010.

Collaboration avec LiLAC (N. Asher, P. Muller, L. Vieu) : apport de l'analyse du discours et d'éléments pragmatiques au repérage d'éléments terminologiques et conceptuels dans les textes. Intégration des travaux sur l'extraction de relations temporelles de LiLAC et de nos travaux sur l'extraction de relations sémantiques au sein d'une plate-forme commune : travail en cours. Participation commune à l'axe « Lexique et ontologies » du LEA ILIKS.

Axe prioritaire « masse de données et calcul » de l'IRIT

Depuis septembre 2007, l'équipe IC3 a contribué à la réflexion relative à cet axe qui fait partie des quatre axes de recherche prioritaires dégagés par le conseil scientifique de l'IRIT pour les années à venir. Cet axe (animé par F. Sèdes et P. Amestoy) s'intéresse entre autres au partage de gros volumes d'information et de connaissances sur le web, et aux algorithmes, représentations, architectures, modèles requis pour les gérer et les exploiter. Dans ce cadre, nous avons proposé un volet consacré au Web Sémantique : analyse, modélisation, annotation et exploitation des contenus informationnels. Ce sujet pourrait fédérer et développer des travaux d'IC3 et ses collaborations avec d'autres équipes de l'IRIT en matière d'ingénierie des connaissances et traitement automatique de textes pour la recherche d'informations dans de gros volumes de données.

2.3.3 Animation scientifique au niveau du laboratoire

Activité passée

SISTEM : groupe toulousain du *PIRTEM* (Pôle Interdisciplinaire de Recherche sur les Technologies, le Travail, l'Emplois et les Modes de vie). Axe 1 : Impact de l'introduction de nouvelles technologies comme les systèmes experts au sein des organisations. De 1988 à 1993, animation du projet consacré à l'acquisition des connaissances.

PRESCOT : Pôle de Recherches en Sciences Cognitives de Toulouse de 1998 à 1996. Responsable d'un projet interdisciplinaire entre 1988 et 1994, portant sur la nature des méthodes de résolution de problème et sur les types de raisonnement en fonction des types de tâches (collaboration avec des psychologues cognitivistes). Membre de l'atelier 'Cognition partagée' animé par B. Pavard et J.L. Soubie entre 1992 et 1994

Activité en cours

Groupe « *Langage naturel* »² depuis 2004 : co-animation avec P. Saint Dizier et P. Muller. Identification des ressources et des besoins au sein du laboratoire. Co-organisation (avec D. Bourigault de l'ERSS) d'une série de séminaires communs avec le laboratoire de linguistique ERSS (devenu CLLE-ERSS) entre 2003 et 2005, co-organisation d'une série de 9 séminaires sur le TAL à l'IRIT tout au long de 2008³. Mise en place avec P. Muller du site web pour partager supports de cours, logiciels, supports de séminaires.

Plate forme RI à l'IRIT (2001- 2005) : PLEXIR intégrée au réseau national EXPRIM. Coordination et animation de 5 séminaires en 2001/2002 en vue d'amorcer des coopérations bilatérales. Collaboration avec les équipes SIG, LaLIC et SMAC de l'IRIT ainsi que l'opération « sémantique et corpus » du laboratoire de linguistique ERSS (UTM) (une vingtaine de chercheurs au total). Objectif : mettre en place une plate-forme de recherche d'information (plate-forme RI), intégrant des approches linguistiques et des ressources terminologiques et conceptuelles, dans le but d'améliorer la formulation des besoins en information, la qualité des réponses, la prise en compte des utilisateurs. L'équipe SIG a pris en charge la gestion de ressources matérielles pour le stockage de gros volumes documentaires et de logiciels de recherche d'information afin de mener des campagnes d'évaluation de niveau international, constituant la plate-forme régionale RFIEC. Un site web⁴ permet l'accès à ces logiciels et ressources.

2.4 Administration et animation de la recherche hors laboratoire

2.4.1 Animation de la communauté nationale

*Ingénierie des connaissances : GRACQ*⁵. Groupe de travail (de 1991 à 1996) puis groupe d'animation de la communauté française d'ingénierie des connaissances (environ 300 personnes), rattaché à l'AFIA⁶ et thème 7 du GDR I3⁷.

Depuis sa formation : co-animation du groupe.

Depuis 1997, membre du bureau du GRACQ : groupe de 12 personnes, présidé par J. Charlet et C. Reynaud jusqu'en 2006, par J. Charlet et moi-même depuis 2007, assurant l'animation de la communauté d'ingénierie des connaissances.

² <http://www.irit.fr/-Langage-naturel->

³ <http://www.irit.fr/Cycle-de-seminaire-IRIT-2008>

⁴ <http://www.irit.fr/RFIEC/>

⁵ <http://www.irit.fr/GRACQ/>

⁶ Association Française d'Intelligence Artificielle : <http://www.afia-france.org>

⁷ <http://www.irit.fr/GDR-I3/>

Depuis 1999, ce bureau assure le pilotage de la conférence IC.

Depuis 1995, gestion du site Web du GRACQ⁸, sur lequel la conférence IC est archivée, une bibliographie du domaine est proposée, des cours sont mis à disposition et de nombreux liens pointent vers d'autres sites relatifs au domaine (conférences, équipes de recherche, etc.).

Depuis 1998, gestion de la liste électronique info-ic@biomath.jussieu.fr (450 inscrits) avec J. Charlet et F.Trichet.

Terminologie et Intelligence Artificielle : TIA (de 1998 à 2008). Groupe de travail associé à l'AFIA et comité de pilotage de la conférence TIA. Membre du groupe depuis 1998, co-animatrice depuis 2002. Ce groupe a rassemblé une dizaine d'équipes : informaticiens de l'IA, linguistes et terminologues autour de l'intégration de leurs démarches pour exploiter les connaissances contenues dans des textes. La notion de bases de connaissances terminologiques est un des supports d'étude des problèmes liés à la représentation et au traitement de ces connaissances. Animation de réflexions sur l'adaptation des méthodes et outils en fonction des types de modèles à concevoir et de leurs usages, sur l'annotation sémantique, l'extraction de relations sémantiques.

ASSTICCOT⁹ « *Constitution de Terminologies à partir de corpus* », Action spécifique 34 du département STIC du CNRS (2002 - 2003). Montage et co-animation avec Anne Condamines (linguiste à CLLE-ERSS, Toulouse). Action rattachée au RTP 33 « Documents : création, indexation, navigation » animé par J.M. Salaün (ENSSIB, Lyon). Fonctionnement en groupe fermé, composé d'une trentaine de chercheurs d'horizons disciplinaires variés : informatique (recherche d'information, traitement automatique des langues, apprentissage automatique et ingénierie des connaissances) ; sciences du langage (terminologie, socio-linguistique et linguistique de corpus) ; sciences de l'information. Diffusion de la réflexion au sein d'ateliers associés à des conférences (CFD 2002 et plate-forme AFIA 2003). Rédaction d'une déclaration d'intention pour organiser un réseau d'excellence européen et de 2 articles scientifiques ainsi que d'un numéro spécial de la Revue I3.

GDR-I3 (*Information, Interaction, Intelligence*)¹⁰ du CNRS : membre du comité directeur depuis 2007. Responsable avec J. Charlet et C. Reynaud de l'animation du thème 7 « ingénierie des connaissances ». Membre du comité éditorial de la revue I3 depuis 2002.

2.4.2 Collaborations scientifiques nationales

Outre les activités d'animation, mes collaborations au niveau national correspondent à des partenariats dans le cadre de projets, à la participation à des jurys de thèse et à l'organisation de journées scientifiques.

CLLE-ERSS, unité CNRS de linguistique à l'Université Toulouse le Mirail

Collaborations avec A. Condamines dans divers projets (depuis 1990)

- 2007-2009 : réflexion fondamentale sur la variabilité de la notion de patron, sur les limites des représentations ontologiques et terminologiques pour rendre compte du fonctionnement de la langue ou même des connaissances ; étude sur les indices permettant de repérer les glissements de sens des termes spécialisés dans le temps ;
- avant 2007 : représentation des connaissances dans les bases de connaissances terminologiques, apports des approches linguistiques à la construction d'ontologies,

⁸ <http://www.irit.fr/GRACQ/>

⁹ <http://www.irit.fr/ASSTICCOT/>

¹⁰ GDR Information, Intelligence et Interaction : <http://www.irit.fr/GDRI3>

automatisation de l'extraction de relations sémantiques à partir de textes, traces linguistiques de l'évolution des connaissances et de la terminologie spécialisée.

Action spécification « Corpus et terminologie » (ASSTICOT) (2004 – 2006) : avec A. Condamines, co-animation et valorisation des résultats de ce groupe : rédaction de deux chapitres de livres, édition d'un numéro spécial de la Revue I3 sur la gestion et l'étude de l'évolution des ressources terminologiques et ontologiques.

Plusieurs séjours post-doctoraux d'étudiant de CLLE-ERSS au sein d'IC3

- *M.P. Jacques (2005-2006) :* évaluation de Caméléon II. Mise au point d'un jeu de patrons lexico-syntaxiques de relations sémantiques pour le français, et évaluation sur un jeu de 8 corpus. Publication de 4 articles communs.

- *C. Pernet (2009) :* étude des anaphores présentes dans des textes de biologie, propositions pour la résolution d'anaphores pour identifier des relations conceptuelles exprimées par plusieurs phrases.

- *M. Laignelet (2009-2010) :* étude de marqueurs de relations sémantiques exprimées à l'aide de la structure de textes et de langage naturel ; cas des relations spatiales et temporelles utilisées pour décrire des itinéraires (projet GEONTO)

SYNTEX (D. Bourigault) (2000-2008), extracteur de terme et analyseur syntaxique : utilisation et évaluation de Syntex dans des projets de construction d'ontologies à partir de textes.

Séminaire TAL commun ERSS-IRIT (2003 – 2005) : co-animation avec D. Bourigault, organisation d'une vingtaine de séminaires mensuels faisant intervenir un chercheur de chaque laboratoire sur une question de recherche commune ou proche.

LIPN, équipe Représentation des Connaissances et Langage Naturel (RCLN)

Méthode Terminae (1998-2005), S. Szulman et B. Biébow : définition d'un cadre méthodologique, spécification de fonctionnalités liées à l'intégration dans Terminae de données extraites de logiciels de TAL, évaluation de Terminae par son utilisation dans divers projets, rédaction d'articles, dont un chapitre de livre paru en 2008, qui fait la synthèse de 10 ans de recherches sur Terminae.

Animation conjointe du groupe TIA (2003 - 2006), avec S. Szulman

Projet DAFOE (depuis 2008), S. Szulman, S. Desprès, A. Nazarenko et L. Audibert : représentation des éléments lexicaux associés aux ontologies ; définition d'une plate-forme de modélisation d'ontologies à partir de textes ; extraction de relations sémantiques à partir de texte (confrontation de nos approches).

Préhistoire et techniques – MAE, Nanterre

Collaboration avec V. Roux et son équipe sur le thème de l'apport d'une ontologie à l'accès à des connaissances scientifiques au sein de collection documentaire au format logique prôné par le projet ARKEOTEK, appelé SCD.

Projet ARKEOTEK (2002-2005) programme « société de l'information » du CNRS. Développement d'un prototype d'annotation sémantique de corpus au format SCD et d'une ontologie de l'archéologie des techniques. Réflexion sur le statut d'un modèle de connaissances par rapport à des corpus logicistes.

Suite d'ARKEOTEK (2006-2007) : Nouveau prototype d'annotation sémantique, enrichissement de l'ontologie. Problématisation des besoins en évolution du corpus, des annotations et de l'ontologie au fil du temps.

Projet ANR-TechLog Dynamo (depuis 2008) : projet Dynamo. Processus d'annotation sémantique à l'aide d'une ontologie prenant en compte l'évolution des corpus, de la terminologie et des connaissances du domaine. ARKEOTEK est l'une des applications de ce projet.

Projet ANR-Corpus « Corpus Logicistes » (2008-2010), porteur : V. Roux. En quoi les textes au format SCD mettent-ils à l'épreuve les approches linguistiques habituelles d'analyse de corpus pour la modélisation des connaissances ?

LRI, université d'Orsay, Chantal Reynaud

Co-animation du GRACQ entre 1989 et 1992.

Co-animation du thème 7 du GDR I3 (depuis 2006) avec J. Charlet

Méthode ascendante de modélisation de connaissances : en s'appuyant sur la complémentarité entre modèles de tâche (thèse de N. Matta, IRIT) et modèles du domaine (thèse de F. Tort, LRI), définition d'une méthode et d'une plate-forme adaptée de MACAO.

Notion de rôle en modélisation des connaissances (1999-2003), plusieurs articles communs avec P. Tchounikine et F. Trichet (IRIN).

Projet ANR-Masse de données GEONTO (2008-2010) : alignement d'ontologies (LRI) construite automatiquement à partir de spécifications de bases de données (IRIT et LIUPPA) pour assurer une interrogation unique de bases de données cartographiques

AP-HP et INSERM, Jean Charlet

depuis 1995 : co-animation de la communauté française d'ingénierie des connaissances via le GRACQ, son site web¹¹, la conférence IC et la liste info-ic.

1995 et 2000 : co-édition d'ouvrages de synthèse, plusieurs articles communs

Depuis 2007 : projet ANR-TechLog DAFOE4App

ACACIA puis EDELWEIS, INRIA Sophia-Antipolis, équipe animée par Rose Dieng (jusqu'en 2008) et depuis 2008 par Olivier Corby

1989 : comparaison de MACAO et 3DKAT

depuis 1993 : collaboration au sein de TIA

depuis 2000 : jurys de thèse, montage de projets (eWok)

mars 2010 : séminaire commun de 2 jours sur « ontologies, ingénierie des connaissances et web sémantique »

LALIC, univ. Paris 4, Philippe Laublet.

1990-1992 : co-animation du GRACQ

Sept. 2007 : co-organisation de l'atelier « ontologies et textes » associé à la conférence TIA 2007.

Juin 2008 : co-organisation avec Y. Prié, A. Giboin de l'atelier « IC2.0 » associé à la conférence IC 2008.

Depuis 2008 : collaboration au sein du projet ANR-TechLog DynamO, co-encadrement de la thèse de Anis Tissaoui depuis nov. 2008.

LIUPPA, Univ. de Pau, Mauro Gaio et Christian Sallabery

Depuis 2008 : Projet ANR-Masse de données GEONTO : extraction de relations sémantiques à partir de textes structurés pour la construction d'ontologies, étude des relations temporelles, confrontation de nos outils et implémentations de la notion de patron. Co-encadrement du contrat post-doctoral de Marion Laignelet.

Autres collaborations

- Groupe DISCO du LAAS, L. Travé-Massuyès : depuis 2004, laboratoire mixte Autodiag
- MAE Nanterre, équipe Documentation (B. Lequeux) projet ANR-Corpus FRANTIQ.

Collaborations passées

- 1996-1999 : Institut de Recherches en Informatiques de Nantes (IRIN) : P. Tchounikine
- 1998 : Laboratoire d'Informatique et de Linguistique (LLI) à Villetaneuse : D. Bourigault

¹¹ <http://www.irit.fr/GRACQ/>

2.4.3 Animation au niveau international

Depuis 2000 : Membre du Comité de pilotage de la conférence European Conference on Knowledge Engineering and Knowledge Management EKAW

Depuis 2005 : Membre du comité éditorial des revues Applied Ontology et IJHCS

2.4.4 Collaborations scientifiques au niveau international

Collaborations récentes :

*Interdisciplinary Laboratory on Interactive Knowledge Systems (ILIKS)*¹² depuis 2005 : laboratoire commun Européen IRIT-LOA-Université de Trento (Italie), L. Vieu. Contribution au champ 5 « lexique, terminologies, ontologies et accès au contenu des textes ». Collaboration avec le LOA sur la complémentarité entre ontologies formelles et ontologies construites à partir de textes, entre étude formelle des relations sémantiques et leur expression en langue. Exposés aux séminaires annuels.

Nicole Tourigny (Université de Laval à Québec) depuis 2005, préparation de cours communs en ingénierie des connaissances (deux séjours de 3 semaines et 1 semaine de N. Tourigny à Toulouse en avril et octobre 2007).

Paul Buitelaar (DFKI, Germany) et *Philipp Cimiano* (AIFB, Univ. of Karlsruhe, Germany) depuis 2006 : TAL pour la construction et le peuplement d'ontologies à partir de textes ; séjour de P. Buitelaar à Toulouse et invitation à TALN 2007 ; invitation de P. Cimiano à l'atelier « Ontologies et Textes » de TIA 2007 ; participation à des ateliers organisé par P. Buitelaar : Ontolex 2007 et OLP 2008 ; contribution sur la méthode Terminae à un chapitre de livre coordonné par P. Buitelaar et P. Cimiano.

Jesualdo Fernando Breis (Departamento de Ingeniería de la Información y las Comunicaciones, Facultad de Informática, Universidad de Murcia, Espagne) (depuis 2002) : ontologies pour la recherche d'information et construction d'ontologies à partir de textes. Accueil de J. Fernando Breis pendant 3 mois ; 3 séjours à Murcia de 1 semaine pour des cours.

Thierry Declerck (DFKI, Allemagne) : échanges scientifiques au sujet de l'extraction de relations sémantiques ; montage en cours d'un réseau de compétence européen, accueil durant 1 mois en décembre 2009.

Collaborations passées

- 2000-2002 : A. Maedche et S. Staab (DFKI, Allemagne), co-organisation d'un workshop
- Depuis 1992 et 1993 : Dept. of social science informatics, Univ. of Amsterdam (NL), J. Breuker
- En 1992 : Laboratoire d'IA, Vrije Univ. of Brussels (F), L. Steels.

2.4.5 Séjours à l'étranger

Avril 2009, Sept. 2008, mai 2005 : Departamento de Ingeniería de la Información y las Comunicaciones, Facultad de Informática, Universidad de Murcia (Espagne), trois séjours d'une semaine pour collaboration sur les ontologies en biomédecine et pour donner des cours en M2R sur « traitement automatique des langues pour l'ingénierie des ontologies ».

Juillet 2008 : the Stanford Center for Biomedical Informatics Research¹³ (BMIR), université de Stanford (CA, USA), dir. Mark Musen. Séjour de 15 jours. Le BMIR développe des recherches en informatique médicale sur les applications à base de connaissances et leur mise en œuvre dans l'activité des médecins, chercheurs, professionnels de santé. Visite en vue de me former sur les ontologies en informatique médicale, d'envisager une collaboration et faire connaître les recherches menées dans IC3 en

¹² <http://www.loa-cnr.it/iliks/>

¹³ <http://bmir.stanford.edu/>

matière de construction d'ontologies à partir de textes, de présenter les développements réalisés au sein de IC3 (prototype d'annotation sémantique TexViz) à l'aide de l'éditeur d'ontologie Protégé développé au BMIR. Découverte du fonctionnement de la recherche américaine.

Avril 2007 : Laboratory of Applied Ontology (LOA) CNR-Université de Trento , Trento (Italie), Dir. N. Guarino, et laboratoire ILIKS¹⁴ (L. Vieu) : séjour de 15 jours pour collaborer avec L. Vieu : apport de l'étude des ontologies formelles à la construction d'ontologies à partir de textes ou de modèles construits à partir de thésaurus. Découverte du fonctionnement de la recherche en Italie.

2.5 Activité d'enseignement

2.5.1 Enseignements en 3^e cycle et écoles d'ingénieur

- Nov. 2009 : Université de Nice, filière "Knowledge and Information Systems" commune au master IFI (Informatique: Fondements et Ingénierie) et à la dernière année d'ingénieur de Polytech'Nice à l'université de Nice Sophia-Antipolis : "traitement du langage naturel et construction d'ontologies à partir de textes" (3h)
- Avril 2009 et Sept. 2008 : Universidad de Murcia (Espagne), departamento de Informática, Master 2 recherche, "De textos a ontologías : herramientas de procesamiento del lenguaje natural para el modelado del conocimiento », professeur invitée, 10h cours et 10h encadrement.
- Depuis 2007 : Université P. Sabatier, M2R MIT en Informatique et Télécommunication, responsable du module BDRI8 - ALRI (Analyses Linguistiques pour la Recherche d'Information), intervention sur « ontologies et web sémantique » (8h à 6h/an)
- Depuis 2006 : Université Toulouse 2, M2R ECIL en Ergonomie Cognitive et Informatique linguistique, intervention sur « ontologies et TAL » (6h/an)
- Depuis 2006 : Université P. Sabatier, M2R ICMST en Information Communication et Médiations Socio-techniques, intervention sur « ontologies et web sémantique » (6h/an)
- 2004-2007 : Université P. Sabatier, DEA en Intelligence Artificielle, responsable du module SEC (Sémantique et Extraction de Connaissances), intervention sur « ontologies, web sémantique et extraction de connaissances à partir de textes » (10h /an).
- 2001-2006 : Ecole Nationale Supérieure en Electronique, Electrotechnique, Informatique et Télécommunications (ENSEEIH, Toulouse). 3^e années Informatique. Module COT (Construction de terminologies). (4h/an).
- 2000-2006 : Université Technologique de Troyes, DEA RACOR (Réseaux Avancés de Connaissances et Organisations) Text Mining (3h/an) et DESS « Connaissances, Réseaux, Communautés » Ontologies (2h/an)
- 1999-2003 : Université P. Sabatier, DEA RCFR (Représentation des Connaissances et Formalisation du Raisonnement), « Ingénierie des connaissances » (6h/an) module AMAC.
- 1996/1997 : ENSTIMAC (Ecole Nationale Supérieure des Mines d'Albi-Carmaux), Introduction à l'Intelligence Artificielle (3 h) et Ingénierie des Connaissances (3 h).
- 1995-1997 : Université P. Sabatier, DEA d'Intelligence Artificielle (Logique, raisonnement, Calcul) : Séminaire sur l'acquisition et la modélisation des connaissances (2 h/an).
- 1991-1997 : Université de Toulouse le Mirail, Diplôme d'Etudes Supérieures Spécialisées (DESS) en sciences cognitives, cours « Acquisition et modélisation des connaissances » et « Intelligence Artificielle » (15 h à 20 h / an)

2.5.2 Ecoles d'été

- Ecole de printemps du programme TCAN du CNRS « *Langue, connaissances, information* », Batz-sur-Mer (F), 23 au 27 mai 2005 : « Ingénierie des connaissances : modélisation et ontologies ». 2h00.

¹⁴ <http://www.loa-cnr.it/iliks/>

- URFIST, formations sur les ontologies et le Web sémantique : Rennes, mai 2005, 6h ; Toulouse, février 2006, mai 2007, juin 2008, mars 2009 6h et Nice, mars 2006, 6h

2.5.3 Tutoriels à des conférences et séminaires

- RDC 2005 : « représentation des données et des connaissances ». Exposé invité sur « modélisation de connaissances à partir de textes ». Paris, 21 mars 2005.
- RFIA 2004 : « Construction d'ontologies à partir de textes », partie d'un tutoriel sur les « Ontologies et web sémantique » présenté avec Charlet J., Laublet P., Toulouse, janvier 2004.
- TALN 2003 : « Construction d'ontologies à partir de textes », avec D. Bourigault. Batz-sur-mer, juin 2003.
- BDA 2003 : « Construction d'ontologies à partir de textes », Lyon, oct. 2003.
- IRIT : « Le langage naturel et ses applications ». Avec P. Saint-Dizier, P. Muller. Toulouse, Juin 2003.
- « Systèmes Experts et leurs applications » 1991 et 1992 : “Acquisition et modélisation des connaissances” (1 journée) avec P. Laublet et J.P. Krivine.

2.5.4 Participation à des commissions de spécialistes

- 2007-2008 : membre de la commission de spécialistes de la 27^e section du département Mathématiques et Informatique de l'université Toulouse le Mirail Toulouse 2.
- 1996-1997 : Membre de la commission de spécialistes de la 27^{ème} section au sein de l'UFR Mathématique, Informatique et Gestion de l'université P. Sabatier Toulouse 3.

2.6 Activité d'encadrement

2.6.1 Direction de thèses terminées

- 2004 – 2008 Axel Reymonet : directrice de thèse et encadrement (100%), « Modélisation de connaissances à partir de textes pour une Recherche d'Information Sémantique ». Contrat Cifre avec l'entreprise ACTIA (Toulouse). Université Toulouse 3, spécialité Informatique, école doctorale MITT, 23 septembre 2008.
- 2004 – 2007 Kévin Ottens : co-encadrement (50%) avec M.-P. Gleizes (IRIT-SMAC, directrice de thèse), « Un système multi-agent adaptatif pour la construction d'ontologies à partir de textes ». Université Toulouse 3, spécialité Informatique, école doctorale EDIT, 2 octobre 2007.
- 2002 – 2005 Mustapha Baziz : co-encadrement (50%) avec M. Boughanem (IRIT-SIG, directeur de thèse), « Apport des ontologies à la recherche d'information à l'aide du moteur Mercure ». Université Toulouse 3, spécialité Informatique, école doctorale EDIT. 14 décembre 2005.
- 2002 – 2005 Nathalie Hernandez : co-encadrement (30%) avec J. Mothe (IRIT-SIG, directrice de thèse), « Ontologies pour l'aide à une activité de veille ou d'exploration d'un domaine ». Université Toulouse 3, spécialité Informatique, école doctorale EDIT. 6 décembre 2005
- 1996 – 2001 Patrick Séguéla : Encadrement à 90%. Thèse Cifre réalisée à l'IRIT et au CEA, CEN de Cadarache. « Extension du logiciel REX par des outils et techniques d'analyse de texte et de traitement automatique du langage naturel. Développement du système Caméléon ». Université Toulouse 3, école doctorale Informatique. Directeur de thèse : Mario Borillo, mars 2001.

1991 – 1995 Nada Matta : Encadrement à 100%, « Représentation des connaissances pour l’acquisition des connaissances ». Université Toulouse 3, école doctorale Informatique. Directeur de thèse : Mario Borillo, octobre 1995.

1991– 1994 David Macchion : co-encadrement, « Intégration de plusieurs modes de raisonnement pour le diagnostic ». Université Toulouse 3, Ecole Doctorale Informatique, nov. 1994.

2.6.2 Direction de thèses en cours

depuis nov. 2008, Anis Tissaoui : directrice de thèse et encadrement (40% avec P. Laublet, Lalic-Paris4 et N. Hernandez, IRIT-IC3). « Annotation sémantique de corpus dynamiques ». Université Toulouse 3, école doctorale MITT.

depuis sept. 2008, Zied Sellami : co-directrice de thèse (avec M.-P. Gleizes, IRIT-SMAC) et encadrement (30%). « Système multi-agents pour l’évolution d’ontologies : traitement des relations conceptuelles ». Université Toulouse 3, école doctorale Informatique.

depuis sept. 2006, Nacim Chikhi : directrice de thèse et encadrement (50% avec B. Rothenburger, IRIT-IC3). « Ontologies et technologies du web sémantique pour l’accès aux données par des communautés scientifiques, cas des observatoires virtuels en astronomie ». Université Toulouse 3, école doctorale Informatique.

depuis déc. 2004, Aurélie Picton : co-encadrement (20 % avec A. Condamines, directrice de la thèse, CLLE-ERSS), « Approche linguistique et terminologique pour le repérage de l’évolution des connaissances spécialisée ». Université Toulouse 2, spécialité Sciences du langage. Soutenance prévue juin 2009.

2.6.3 Participation à des jurys

Jurys d’HDR

2009 : Juliette Dibie Barthélémy (AgroParisTech, univ. Paris Dauphine) - Rapporteur

2008 : Fabien Gandon (Edelweis, INRIA Sophia Antipolis) - Rapporteur

2004 : Nada Matta (TechCico,UTT – Troyes)

Jurys de thèses

2009 : Julien Lafalquière (UTT – Troyes)

2009 : Aurélie Picton (thèse en linguistique, UTM – Toulouse)

2009 : Sidonie Christophe (COGIT – Paris)

2008 : Dinh Quoc Truong (UTM – Toulouse)

2008 : Axel Reymonet (IRIT – Toulouse) – directrice de thèse

2008 : Bruno Richard (LIUM – Le Mans)

2007 : Kévin Ottens (IRIT – Toulouse) -- co-directrice de thèse avec M.P. Gleizes

2007 : Sabine Bruaux (LARIA – Amiens) -- Rapporteur

2007 : Florence Amardeilh (LaliCC – Paris X Nanterre) -- Rapporteur

2007 : Audrey Baneyx (INSERM – Paris 8) – Rapporteur

2006 : Khaled Khelif (INRIA - Sophia Antipolis)

2005 : Nathalie Hernandez (IRIT – Toulouse)

2005 : Mustapha Baziz (IRIT - Toulouse)

2002 : Joanna Golebiowska (INRIA – Sophia Antipolis)

2001 : Patrick Séguéla (IRIT – UPS) – directrice de thèse

2001 : Vincent Pautret (ENSSAT/IRISA – Rennes)

1999 : Jérôme Nobecourt (LIPN - Paris 13)

1999 : Florence Sellini (ISCMCM - Centrale Paris)

1998 : Francky Trichet (IRIN - Nantes)

1997 : Zoltan Isténès (IRIN – Nantes)

1996 : Françoise Tort (LRI – Orsay Paris Sud)

1995 : Nada Matta (IRIT – Toulouse 3)

1992 : Catherine Gréboval (UTC – Compiègnes)

2.6.4 Encadrement d'étudiants et stagiaires

- M2R IT: 2004 K.Ottens ,
- DEA IIL et IA : 2002, M. Baziz , 1998 : S. Simon, 1996 : P. Séguéla, 1994 N. Matta, 1991 : E. Courbon, G. Testemale, Th. Vidal. DEA RACOR (UTT) : 2001 : F. Amardeilh.
- DESS de terminologie 2002 A. Busnel, 2001 E. Cannesson,
- DESS de psychologie sociale et de psychologie du travail (Univ. Toulouse Le Mirail) : 9 étudiants
- DESS de sciences cognitives (Université de Toulouse Le Mirail), 2 étudiants
- ingénieurs 3A ENSEEIHT (Ecole Nationale Supérieure d'Electronique, Electrotechnique, Informatique et Hydraulique de Toulouse) : 12 étudiants entre 1991 et 1998.
- Ingénieurs CNAM (Centre National des Arts et Métier) de Paris et de Toulouse : 5 stagiaires ingénieurs (1988, 1989, 1990, 1998, 1999) : logiciels MACAO, MACAOII, SADE, GEDITERM et CONSULTERM.
- depuis 2002, Encadrement de 14 étudiants en maîtrise M1 d'IUP ISI et 5 étudiants en licence L3, développement et maintenance des nouvelles versions de Caméléon et du prototype Arkeotek.
- ingénieurs contractuels (sans préparation de diplôme) : au total 4,4 ans de travail, 9 personnes.

2.7 Actions de diffusion de l'information scientifique et technique

2.7.1 Participation à des comités de programme

Reuves, membre du comité éditorial

depuis 2004 (création) : Applied Ontology (Ed. : M. Musen, N. Guarino).

depuis 2007 : International Journal of Human-Computer Studies (IJHCS) (Ed. : E. Motta).

depuis 2001 (création) : Revue I3 (Ed. M. Boughanem, S. Benferhat et G. Mélançon).

Relectures régulières pour des revues internationales / numéros spéciaux

2009 et 2008 IEEE Transactions on Knowledge and Data Engineering (TKDE)

2009 Data and Knowledge Engineering (DKE)

2008 International Journal of Semantic Web and Information Systems (IJSWIS)

IEEE Intelligent Systems: "AI & Cultural Heritage".

2007 et 2006 International Journal of Data Semantics (IJoDS): "best 2006 conf papers"

2007 Journal of Universal Computer Science (JUCS): "ontologies and their applications"

2004 et 2008 Terminology : "Application-driven Terminology engineering"

« Pattern based approaches for relation extraction »

2001 Knowledge and Information Systems (KAIS): an international Journal

International Journal of Operational Research (IJOR)

Demandes ponctuelles par des revues nationales / numéros spéciaux

TSI (Techniques et Sciences Informatiques) : 2006 « Document Numérique » ; 2008 « Web Sémantique », 4 relectures ponctuelles

Revue des Nouvelles Technologies de l'Information RNTI, « fouille de données d'opinions » 2009

« Fouille de données complexes » 2004, 2006, 2008

Revue d'Intelligence Artificielle (RIA) (1999, 2002, 2003)

Document numérique : « Création et Gestion Coopératives de Documents Numériques d'Information et de Communication » (2002 et 2004)

Documents et Connaissances (automne 1999)

Membre du comité de pilotage de conférences

depuis 1997 Ingénierie des connaissances
depuis 1999 Terminologie et IA
depuis 2000 European Conference on Knowledge Acquisition and Management EKAW
2002 et 2004 Reconnaissance des formes et Intelligence Artificielle RFIA

Présidences et co-présidences de comités de programmes

2008 Atelier « IC 2.0 : Vers une ingénierie "sociale" des connaissances »¹⁵, IC 2008, organisé avec A. Giboin, P. Laublet, A. Passant, Y. Prié. Nancy (F), Juin 2008.
2007 Atelier "Ontologies et textes" du GDR-I3, TIA 2007, Nice, avec P. Laublet (LaLIC).
2006 - Semaine de la Connaissance 2006 (SDC 2006)¹⁶, juin 2006, Nantes (F) regroupant 18 manifestations scientifiques. Présidente du comité de pilotage..
- Atelier international « Indexation et Connaissances en Sciences Humaines »¹⁷ à la Semaine de la Connaissance 2006, juin 2006, Nantes (F), avec S. Calabretto (LISI).
2002 Workshop OLT 2002 associé à ECAI 2002¹⁸ : NLP and ML for Ontology Engineering
2000 Workshop associé à EKAW 2000 : Ontologies and Texts
1993 European Knowledge Acquisition Workshop (EKAW)
1991 Journées d'Acquisition des Connaissances (JAC)

Membre du comité de programme de conférences internationales

2009 EACL 2009, natural language processing track
ECML-PKDD¹⁹ NLP track
depuis 2006 International Conference on Knowledge Science, Knowledge Engineering, and Knowledge Management (KSEM)
2006, 2007 Intern. Conf. on Artificial Intelligence: Methodology, Systems, Applications (AIMSA)
2006 ECAI European Conference on Artificial Intelligence
2004, 2005 Knowledge Discovery and Ontologies (KDO)
Formal Ontologies and Information Systems (FOIS)
TAsk MOdels and DIAGrams for user interface design (TAMODIA)
Conférence Internationale sur le Fouille de Textes (CIFT)
2001 Conférence Internationale sur les Technologies de l'Entreprise (CITE), 2001 et 2003
Conférence Internationale sur le Document Electronique (CIDE)
1997 International Conference on Data and Expert Systems Applications (DEXA)
1996 Japanese Knowledge Acquisition Workshop (JKAW)
1994 à 1996 Knowledge Acquisition Workshop (KAW)
depuis 1994 European Knowledge Acquisition Workshop (EKAW)

Membre du comité de programme de conférences nationales

2007 et 2008 Journées Francophones sur les Ontologies (JFO)
Traitement Automatique du Langage naturel (TALN)
2007 Colloque ARCo, association pour la Recherche Cognitive
2004 Informatique des Organisations et Systèmes d'Information (INFORSID)
2001, 2008, 2009, 2010 Extraction et gestion des Connaissances (EGC)
2000 à 2004 Reconnaissance des formes et Intelligence Artificielle (RFIA) (biannuelle)
depuis 1999 Terminologie et Intelligence Artificielle (TIA) (biannuelle)
depuis 1997 Journées Francophones d'Ingénierie des Connaissances (IC) (anciennement JAC)
de 1991 à 1996 : Journées d'Acquisition des Connaissances (JAC).

¹⁵ <http://apassant.net/home/2008/05/ic/>

¹⁶ <http://www.sdc2006.org/>

¹⁷ <http://www.sdc2006.org/cdrom/ICSH2006-cd.html>

¹⁸ <http://www.inria.fr/acacia/OLT2002>

¹⁹ <http://www.ecmlpkdd2009.net/>

Membre du comité de programme de workshops internationaux et nationaux

- 2010 MSW 2010²⁰ associé à WWW2010 (Raleigh, USA) Multilingual Semantic Web
 2009 RISE 2009 associé à INFORSID: Atelier Recherche d'Information SEMantique RISE
 IWOD 2009²¹ : Ontology Dynamics: Ontology Evolution in Practice
 2008 OntoLex 2008²² associé à LREC: The Lexicon/Ontology Interface
 OLP 2008²³ associé à ECAI (Patras, Grèce) : Ontology Learning and Population
 FODOP 2008 Atelier FOuille des Données d'OPinions associé à INFORSID
 2006 OLP 2006 associé à COLING (Sidney, Aust.) : Ontology Learning and Population
 2005 à 2007 Atelier Défi Fouille de textes (DEFT)
 2005 KDO 2005²⁴ associé à ECML/PKDD (Porto,P) Knowledge Discovery and Ontologies
 2004 TAMODIA 2004 : Intern. Ws on TAsk MOdels & DIAGrams for user interface design
 OLP 2004²⁵ associé à ECAI (Valencia, E.) : Ontology Learning and Population
 KDO-2004²⁶ associé à ECML/PKDD (Pisa, I) Knowledge Discovery and Ontologies
 2003 IEBO 2003 associé à EUROLAN : Information Extraction for Building Ontologies
 EON 2003 associé à KCAP (USA): Evaluation of Ontology Engineering Tools
 2002 EON 2002 associé à EKAW 2002 : Evaluation of Ontology Engineering tools
 NLP and ML for Ontology Engineering, associé à ECAI 2002
 OntoLex 2002 à LREC: From Text to Knowledge: The Lexicon/Ontology Interface
 2000 Knowledge Management, Theory and Applications associé à PKDD2000
 Ontologies and Texts associé à EKAW'00
 Application of Ontologies and Problem-Solving Methods associé à ECAI'00
 1999 Applications of Ontologies and problem-solving methods associé à IJCAI99²⁷
 Ontological Engineering on the Global Information Infrastructure associé à EKAW
 1998 Application of Ontologies and Problem-Solving Methods associé à ECAI 98

Membre du comité de programme de journées scientifiques

- janv. 2009 XVIe rencontres de Rochebrune 2009 : « Ontologie et dynamique des systèmes complexes, perspectives interdisciplinaires »
 mars 2004 Journée ATALA, AGENTAL « Agents et traitement automatique des langues », Paris
 janv. 2004 Journées « Terminologie, ontologie et représentation des connaissances », Lyon.
 Oct. 2003 Journée « document et connaissances » du GDR-I3.

2.7.2 Organisation de conférences et journées scientifiques

Présidence et co-présidence de comité d'organisation

- 2009 : Conférence TIA 2009 à l'IRIT - Toulouse, nov. 2009 ; avec A. Condamines (CLLE-ERSS).
 2008 : Workshop " IC 2.0. Vers une ingénierie "sociale" des connaissances : dans quelle mesure les usages du Web 2.0 font-ils évoluer les pratiques d'IC ? "28, IC 2008, Nancy ; avec P. Laublet (LaLIC), A. Passant (EDF), A. Giboin (INRIA) et Y. Prié (LISI).
 2007 : Workshop " Ontologies et textes"29, TIA 2007, Sophia-Antipolis ; avec P. Laublet (LaLIC).
 2006 : Atelier "Indexation et Connaissances en Sciences Humaines", Semaine de la Connaissance30, Nantes (F) ; avec S. Calabretto (LISI).

²⁰ <http://msw.deri.ie/>

²¹ <http://www.ontologydynamics.org/od/index.php/iwod/iwod2009>

²² <http://www.lrec-conf.org/lrec2008/>

²³ http://olp.dfki.de/olp2/olp2_cfp.htm

²⁴ <https://webhosting.vse.cz/svatek/KDO05/>

²⁵ <http://olp.dfki.de/ecai04/cfp.htm>

²⁶ <http://olp.dfki.de/pkdd04/cfp.htm>

²⁷ <http://www.swi.psy.uva.nl/usr/richard/workshops/ijcai99/home.html>

²⁸ <http://apassant.net/home/2008/05/ic/>

²⁹ <http://www-sop.inria.fr/acacia/tia2007/atelier.html>

³⁰ <http://www.sdc2006.org/>

2002 : Workshop “Natural Language Processing & Machine Learning for Ontology Engineering”³¹, ECAI 2002, Lyon ; avec A. Maedche (FZI, G) et S. Staab (DFKI, G).

2000 : Workshop “ Ontologies and Texts ”³², EKAW 2000, Juan les Pins ; avec B. Biébow et S. Szulman (LIPN), initiative du groupe TIA.

2000 : Conférence IC’2000³³ (Journées francophones d’Ingénierie des Connaissances), Toulouse.

Membre de comités d’organisation

- *TALN 2007*, Toulouse, juin 2007.
- *Workshop SEMWEBKR2002* (Formal Ontology, Knowledge Representation and Intelligent Systems for the World Wide Web) associé à KR 2002, Toulouse, mai 2002.
- *ACL 2001*, Toulouse, juillet 2001
- *CIDE 2001* (Conférence Internationale sur le Document Electronique). Toulouse, oct. 2001.

³¹ <http://www.inria.fr/acacia/OLT20002>

³² <http://www.irit.fr/wsontologies2000>

³³ <http://www.irit.fr/IC2000/>

3 LE DOMAINE DE RECHERCHE

Mes recherches se situent dans le domaine de l'ingénierie des connaissances, champ de l'informatique qui s'intéresse à la mise au point de logiciels s'appuyant sur des connaissances pour assister un utilisateur dans sa tâche. L'ingénierie des connaissances (IC) définit des méthodes, des techniques et des outils pour construire et utiliser des modèles conceptuels de domaines spécialisés et de tâches en amont du développement de systèmes opérationnels (Charlet, 2002).

3.1 L'ingénierie des connaissances

3.1.1 L'ingénierie des connaissances en informatique

Quelle que soit la technologie informatique visée (systèmes experts, gestion documentaire, support au travail coopératif, etc.), la réalisation d'un système dont on attend un comportement 'intelligent' vis à vis de ses utilisateurs nécessite non seulement une analyse fine de leur besoin, mais aussi la prise en compte des savoir-faire et des pratiques du domaine concerné. Ainsi, un large éventail d'applications de l'informatique fait appel à l'intelligence artificielle (IA) et concerne l'IC. Depuis 10 ans, le champ de l'IC s'est élargi à toute application manipulant des connaissances, même si le système informatique final ne les rend pas opératoires. Le comportement intelligent du système est garanti dès qu'il apporte à son utilisateur la bonne connaissance au bon moment, qu'elle ait été trouvée après un raisonnement formel ou par accès à des informations documentaires par exemple. L'enjeu est de parvenir à une bonne analyse conceptuelle qui assure une meilleure spécification des connaissances puis en facilite la représentation formelle.

De manière complémentaire au génie logiciel, l'ingénierie des connaissances propose des méthodes pour définir le comportement attendu du système à concevoir, les buts qu'il doit atteindre et les connaissances qui lui sont nécessaires. Elle fournit également des langages de modélisation et d'opérationnalisation pour représenter ces modèles indépendamment du langage de programmation choisi, à un niveau d'abstraction qui favorise l'intelligibilité plus que la performance ou la calculabilité. Ces modèles et langages servent aujourd'hui à définir de manière structurée des métadonnées pour annoter des documents (annotation sémantique) et exploiter la complémentarité entre les connaissances déjà formulées dans des documents et celles représentées dans des modèles.

3.1.2 Un domaine de recherche pluridisciplinaire

En tant que discipline prenant en compte l'introduction d'outils informatiques dans des situations de travail, au sein d'organisations humaines, le plus souvent professionnelles, l'IC a besoin de collaborer avec de nombreuses disciplines comme l'ergonomie, les sciences de l'organisation, la psychologie et la sociologie, mais aussi la linguistique et la philosophie. Les résultats établis par l'IC, même s'ils ne sont pas complètement transférables à ces domaines, ouvrent des perspectives de collaboration. De manière symétrique, l'IC se nourrit de résultats de l'intelligence artificielle et de l'informatique en matière de technologies, méthodes, représentation des connaissances, algorithmes. Elle adapte, fait évoluer ou évalue l'intérêt de ces résultats, et contribue ainsi à leur diffusion et à leur ajustement opérationnel. L'activité de recherche en ingénierie des connaissances tient donc une place relativement indépendante par rapport à l'intelligence artificielle, et se caractérise par son besoin vital de collaborations interdisciplinaires.

3.1.3 L'ingénierie des connaissances en France

Domaine actif en France depuis 1988, l'IC a acquis plus de reconnaissance lorsqu'une première vague de travaux sont parvenus à maturité vers 2000. Son dynamisme s'est d'abord traduit par la mise en place d'un groupe de travail, le GRACQ, rattaché à l'AFIA³⁴ et au GDR-I3³⁵, qui s'est réuni tous les 2 mois de 1990 à 1994, puis 2 fois par an depuis 1995. Une démarche parallèle a

³⁴ Association Française d'Intelligence Artificielle (AFIA) : <http://www.afia-france.org/>

³⁵ <http://www.irit.fr/GDR-I3>

donné naissance à une conférence annuelle, les Journées d'Acquisition des Connaissances en 1991, devenues les Journées francophones d'Ingénierie des Connaissances en 1997. Entre 2000 et 2005, l'activité scientifique du domaine s'est traduite par la publication d'ouvrages de synthèse ; l'animation de groupes de travail (Actions Spécifiques), rattachés aux réseaux pluridisciplinaires du département STIC du CNRS ; l'implication de la conférence IC dans la plate-forme de conférences de l'AFIA ; enfin, l'intérêt manifesté par d'autres communautés scientifiques, comme celles du document, de la gestion des connaissances et de la recherche d'information. Après 2005, l'afflux d'informations et l'importance d'internet dans les questions de modélisation et gestion des connaissances a conduit à une rencontre entre des approches cognitives et qualitatives, classiques en IC, et des approches quantitatives développées pour fouiller textes et données, rechercher des informations, etc. Cette convergence s'est traduite par exemple par la naissance de la conférence EGC (Extraction et Gestion des Connaissances), mais aussi par la publication d'un nombre croissant d'articles d'IC, y compris dans des conférences générales d'IA comme RFIA. En 2006, la « semaine de la connaissance », organisée à l'initiative du comité de pilotage de la conférence IC, a regroupé 3 conférences et 8 ateliers pour favoriser les échanges avec d'autres disciplines concernées par les connaissances, leur modélisation et leur utilisation dans des systèmes informatiques.

3.2 Problématiques actuelles de l'ingénierie des connaissances

L'ingénierie des connaissances est un domaine en renouvellement constant. En effet, autour d'une problématique générique stable, celle de l'ingénierie de systèmes faisant appel à des connaissances, visant à les gérer, les diffuser, les pérenniser, l'IC doit s'adapter régulièrement à des technologies nouvelles à utiliser ainsi qu'à de différents objectifs d'application. Cela rend sans doute difficile la lecture des frontières du domaine au sein de l'informatique. L'apport de la démarche et de la réflexion de l'IC est alors perçu avec plus ou moins d'intérêt suivant les périodes et les applications. Quatre thèmes ressortent des travaux actuels en IC :

- La mémoire d'entreprise
- La gestion des connaissances sur support électronique
- Le web sémantique
- L'acquisition de connaissances à partir de textes

Nous revenons sur les deux derniers car ils concernent nos travaux. La multiplication des recherches sur l'acquisition des connaissances à partir de textes s'explique par le besoin d'accéder aux données sur le web, toujours plus nombreuses (en particulier des textes) via des techniques sémantiques. D'une certaine manière, elle découle de l'importance croissante du Web Sémantique

3.2.1 Le Web Sémantique

Le Web Sémantique est une proposition lancée par Tim Berners Lee et soutenue par le W3C, qui vise à « doter le web des éléments nécessaires pour que les machines deviennent aussi pertinentes que les humains pour retrouver et combiner des informations présentes sur les web ». Cette ambition semble viser des tâches de recherche complexes, faisant appel à plusieurs sources par exemple, ou à des contenus difficiles à caractériser. La solution envisagée par le W3C avec l'appui de la communauté scientifique, passe par l'annotation systématique des données du web à l'aide d'éléments de connaissance d'un format standard et d'identifiants uniques. Elle passe aussi par la définition de listes et de modèles conceptuels contenant ces connaissances et les mots-clés adaptés. Elle suppose enfin de définir des applications réparties, sous forme de services web, capables de manipuler ces standards et d'accéder à des données hétérogènes issues de sources diverses. Il s'agit là d'un choix très technologique, lié à l'état du web au moment où cette solution a été formulée, vers 1999. Il reflète une analyse du problème en termes de représentation de connaissances sous une forme exploitable par la machine, les technologies de l'IA pouvant apporter une réponse efficace.

L'IC à l'heure du Web Sémantique

L'IC est ainsi une des disciplines directement concernées par les technologies du Web Sémantique : elle contribue à définir des formats de représentation, comme RDF et RDFs, les types

de modèles pertinents pour l'accès à l'information et la production de raisonnements, dont les ontologies, et enfin les conditions de leur utilisation. En 2006, trois caractéristiques ressortaient de ces recherches :

- la prédominance d'une vue informatique, au détriment d'une analyse ergonomique et sociale des pratiques et usages des communautés qui se forment autour du web ;
- la convergence nécessaire entre plusieurs domaines de l'informatique (recherche d'information (RI), ingénierie documentaire, représentation des connaissances, IC, ...). A titre d'exemple, sur le terrain des ontologies, les collaborations entre RI et IC doivent se poursuivre pour définir des solutions vraiment adaptées.
- une influence anglo-saxonne forte caractérisée par (a) une approche technique du problème, focalisée sur la représentation des connaissances, l'interopérabilité et la réutilisation, (b) des hypothèses de généralité et d'universalité des ressources ontologiques (c) la recherche de l'automatisation des processus.

Cette dynamique découle en partie de projets aux budgets ambitieux sur ces sujets en Grande-Bretagne et en Allemagne, ainsi que de projets européens, qui permettent d'aborder le problème en vraie grandeur. L'évaluation de propositions scientifiques en la matière passe par des expérimentations, suppose des développements informatiques coûteux, qui ne sont possibles qu'au sein de projets conséquents, regroupant plusieurs laboratoires.

Ainsi, l'engouement que suscite le Web Sémantique a transformé l'IC radicalement au niveau international. On assiste sans doute là à un phénomène comparable à celui des systèmes experts dans les années 80 : beaucoup d'espoirs sont placés dans des technologies dont on promet des avancées sans doute exagérées et idéalisées ; la recherche est en partie doublée par des start-up spécialisées dans la mise en œuvre de ces technologies. Mais ce phénomène est compliqué par les particularités du web : les technologies et les innovations avancent à une vitesse fulgurante grâce à des entreprises innovantes et à de nouveaux modèles économiques ; les utilisateurs diffusent ou rejettent eux-mêmes les technologies selon leur pertinence à l'usage ; il y a un risque de confusion entre la standardisation des formats de mise en correspondance des contenus et l'harmonisation de ces contenus (mythe de l'ontologie universelle). Rapidement, le verdict va tomber sur une proposition technologique qui deviendra un succès mondial ou un échec immédiatement oublié.

Entre web sémantique et web2.0

Finalement, on peut être dubitatif sur les suites du Web Sémantique, car ce web demande la maîtrise de technologies complexes et surtout un effort considérable pour décrire des connaissances formellement selon des règles assurant leur qualité. Depuis 2005 environ, ce choix et cette interprétation sont bousculés par les propositions du Web 2.0, dit « web social ou collaboratif ». Le succès des wikis, blog, folksonomies, sites de partage et d'annotation libre de ressources démontre que des alternatives sont possibles et largement pratiquées par les internautes. Ces alternatives donnent une part plus grande à l'utilisateur, qui prend en charge l'inventaire de mots clés personnels ou partagés, puis l'annotation de ses ressources à l'aide de ces mots-clés. Plus simples, ces solutions constituent des descriptions de connaissances moins formelles, souvent moins rigoureuses et ne permettant pas de raisonnements. Mais elles suffisent dans une perspective peu automatisée, où c'est l'utilisateur qui prend en charge une partie de la recherche d'information.

Ainsi, paradoxalement, parce que de plus en plus de vocabulaires d'annotation et de documents annotés (images, films, profils de compétences, textes) sont disponibles, le web actuel se rapproche chaque jour un peu plus de la « vision » du futur web de T. Berners Lee. Mais ce qui ressort, c'est que l'approche sémantique rigoureuse et formelle qui fait appel aux ontologies s'avère trop lourde par rapport à la dynamique sociale des folksonomies. La répartition de l'effort requis pour l'accès à l'information, à savoir la description structurée, organisée de connaissances, l'inventaire de vocabulaires et la formalisation des raisonnements, n'est pas celle prévue par le web sémantique. Conscients de ces limites, les chercheurs en IC et la communauté du web sémantique cherchent de nouvelles voies. La solution à venir sera certainement un compromis entre une approche ascendante basée sur les contributions des utilisateurs et des modèles accessibles mais peu

rigoureux, et une approche plus élaborée, au potentiel de traitement plus puissant, gérée de manière plus centralisée par les développeurs d'applications. C'est là l'objet de plusieurs travaux actuels.

Place de l'IC française dans les recherches sur le Web Sémantique

Dans ce tourbillon, que devient le « knowledge engineering » ? Est-il dans, à côté, à la traîne, à la marge du web sémantique ? Que devient l'ingénierie des connaissances en France ? Quelle est la place d'une démarche scientifique à moyen terme qui essaierait de se dégager un peu de la contrainte technologique pour y retrouver des questions récurrentes et plus fondamentales ? L'ingénierie des connaissances peut avoir une place claire et originale si elle la défend, sur l'articulation langue – connaissances en particulier, que ce soit pour construire ou utiliser des modèles comme les ontologies, sur l'évolution, la prise en compte des usages et la dimension technologique de la maintenance de ces infrastructures sémantiques. Les contributions historiques des sciences humaines à l'IC devraient ressortir plus et être mises en avant pour faire des propositions plus proches des usages. Avec plusieurs collègues, nous avons organisé un atelier « IC2.0 » associé à la conférence IC 2008 dans ce sens.

Cependant, seules environ 5 équipes françaises sont visibles dans le paysage du web sémantique, contribuant à des recherches formelles, très informatiques. De fait, très peu d'équipes françaises ont réussi à être impliquées dans des projets ou réseaux européens. De plus, contrairement à l'Angleterre, ce n'est que depuis 2005 que des projets nationaux (ANR) sont financés sur ce sujet (WebContent, eWok, Dafoe4App ...). Or, des compétences existent et sont reconnues, en particulier dans l'acquisition supervisée de connaissances à partir de documents spécialisés, ou dans l'étude du contenu des ontologies. La communauté française défend une position à part, qu'elle fait entendre au niveau européen, en problématisant les questions au lieu d'y répondre rapidement et pragmatiquement, et enfin, en abordant le web sémantique selon une approche pluridisciplinaire pour prendre en compte toutes ses dimensions.

3.2.2 Modélisation de connaissances, ontologies et textes

Depuis 1995 environ, la convergence des besoins en identification de connaissances à partir de textes et les avancées du traitement automatique des langues, de la linguistique de corpus et les évolutions des théories et pratiques en terminologie ont permis d'envisager de nouvelles approches pour exploiter les textes comme sources de connaissances. En France, le groupe Terminologie et Intelligence Artificielle s'est fondé en 1993 autour de cette problématique, ouvrant de nombreuses perspectives (méthodes et outils) pour construire terminologies et ontologies, identifier des concepts et des relations conceptuelles à partir de textes. Grâce aux réflexions menées dans ce groupe (que j'ai co-animé de 1998 à 2008), les chercheurs concernés ont produit des résultats publiés (méthodes Archonte et Terminae, logiciels Syntex, Doe, Ana ...), organisé le workshop international « Ontology and text » en 2000 et animé l'action spécifique « Sémantique et corpus » en 2003-2004.

Depuis, ces questions bénéficient de nouvelles convergences entre recherche d'information, extraction d'information, traitement automatique des langues, apprentissage et IC. L'articulation langue-connaissances se pose doublement : depuis les textes vers les modèles, pour la construction de ces modèles ou l'inventaire d'instances des modèles ; inversement, au moment d'utiliser un modèle pour annoter, caractériser un document en fonction de son contenu. La question de l'annotation est au cœur du web sémantique autant que celle de la disponibilité des ontologies : qui annoter, quand, avec quelles méta-données, pour quel objectif, avec quels outils et selon quels principes ? Le groupe TIA était précurseur en matière de construction d'ontologies à partir de textes. Depuis 2004, cet objectif fait aussi l'objet de collaborations d'équipes de TAL et d'IC en Angleterre (Sheffield, KMI) et en Allemagne (DFKI). Les livres de Maedche (2001), Buitelaar et al. (2005) et Cimiano (2007) ainsi que les workshops OLP (Ontology Learning and Population) et OntoLex (Ontology and Lexicon) témoignent du dynamisme et des avancées de ces questions. A la différence de ceux de TIA, ces travaux cherchaient à automatiser le plus possible la construction de noyaux de modèles à valider. Des articles récents considèrent les résultats du TAL comme des éléments à interpréter pour prendre des décisions de modélisation.

4 METHODES ASCENDANTES POUR L'INGENIERIE DES CONNAISSANCES

Parmi les problématiques de l'IC, c'est la partie amont, la mise en place de modèles adéquats à partir de traces de connaissances, qui m'intéresse particulièrement. De ce fait, je me suis focalisée sur trois des rôles que jouent les modèles conceptuels : cible de la modélisation pour définir le système à construire, structure permettant d'organiser et présenter ce qui est analysé et enfin, grille pour repérer les lacunes et orienter la suite du processus. Pour aborder ce problème, j'ai développé plusieurs propositions relevant toutes de ce qui est appelé une « démarche ascendante », à savoir une méthode, des logiciels et langages définis en vue d'une meilleure localisation, explicitation et mise en forme de connaissances à partir de leurs usages afin de construire des modèles. Je présente mes différentes contributions selon l'historique de mes travaux.

4.1 Modélisation de connaissances expertes

4.1.1 MACAO, une méthode pour l'acquisition de connaissances expertes

La première partie de mes travaux s'est focalisée sur l'expert humain et les activités des spécialistes comme sources de connaissances. L'orientation choisie était déjà interdisciplinaire : d'une part, s'appuyer sur les travaux de la psychologie cognitive sur la nature des connaissances expertes pour mieux connaître les processus de résolution de problème, savoir les identifier et faire expliciter les connaissances mises en œuvre par des experts ; d'autre part, évaluer et adapter différentes techniques d'analyse de la tâche utilisées en psychologie et surtout en ergonomie pour les intégrer dans une méthode de construction de systèmes à base de connaissances. Ces recherches, menées au cours de ma thèse (1986-1989), ont produit une des premières méthodes d'acquisition de connaissances, MACAO, qui proposait des repères et des supports à différentes techniques d'entretiens. La méthode est accompagnée d'une plate-forme de modélisation intégrant des outils de recueil de connaissances ainsi qu'une représentation de connaissances au niveau conceptuel, à l'aide de schémas. La méthode a été utilisée sur des cas d'école (Aussenac-Gilles, Matta, 2004) et dans le cadre du projet SAMIE³⁶. Elle a été présentée aux toutes premières journées du PRC-IA sur ce thème (1988 et 1989), à la première édition des Journées d'Acquisition des Connaissances (JAC) en 1990 (Aussenac et Soubie, 1990) et aux conférences internationales EKAW et KAW dès 1988 (Aussenac et al 1988) (Aussenac et al, 1989a et b). Enfin, une collaboration avec R. Dieng a consisté à comparer son outil 3DKAT à MACAO (Aussenac et Dieng, 1991).

4.1.2 MACAO-II, modélisation de connaissances et opérationnalisation

La thèse de N. Matta (que j'ai co-encadrée de 1991 à 1995) a fait évoluer MACAO de manière à mieux rendre compte de la méthode de résolution de problème au sein d'un modèle conceptuel. Pour cela, elle a défini une représentation des connaissances qui assure la continuité de l'expression des connaissances depuis le langage naturel jusqu'au système opérationnel, le langage MONA. La nouvelle version de la méthode et de la plate-forme associée, MACAO-II, permettent de gérer des modèles de tâches de la bibliothèque de KADS (Schreiber et Wielinga, 1993) (Breuker et Van de Velde, 1994) et de les adapter pour construire le modèle conceptuel d'une expertise, écrit en MONA. Plusieurs collaborations avec d'autres laboratoires ont débouché sur une opérationnalisation des modèles à l'aide de langages de tâche : la version de base utilise LISA, langage défini à la DER d'EDF par I. Delouis et J.P. Krivine ; un module s'appuie sur le langage ZOLA développé à l'IRIN, équipe de P. Tchounikine (Beaubeau et al, 1996) ; une coopération avec F. Tort et C. Reynaud du LRI a permis de reprendre certaines propositions d'ASTREE (LRI), pour enrichir la représentation des connaissances du domaine, et mieux formaliser les relations (Reynaud et al., 1998) ; enfin, l'articulation entre connaissances du domaine et raisonnement, à travers la

³⁶ Projet mené avec la société MMS au sein du laboratoire ARAMIHS au cours de mon stage post-doctoral

notion de rôle, a été étudiée avec le LRI et l'IRIN (Reynaud et al., 1997). Plusieurs évaluations expérimentales de MACAO-II, comme le projet SADE (1993), ont montré l'intérêt d'exploiter la complémentarité entre méthodes ascendantes (constructives) et descendantes (par réutilisation de modèles de résolution) (Aussenac-Gilles, 1994) (Lépine et al, 1996). Enfin, dans le cadre d'une collaboration avec J. Breuker de l'Univ. d'Amsterdam, des outils ont été définis pour assurer une maintenance aisée et cohérente du modèle conceptuel et de la base de connaissances associée. À partir du module d'opérationnalisation en ZOLA, le langage MONA a été enrichi afin de mieux tracer le processus de modélisation et faciliter la maintenance du système.

4.2 Modélisation à partir de textes

4.2.1 Une évolution thématique

A partir de 1993, mes travaux ont pris un tournant en se focalisant aussi sur les documents comme sources de connaissances, sur les outils d'analyse terminologique comme moyen de les exploiter mais aussi sur les modèles conceptuels (du domaine ou de la tâche) pour faciliter leur « lecture » ou la recherche d'information. La motivation initiale à exploiter des textes visait un gain de temps pour le repérage de la terminologie et pour la structuration des concepts du domaine. L'intérêt d'étudier l'utilisation des modèles pour parcourir des textes et y rechercher de l'information est d'élargir les perspectives d'utilisation des modèles conceptuels au-delà de la mise au point de systèmes formels de raisonnement.

Ce glissement thématique a bénéficié de plusieurs projets menés avec la DER d'EDF. Dès 1993, dans le cadre du projet SADE, nous avons utilisé le logiciel extracteur de termes LEXTER de D. Bourigault, alors chercheur à la DER-EDF. LEXTER a permis de dégager la terminologie du domaine à partir de documents techniques, et, à partir de ces termes, de définir des concepts et de regrouper des synonymes. Ainsi, le modèle conceptuel est relié aux textes « sources ». Dans le projet Mougis, nous avons défini un modèle de tâche intégré dans l'interface d'un document technique électronique à côté de la table des matières et de l'index (Gros et al., 1996).

Les résultats prometteurs obtenus m'ont amenée à explorer plus systématiquement la manière de conduire des analyses terminologiques en amont de l'acquisition des connaissances, en étroite collaboration avec des linguistes du laboratoire ERSS (Toulouse2). Une convergence d'intérêt avec la linguistique de corpus et la structuration de terminologies, ainsi que l'émergence des ontologies comme support à la modélisation d'un domaine, ont donné un caractère plus ambitieux à cette piste. La période entre 1993 et 1998 correspond donc à celle d'une évolution thématique qui s'est stabilisée avec la confirmation de l'intérêt de cette approche pour la construction d'ontologies et de l'importance (provisoirement exagérée sans doute) des ontologies dans les applications comme la recherche d'information et la gestion documentaire.

4.2.2 Vers des ressources termino-ontologiques

Des bases de connaissances terminologiques ...

Ces recherches ont porté tout d'abord (1995) sur la notion de *bases de connaissances terminologiques* (BCT), leur représentation et les supports logiciels requis pour les construire. Ces bases contiennent des connaissances sur la terminologie d'un domaine, sous la forme d'un réseau conceptuel associé à des fiches terminologiques (Aussenac-Gilles et Condamines, 1997). Grâce au financement de projets de valorisation avec la DER d'EDF Mougis puis Hyperplan) et la région Midi-Pyrénées (DDE, 1998), j'ai encadré le développement de logiciels pour leur construction (GEDITERM, Fournier 1998) et leur utilisation (CONSULTERM, Lecorgne 1998). GEDITERM est un des rares systèmes opérationnels de ce type et constitue une contribution significative au niveau national et international sur deux points (Aussenac-Gilles, 1999) :

- il comporte un modèle de données permettant d'associer de manière souple termes et concepts, d'organiser les concepts au sein d'un réseau sémantique et de conserver un lien vers la justification linguistique de cette modélisation ;

- il est un des tout premiers éditeurs de modèle permettant d'importer les résultats d'un extracteur de termes (Lexter en l'occurrence).

Les expériences de spécification puis d'utilisation de ces outils m'ont permis d'étudier le passage de données lexicales à un modèle conceptuel, la traçabilité des choix de modélisation et le rôle des textes dans ce cadre. Du point de vue méthodologique, j'ai évalué en quoi une BCT pourrait être un produit intermédiaire utile pour la construction de modèles du domaine. Une réflexion approfondie a porté sur la nature des connaissances représentées à l'aide d'une BCT, leur distance par rapport au texte et la part d'interprétation faite par le linguiste au moment de définir des concepts et organiser le réseau conceptuel (Aussenac-Gilles et Condamines, 2001). Il en est ressorti que la spécificité des BCT n'est pas de refléter fidèlement les connaissances identifiées dans un texte, mais d'être des modèles faiblement formalisés d'un domaine, construits pour une application précise (rendre explicite des notions, faciliter la communication ou l'accès à des documents) (Aussenac-Gilles et al, 2002).

Ces projets ont également donné lieu à une première étude sur l'extraction de relations lexicales à partir de textes grâce à une collaboration avec D. Garcia (Garcia et al, 1996) (Garcia et al 2000). Son logiciel COATIS applique la méthode d'exploration contextuelle pour repérer des relations causales entre événements. L'étude des relations est un des éléments qui montre à quel point le modèle construit à partir de textes, y compris dans le cas d'une BCT, résulte d'un processus de décision. L'hypothèse sur le rôle des BCT a été alors revue, suite à l'impossibilité de rendre compte de manière neutre et exhaustive du contenu d'un texte, et pour des motivations plus théoriques liées à l'importance *de l'interprétation* au cours de la construction de modèle.

En parallèle, en collaboration avec des linguistes (A. Condamines et D. Bourigault de l'ERSS), nous avons étudié l'intégration de l'analyse terminologique et de ses résultats dans le processus de modélisation, et montré ainsi comment les avancées récentes du TAL permettaient d'envisager l'acquisition de connaissances à partir de textes sous un angle nouveau (Aussenac-Gilles et al., 1995). Au cours de différents projets, nous avons expérimenté ou validé des logiciels d'aide à l'extraction d'éléments linguistiques porteurs de connaissances, et donc utiles pour la construction de modèles, comme les concordanciers (YAKWA), les extracteurs de termes et de réseau terminologique LEXTER puis SYNTAX, des logiciels d'analyse distributionnelle (UPERY). Il en est ressorti un manque d'intégration entre ces logiciels, et l'absence de système spécialisé dans la recherche de relations sémantiques. Or les relations sémantiques sont un des moyens de repérer des concepts et de contribuer à leur définition. La thèse de P. Séguéla a donc été lancée sur ce sujet. L'approche retenue utilise des patrons linguistiques pour le repérage de relations lexicales puis la mise en relation conceptuelle. Elle a débouché en 2000 sur la mise au point du logiciel CAMELEON, dont deux nouvelles versions ont été depuis développées afin de repérer des relations dans des textes étiquetés grammaticalement. Cette thématique est des points forts actuels de l'équipe IC3.

... aux ontologies

En parallèle à cette focalisation sur les relations, nous avons établi un rapprochement naturel entre base de connaissances terminologiques et ontologies, alors que ce type de modèle de domaine devenait l'objet de toutes les focalisations en IC. Grâce aux échanges scientifiques et collaborations menés dans le groupe TIA entre 2000 et 2005, en particulier sur la nature des termes et des concepts ainsi que le statut de modèles comme les terminologies, les BCT et les ontologies, de nouvelles perspectives se sont ouvertes pour mes travaux. J'ai repris les travaux et résultats établis jusque là sur la modélisation de connaissances d'un domaine à partir de l'analyse de textes pour en étudier l'adaptation pour la construction d'ontologies.

La complémentarité de nos résultats et des recherches menées au LIPN par B. Biébow et S. Szulman a permis de spécifier une méthodologie de construction de modèles terminologiques et ontologiques à partir de textes : TERMINAE. Ces modèles, appelés « ressources termino-ontologiques », correspondent aux ontologiques lexicales de (Maedche, 2000). Cette méthode s'appuie sur un logiciel de modélisation dédié qui permet de construire une ontologie représentée en logique de description. La collaboration avec le LIPN a débouché sur l'intégration de plusieurs

logiciels d'analyse de texte (extracteur de termes et de relations) au sein d'une chaîne de traitements dont TERMINAE récupère les résultats et à partir desquels on peut définir une ontologie à composante terminologique.

4.2.3 Problématique et partis pris théoriques

Les problématiques abordées concernent avant tout l'exploitation des textes comme sources de connaissances pour la construction d'ontologies et de modèles de domaines, sous les angles méthodologiques, technologiques (logiciels d'analyse de textes), de la représentation des connaissances et plus fondamentalement, de la nature des éléments manipulés (textes, termes, concepts, ontologies etc.). Dans la plupart des applications ciblées, j'ai abordé une problématique symétrique, liée à l'utilisation des modèles pour caractériser les informations contenues dans des textes, constituer des index et faciliter la recherche d'informations.

Les travaux et résultats associés concernent d'abord les aspects méthodologiques de la construction et de l'évolution d'ontologies à composante terminologique, l'extraction de relations sémantiques à partir de textes et le modèle de données de ces modèles. Une autre série de travaux portent sur l'annotation sémantique et l'utilisation d'ontologies pour l'accès au contenu de documents. J'ai formulé ces deux problématiques conjointement de manière à tirer profit de leur dualité. Enfin, j'aborde ces questions en référence à des choix théoriques confirmés au fil des réflexions menées au sein de TIA et de mes collaborations (Aussenac-Gilles et Condamines, 2007):

- sémantique textuelle / sémantique de corpus : c'est l'usage des termes dans la globalité du corpus qui leur donne du sens et oriente vers la définition de concepts. Ce type de sémantique est pertinent pour rendre compte des phénomènes complexes, observés en corpus, de polysémie (y compris dans les corpus spécialisés), de glissement et de stabilisation de sens des termes dans le temps. Ces mêmes constats ont conduit à l'évolution théorique de la terminologie, qui remet en question la vision « classique » de Wüster.
- les ontologies sont vues comme des modèles régionaux, adaptés à un usage particulier : même au sein d'un domaine, les raisonnements ou des tâches qui seront faits avec l'ontologie en déterminent le contenu ; ainsi, on peut établir un lien entre la nature des modèles, leur contenu et leur degré de formalisation d'une part, et le type d'application visé d'autre part.
- la structuration des connaissances est donc justifiée à la fois par l'usage dans la langue et par l'application ciblée. Les ontologies construites ne sont pas linguistiques mais bien « de domaine » au sens où les modèles ne prétendent pas rendre compte de différenciations exprimées dans la langue mais de celles pertinentes pour le modèle et l'application visée.

4.2.4 Représentation des connaissances intégrant une composante terminologique

Depuis les travaux sur les bases de connaissances terminologiques (BCT), nous défendons l'intérêt de termes et d'informations terminologiques pour documenter un modèle conceptuel, mieux le maintenir, l'utiliser dans des interactions homme-machine et l'utiliser pour annoter des documents (Aussenac-Gilles et Condamines, 1997) (Séguéla et Aussenac, 1997). Il s'agit d'informations linguistiques, de nature différente du réseau conceptuel, qui se justifie quel que soit le degré de formalisation du modèle (Aussenac-Gilles et Condamines, 2001). Nous avons implémenté un modèle de donnée pour gérer informatiquement des BCT, au sein de Géditerm (Aussenac, 1999) et Consulterm. Ce modèle permet de rendre compte de la polysémie, de conserver des justifications linguistiques des concepts grâce aux termes, et des relations grâce à des patrons de relation. Il documente également les structures conceptuelles par des fragments du corpus à partir desquelles elles sont définies.

Sur la base du modèle de BCT, une de mes contributions à Terminae a consisté à proposer d'associer une composante terminologique à une ontologie du domaine, et de capitaliser le savoir-faire acquis en matière de construction et représentation des BCT (Szulman et al, 2001). C'est ainsi qu'a vu le jour au sein de TIA la notion de Ressource Termino-Ontologique (RTO) qui associe des

éléments de lexique ou des informations linguistiques aux concepts et relations d'une ontologie. Ce premier modèle a été repris et affiné par S. Szulman (2004).

Ce modèle a ensuite été redéfini par A. Reymonet dans sa thèse afin de pouvoir manipuler directement les termes au sein d'une ontologie écrite avec le langage standard OWL. Une première version respecte la syntaxe de OWL-DL (Reymonet et al., 2007a et b). La classe Terme, définie comme sous-classe de OWL-Thing, est reliée à la classe Concept par un lien « dénote ». Ce modèle, manipulable dans tout éditeur important le format OWL, n'entrave pas l'utilisation de raisonneurs. Ainsi, une extension de Protégé-OWL a été développée pour gérer ce modèle de RTO. Une 2^e version a été proposée en utilisant les méta-propriétés selon la syntaxe OWL-full. Il a fait l'objet d'une publication et d'une conférence invitée (Reymonet et al., 2009).

Ce travail est repris actuellement pour définir le modèle de données des projets DAFOE4App et Dynamo. Dans la plate-forme DAFOE, le modèle de données comporte trois niveaux : l'articulation terme-concept se joue au niveau du modèle conceptuel, alors que les occurrences des termes et les textes se situent au niveau corpus. Afin de pouvoir exporter une représentation en OWL de l'ontologie et de sa composante terminologique, le modèle défini par A. Reymonet est repris au niveau formalisation. Il est enrichi par la représentation d'information associée au repérage des relations dans les textes, évolution également nécessaire dans Dynamo.

4.2.5 Des textes aux applications : méthodes et plate-forme de modélisation

La méthode Terminae

Après les premières études méthodologiques sur les BCT (Aussenac-Gilles et Condamines, 2004), j'ai participé à la mise en forme de la méthode Terminae entre 1999 et 2007 en collaboration avec le LIPN et le groupe TIA. Terminae suppose que les textes seuls ne peuvent conduire à des ontologies au sens strict, et qu'une intervention humaine est indispensable pour assurer une cohérence dans la définition des concepts. La méthode définit donc un processus supervisé.

Mon intervention a consisté à intégrer une étape de constitution de corpus (selon des critères précis liés à l'ontologie visée) ; à préciser les étapes d'analyse terminologique du corpus puis de normalisation ; à faciliter la traçabilité du processus par l'utilisation de fiches terminologiques et la représentation de termes ainsi que la documentation de l'ontologie par des fragments de texte (Szulman et al., 2002). Terminae marque une différence significative entre une phase dite de modélisation conceptuelle, où l'accent est mis sur la nature des connaissances à modéliser, et la formalisation proprement dite (selon un langage comme OWL), où les capacités d'inférence prennent plus d'importance. L'analyse terminologique fait appel à l'extraction de termes pour l'identification de concepts puis à l'étude des relations lexicales pour définir des relations sémantiques au niveau conceptuel. Le point fort de la plate-forme associée est de récupérer facilement les résultats d'extracteurs mais aussi, de s'appuyer sur un modèle de données qui assure la traçabilité du modèle vers les textes sources, et donc des éléments conceptuels vers les termes. Terminae permet de construire aussi bien des ontologies à composante lexicale que des terminologies. Nous appelons ces modèles « ressources termino-ontologiques », et elles correspondent aux ontologiques lexicales de (Maedche, 2000).

Réflexions méthodologiques

Plus fondamentalement, l'ensemble de ces travaux a permis une réflexion sur l'apport d'éléments terminologiques et linguistiques pour améliorer la qualité des modèles, mieux répondre aux besoins des utilisateurs, en faire valider et adopter le contenu. D'autres éléments méthodologiques ont pu être étudiés :

- *influence de l'application* pour laquelle une RTO est construite sur son contenu. Un premier travail a été grâce à la confrontation de plusieurs expériences au sein de TIA (Aussenac-Gilles et al., 2002) puis (Bourigault et al, 2004). Une étude plus fine a été présentée par A. Reymonet sur les données du projet Mode (Reymonet et al., 2006).

- *validation des modèles* : il est difficile d'identifier les classes d'applications pour lesquelles une ontologie construite avec Terminae est pertinente, son adéquation aux besoins des utilisateurs ou encore la possibilité de la réutiliser pour d'autres applications. J'ai participé au premier atelier EON sur l'évaluation d'ontologie et de méthode de construction par l'utilisation de l'ontologie dans une application (Aussenac-Gilles, 2002). Ce travail n'a pas été poursuivi faute de temps.
- *part des connaissances qui doivent être formalisées* (et qui seront accessibles au système informatique) : dans le cas d'utilisation d'ontologies pour l'accès à des textes, on peut se demander jusqu'à quel degré formaliser les connaissances, et lesquelles peuvent être laissées dans leur forme d'origine.
- *terminologie et évolution des connaissances dans le temps* : une expérience, menée avec D. Bourigault et R. Teulier (2003), a consisté à comparer des termes extraits de deux corpus scientifiques d'un même domaine à 5 ans d'intervalle ; la question de l'évolution dans le temps a été identifiée comme un des axes à étudier par le groupe ASSTICCOT, et a donné lieu à un numéro spécial de la revue I3 (Aussenac-Gilles, Condamines, Sedes 2006).

Plusieurs retours d'expérience de projets académiques et industriels ont conduit à faire évoluer Terminae et en affiner les principes. Ces évolutions ont bénéficié des avancées du TAL (nouveaux extracteurs de termes, meilleure efficacité des analyses syntaxiques) ainsi que de travaux sur l'ingénierie des ontologies (export et import en format OWL, mapping d'ontologies et réutilisation). Ainsi, la méthode Terminae permet désormais la réutilisation d'ontologies de référence (core ontologies) de domaines généraux comme le droit. Un article récapitulatif l'évolution de la méthode depuis 10 ans a été publié dans un ouvrage de synthèse sur l'apprentissage d'ontologies à partir de textes (Aussenac-Gilles et al., 2007).

Enfin, ces travaux ont eu une visibilité au niveau international grâce à la co-organisation d'ateliers et la participation à des ateliers et séminaires. Une confrontation de nos points de vue avec ceux du chercheur américain en sciences de l'information D. Soergel a été retenue pour publication dans le numéro de lancement de la revue Applied Ontology (Aussenac-Gilles, Soergel, 2005).

La plate-forme DAFOE

L'ensemble de ces résultats et choix relatifs aux méthodes, outils et représentations pertinents pour construire des ontologies à partir de textes ne sont pas le fruit d'un travail individuel, mais ils reflètent la convergence de travaux menés par plusieurs équipes francophones. La maturité de la recherche a permis de poser les bases du projet DAFOE4App, mené par J. Charlet et impliquant plusieurs équipes française dont le LIPN, l'AP-HP, l'ENSMA et la société Mondeca. L'objectif du projet est de définir une plate-forme offrant une architecture ouverte pour accueillir une boîte à outil modulaire destinée à faciliter la construction d'ontologies à partir de textes et par réutilisation d'autres ressources, en utilisant des logiciels de traitement du langage. La plate-forme met donc à disposition de ses utilisateurs des langages de représentation des connaissances et des modules intégrant des outils de TAL, selon une approche où l'utilisateur reste à l'initiative du processus de modélisation. La présence de PME concernées par la diffusion et la pérennité de cette plate-forme en garantit le caractère opérationnel. Un prototype est en cours de développement, dont les principes ont été présentés par des posters et démonstrations aux conférences IC, JFO, EGC, TALN en France ; TIA et KEOD à l'international (Charlet et al., 2008) (Charlet et al., 2009a, b, c) (Szulman et al., 2009) (Charlet et al., 2010).

4.2.6 Evolution des modèles de connaissances dans le temps

Maintenance d'une hiérarchie de concepts à l'aide d'un SMA : DynamO

Pour demeurer pertinentes dans les systèmes qui les utilisent, et pour qu'elles restent cohérentes avec les documents qu'elles servent à annoter, les ontologies doivent être mises à jour. Si l'on considère le cycle de vie d'une ontologie comme un système complexe constitué de concepts, on peut l'implémenter dans un système multi-agents (SMA) adaptatifs. Le projet Dynamo, défini avec l'équipe SMAC de l'IRIT, propose de gérer par un SMA le cycle de

construction et évolution des ontologies à partir de texte. L'état stable du système résulte des interactions coopératives entre les agents logiciels qui les constituent. Dans notre cas, les agents utilisent des algorithmes distribués de clustering statistique pour trouver la structure la plus satisfaisante d'après une analyse syntaxique, terminologique et distributionnelle des textes. L'utilisateur peut alors valider, critiquer ou modifier des parties de cette structure d'agents, qui est la base de l'ontologie en devenir, pour la rendre conforme à ses objectifs et à sa vision du domaine modélisé. En retour, les agents se réorganisent pour satisfaire les nouvelles contraintes introduites. Dans le projet Dynamo, les ontologies, habituellement vues comme des structures figées, deviennent des ensembles dynamiques, capables de s'adapter à leur contexte (textes et utilisateurs).

La pertinence de cette approche a été mise à l'épreuve par des expérimentations visant à évaluer la complexité algorithmique du système, et par son utilisation sur un cas d'étude. Ce travail a été réalisé par K. Ottens dans le cadre de sa thèse, et l'évaluation a été menée avec N. Hernandez. Il a donné lieu à 2 publications à EGC 2007 (Ottens Aussenac-Gilles, 2007) et à l'atelier ESOE associé à ISWC2007 (Ottens et al., 2007), et à des articles relatifs à l'implémentation des agents (AMAW 2007) et à une article de revue internationale (Ottens et al., 2008).

Le projet DynamO a permis de reprendre cette étude en confrontant une approche à base d'agents à une approche « manuelle » réalisée par l'ontographe en fonction des besoins d'annotation sémantique de nouveaux documents. Ces deux approches font l'objet de deux thèses commencées fin 2008 : A. Tissaoui étudie l'influence réciproque des évolutions d'une ontologie et d'annotations sémantiques, pour mieux définir une aide à la gestion de ces évolutions (Tissaoui, 2009) ; Z. Sellami poursuit les recherches sur l'apport de systèmes multi-agents pour gérer les évolutions d'une ontologie en fonction de nouveaux corpus dont le contenu doit être reflété dans l'ontologie. L'approche de Z. Sellami repose sur des hypothèses différentes de celles de K. Ottens : l'analyse de textes repose sur une approche par patrons lexico-syntaxiques, de manière à organiser les concepts non seulement sous forme de hiérarchie, mais aussi en définissant de relations sémantiques de différents types entre concepts. Les premiers résultats sont assez prometteurs malgré des difficultés inhérentes à la faible taille des corpus étudiés (Sellami et al, 2009 a, b, c).

4.2.7 Outils pour l'extraction des relations conceptuelles

Recherche de relations sur corpus étiquetés : Caméléon

Défini en 2000 dans le cadre de la thèse de P. Séguéla, Caméléon implémente une approche supervisée d'extraction de relations sémantiques à l'aide de patrons lexico-syntaxiques et d'enrichissement d'ontologie. Les principes à la base de Caméléon s'avèrent des choix pertinents et toujours d'actualité (Séguéla, Aussenac-Gilles 1999) (Aussenac-Gilles Séguéla, 2000) :

- *un format précis de patrons* : la nécessité de définir des patrons riches, caractérisant le contexte d'apparition d'une relation entre deux termes x et y sous la forme $A \ x \ B \ y \ C$ où A , B et C caractérisent des formes d'expression d'une relation, et x et y caractérisent les termes recherchés ; suivant le niveau d'analyse des textes, ces caractérisations sont syntaxiques, lexicales, sémantiques ou liées à la mise en forme
- *un processus supervisé* : la variabilité sémantique des résultats retournés par la projection d'un patron suppose une interprétation humaine pour fixer précisément le sens de la relation et les termes reliés. L'objectif serait, à terme, de réduire cette intervention. Ce point de vue a été discuté au cours d'un atelier international sur l'apprentissage pour la construction d'ontologies (Aussenac-Gilles, 2005).
- *la nécessité d'adapter les relations recherchées et les patrons associés à chaque corpus* : les résultats de linguistique de corpus ont très tôt montré la forte variabilité du sens de la relation présente dans les phrases retournées par un patron, la variabilité d'efficacité d'un patron d'un corpus à un autre, et la grande diversité des relations présentes ; une extraction doit donc comporter une interface simple de mise au point de patron et si possible, assister ou automatiser la découverte de patrons pertinents dans un corpus donné.

- l'intérêt de rendre disponible une *base de patrons réutilisables et adaptables* : P. Séguéla a formalisé une base d'environ 200 patrons pour le français, repris de la littérature, couvrant une vingtaine de types de relations. Pour chaque corpus étudiés, les patrons peuvent être utilisés tels quels ou modifiés ou associés à un autre type de relation.

Deux nouvelles versions du logiciel, respectant ces mêmes principes, ont été définies pour rendre plus précise la formulation des patrons et plus efficace le repérage de termes reliés. Développé entre 2004 et 2005, CaméléonII correspond à une évolution majeure car il exploite des textes étiquetés par un analyseur syntaxique. Pour cela, il intègre le concordancier Yakwa, dont il reprend le format pour définir des patrons, et qu'il enrichit d'une interface d'évaluation et de validation des patrons ainsi que d'une interface d'enrichissement d'ontologie. Comme Yakwa ne permettait pas de caractériser facilement les termes mis en relation par le patron, CaméléonIII a été réalisée en 2005-2006 à l'aide d'un nouveau concordancier que nous avons développé : KesKya. En parallèle, l'écriture d'une partie des patrons de la base initiale au nouveau format a été entreprise.

Evaluation de CaméléonIII et mise au point d'une base de patrons réutilisables

Dans le cadre d'un contrat post-doctoral, Marie-Paule Jacques a mis au point une base de marqueurs de relations pour le français, évalués sur huit corpus différents. Globalement, cette étude a souligné l'importance du corpus sur l'efficacité et la sémantique des patrons. En matière d'ingénierie des connaissances, ce résultat confirme deux des hypothèses retenues dans Caméléon : nécessité d'adapter les patrons et les relations au corpus ; importance de l'interprétation humaine. En matière de TAL, la grande variation de performances selon le corpus testé conduit à deux conclusions : nécessité de fournir à l'utilisateur des informations sur les corpus avec lesquels les patrons ont été mis au point ; exploiter les outils de TAL pour contribuer à une meilleure caractérisation des textes et des genres textuels en relation avec les traitements possibles, au-delà d'une classification rigide des textes. Les résultats ont été publiés dans deux revues (TAL en 2007 et Terminology début 2008) et présentés à la conférence EKAW 2006. Une étude analogue est envisagée pour l'anglais.

De Caméléon III à la plate-forme DAFOE

Au sein du projet DAFOE4App, une des contributions de l'équipe IC3 est de définir un module de la plate-forme DAFOE, chargé d'aider à l'identification de relations sémantiques à partir de textes. Les modules logiciels assurant les traitements et analyses de textes requis pour la construction d'ontologies constituent autant de plug-ins associés au cœur de la plate-forme. Un travail est en cours pour faire de CaméléonIII un plug-in opérationnel de DAFOE. Une extension prévue de ce travail est de faire évoluer CaméléonIII pour mieux assurer la complémentarité avec les interfaces de définition de relations sémantiques aux 3 niveaux du modèle DAFOE (Aussenac-Gilles et Hernandez, 2009).

Une réflexion plus ambitieuse a été menée avec des linguistes pour revoir les enjeux linguistiques et informatiques de l'identification de relations sémantiques grâce à des patrons lexico-syntaxiques. En particulier, la variabilité observée d'un type de corpus à l'autre (suivant les domaines, les genres textuels, les langues, etc.) semble aujourd'hui confirmée par différentes analyses. Elles invitent à une grande flexibilité dans la définition des patrons et des types de relations, et soulignent la nécessité de disposer d'une capacité d'adaptation (ou d'apprentissage) des patrons à chaque nouvel usage pour un nouveau corpus (Aussenac-Gilles et Condamines, 2009).

A terme, une perspective plus ambitieuse est d'intégrer ces besoins pour définir de nouvelles fonctionnalités dans un logiciel de recherche de relations, ainsi que les différents travaux de l'équipe sur les relations au sein d'un ou plusieurs modules complémentaires pour l'extraction de relations dans DAFOE. Parmi les travaux prioritaire, il y aura en particulier l'apprentissage de patrons et l'étude de relations formulées sur plusieurs phrases ou à partir de la structure de documents (Aussenac-Gilles, Kamel et Hernandez, 2008).

Extraction d'information et recherche d'instances de relations conceptuelles

L'extraction de relations à partir de textes peut concerner l'identification de relations entre concepts pour construire une ontologie, mais aussi l'identification de relations entre instances de concepts pour « peupler » une ontologie ou annoter des documents. Caméléon vise le premier objectif. Au sein de l'équipe, nous avons commencé à nous intéresser au deuxième objectif, plus proche de celui de l'extraction d'information. Pour cela, M. Kamel a identifié une application nécessitant une extraction d'information s'appuyant sur une ontologie dans un domaine spécialisé. Les informations recherchées correspondent à des instances de relations entre concepts de l'ontologie. L'application a été choisie en collaboration avec des médecins chercheurs en biologie médicale. Il s'agit d'extraire des informations précises (des noms de gènes, leur localisation, des pathologies et des études scientifiques) à mettre en relation à partir de résumés d'articles scientifiques (Kamel et Perret, 2007). Pour cela, la plate-forme Gate a été évaluée et s'avère pertinente pour définir une chaîne de traitement depuis l'étiquetage grammatical des textes jusqu'à l'identification de relations. Gate propose d'utiliser des règles JAPE pour formuler des patrons caractérisant des relations, les projeter et enrichir ainsi une ontologie de nouvelles instances.

Cette étude a permis de souligner certaines limites de l'approche par patron, comme l'incapacité à trouver des relations formulées par des références entre plusieurs phrases, et de discuter la manière d'implémenter et projeter des patrons pour parvenir à plus d'efficacité. Il en ressort également le besoin d'élargir l'extraction de relations dans des domaines spécialisés (comme la biologie) aux cas de relations n-aires (Kamel, 2008).

Exploiter la structure de documents pour identifier des relations conceptuelles

Un travail original s'appuyant sur la structure XML de documents très particuliers (des spécifications de bases de données géographiques) a permis d'exploiter le texte et les relations pouvant être déduites soit de la structure, soit des relations syntaxiques dans les phrases. Ce travail, réalisé dans le projet GEONTO, permet de restituer des ontologies très fidèles aux documents et facilite la mise en place de relations d'alignement entre plusieurs ontologies associées à différentes bases de données géographiques (Aussenac-Gilles et Kamel, 2009) (Kamel et Aussenac-Gilles, 2009 a, b) (Kergozien et al., 2009) (Mustière et al., 2009).

4.3 Ontologies et terminologies pour la recherche d'information

4.3.1 Historique

Via des contrats de valorisation (laboratoire Autodiag, projet Saint-Gobain 2002), j'ai élargi l'éventail des applications permettant d'évaluer ces méthodes et outils, au delà de l'aide à la résolution de problèmes : gestion documentaire, mémoire d'entreprise, modélisation des utilisateurs, construction de systèmes coopératifs. Or chaque type d'application soulève des problèmes de recherche spécifiques, qui vont au delà de l'adaptation des logiciels. La majorité de ces expériences a concerné l'intérêt de modèles conceptuels ou terminologiques, donc de données sémantiques structurées, pour accéder à ou naviguer dans des éléments documentaires. La répartition de la « résolution de problème » au sein du couple système-utilisateur final est ici tout à fait inversée par rapport aux systèmes à base de connaissances. Le système exploite des connaissances du domaine pour orienter au mieux un utilisateur qui a l'initiative de la recherche, et surtout de l'interprétation du contenu documentaire en fonction du contexte dans lequel il réalise sa tâche. Au cours de différents projets, cette question s'est déclinée en mesurer l'intérêt d'un modèle de tâches pour la consultation de guides de procédures (projet Mougis), juger de l'apport d'un modèle conceptuel pour faciliter la sélection des termes et leur structuration dans un index (projet Hyperplan), structurer un index de site web selon une approche terminologique (IndexWeb), évaluer l'apport des ontologies pour la reformulation de requêtes (thèse de M. Baziz, 2005), la classification de documents spécialisés (thèse de N. Hernandez, 2005) ou encore pour la consultation de documents structurés (projet Arkeotek). Ces questionnements exigent une réponse interdisciplinaire entre spécialistes des sciences de l'information ou de recherche d'information, du TAL et de l'IC, linguistes et ergonomes. Ma participation à des groupes de travail (TIA, action spécifique « corpus

et terminologies ») et mes collaborations avec le laboratoire CLLE-ERSS m'ont permis de mener ce type de réflexion avec des chercheurs de ces disciplines.

Grâce au co-encadrement des thèses de N. Hernandez et M. Baziz, j'ai abordé la question de l'utilisation des ontologies pour la recherche d'information du point de vue de l'IC, en cherchant à identifier les types d'ontologie les mieux adaptées, leur mode de construction et la manière de les utiliser. Ces travaux ont montré que les ontologies sont des représentations d'autant plus pertinentes dans ce contexte qu'elles comportent une dimension terminologique, et même qu'elles peuvent être rattachées à des éléments linguistiques comme des patrons d'extraction d'information. La richesse terminologique est en effet à la fois une conséquence de la construction des ontologies à partir de textes et un atout pour les utiliser dans une indexation conceptuelle. Enfin, les outils du Traitement Automatique du Langage (TAL) reposent sur des éléments linguistiques comme les patrons d'extraction de relations ou de concepts qui peuvent faciliter le processus d'annotation ou d'indexation. L'indexation revient alors à chercher des instances de concepts dans les textes à l'aide de ces patrons : elle permet à la fois de "peupler" l'ontologie de nouvelles instances de concepts et d'indexer les textes. Cette analyse est synthétisée dans deux chapitres de livre (Aussenac-Gilles, Baziz Hernandez, 2006) et (Aussenac-Gilles 2008).

4.3.2 Ontologies et recherche d'informations générales

Afin d'évaluer l'apport de modèles conceptuels à la recherche d'information en amont d'un moteur de recherche général, deux études ont été menées successivement par M. Baziz dans son DEA puis sa thèse que j'ai co-encadrée. L'une porte sur la reformulation de requêtes en exploitant les relations entre concepts (Baziz et al. 2003), l'autre sur la représentation de documents sous forme d'un réseau puis d'un arbre de concepts (Baziz et al 2005).

La première expérience a donc consisté à utiliser une base de données lexicale pour la reformulation des requêtes utilisateurs. Différents tests ont été effectués pour évaluer ce processus qui conduit à une amélioration significative de la pertinence des réponses fournies par le moteur. Les expérimentations ont été réalisées en utilisant le moteur Mercure développé à l'IRIT, WordNet comme base de données lexicales et Clef2001 comme collection de test (Baziz et al, 2003). Pour assurer un gain dans les résultats retournés, un processus d'"expansion prudente" a été défini en amont d'un moteur de recherche. Ce processus, transparent à l'utilisateur, exploite d'abord la notion de concepts multi-termes pour désambigüiser les mots de la requête (au sens de WordNet). Il s'appuie ensuite sur les relations sémantiques entre concepts pour élargir la requête. Les modules d'expansion de requête en choisissant le type de relation ont été intégrés à la plate-forme de recherche d'information RFIEC³⁷, afin de pouvoir reproduire l'expérience sur d'autres corpus.

De manière symétrique, M. Baziz a défini une représentation sémantique des documents (Baziz et al, COLIS 2005) Cette représentation, appelée *noyau sémantique du document*, prend la forme d'un réseau de concepts jugés représentatifs du document et tirés d'une ressource générale ou « ontologie » (ici WordNet). Pour identifier automatiquement ces concepts, les documents sont « projetés » sur cette ressource : les concepts *représentant* un document sont choisis à partir des termes qu'il contient. Ensuite, les concepts sont pondérés (deux poids ont été étudiés : cf.idf inspiré du tf.idf, et le C_score, tenant compte des relations entre concepts dans l'ontologie). Seuls les concepts d'un poids supérieur à un seuil sont retenus. Le *noyau sémantique du document* représente le contenu informationnel du document à l'aide de nœuds (les concepts désambigüisés) et d'arcs (liens de similarité sémantique calculés à partir de relations présentes dans WordNet et d'une distance sémantique choisie). Le calcul de ce noyau, long et coûteux, est fait une fois pour toutes pour une distance donnée. Ainsi, la collection interrogée est représentée par l'ensemble des noyaux sémantiques des documents qui la composent. Lors d'une recherche, la requête est traduite sous forme de concepts et étendue selon les principes d'expansion prudente. Puis elle est comparée aux différents noyaux sémantiques pour identifier les documents les plus pertinents.

³⁷ <http://www.irit.fr/RFIEC>

Six distances ont été comparées sur un jeu de test pour calculer la proximité sémantique entre concepts. Il en ressort que la mesure de Resnik est la plus efficace sur la collection utilisée, combinée au calcul du C_Score (Baziz et al., COLIS 2005). L'ajustement des poids associés aux concepts s'avère d'un impact presque aussi important sur la qualité des résultats que le choix des concepts eux-mêmes. En effet, pour le moteur de RI utilisé, la représentativité des concepts (explicitée par leur pondération) est importante pour classer les documents répondant à une requête. Le système intégrant ces 2 modules, DocCore, a été évalué dans le cadre de CLE 2004.

M. Baziz a ainsi établi plusieurs résultats réutilisables, publiés au niveau national (inforsid 2003) et international (dont 5 conférences et 2 workshops spécialisés, Baziz, Boughanem et Aussenac-Gilles 2004 et 2005) :

- la nature des relations sémantiques a une influence significative sur l'expansion de requête : les résultats ne sont améliorés que si on n'exploite que les relations hiérarchiques est-un (les relations de méronymie ou d'antonymie au contraire dégradent les résultats) ;
- principe « d'expansion prudente » : l'expansion n'améliore les résultats que si on minimise aussi le nombre de concepts. Pour représenter le document et la requête, l'algorithme sélectionne les concepts les « mieux reconnus », et en cas d'expansion, substitue le concept relié au concept reconnu s'il semble plus approprié.
- proposition de pondération pour qualifier l'importance des concepts identifiés dans un document, le C-Score. Ce poids intervient pour sélectionner les concepts représentatifs d'un document.
- utilisation des termes synonymes et des termes de la glose d'un concept pour désambiguïser les différents sens possibles d'un terme trouvé dans un document
- la représentation du document sous forme de réseau, où les concepts sont reliés entre eux selon leur proximité sémantique dans l'ontologie, est plus efficace qu'une représentation sous forme de liste de concepts.

4.3.3 Ontologies pour naviguer dans des collections de documents spécialisés

La navigation au sein de collections documentaires à l'aide d'une ontologie passe l'annotation sémantique de ces documents. Cela revient à associer au document une représentation plus ou moins riche basée sur les concepts (Handschuh, 2004).

Principes pour une annotation sémantique élémentaire

Associer termes et concepts pose la question de l'identification des formes linguistiques des concepts et relations de l'ontologie, ou du repérage d'instance de concepts et de relation. Définir le format de l'annotation revient également à décider si ce sont des termes, des concepts ou des instances de concepts ou même des parties de réseau conceptuel qui caractérisent le texte. Les annotations peuvent être ou non pondérées. Enfin, il faut choisir la nature des fragments à annoter et la portée des annotations, autant que la manière de poser ces annotations (automatique ou supervisée). Nous abordons cette problématique sous deux angles proches dans les projets Arkeotek et Mode. Dans les deux cas,

- le point de départ est une ontologie relativement simple à composante lexicale (une question de recherche est donc de représenter cette ressource) ;
- l'ontologie est définie en fonction des besoins d'annotation (une 2^e question de recherche est de définir un processus pour la faire évoluer en fonction des corpus à annoter) ;
- Le processus d'annotation est supervisé : le système génère automatiquement une proposition d'annotation, basée sur la reconnaissance dans le corpus des termes associés aux concepts, mais aussi sur l'exploitation des relations entre concepts ; le système applique des critères d'évaluation de l'annotation et présente à l'annotateur les documents « mal annotés » ; cette proposition est corrigée par un spécialiste du domaine.

Annotation à l'aide de hiérarchies de concepts : OntoExplo

Les ontologies semblent une alternative pertinente aux listes de termes habituellement utilisées pour l'exploration de collections dans le cadre de la veille scientifique. Dans sa thèse, N. Hernandez (que j'ai co-encadrée avec J. Mothe) a défini un environnement d'exploration de collections (OntoExplo) à partir de plusieurs hiérarchies de concepts, les unes décrivant le domaine de la collection et les autres décrivant la tâche de veille (Hernandez, 2005). Ces hiérarchies (appelées ontologies dans OntoExplo) ne forment pas une ontologie unique mais autant de dimensions d'interrogation des collections. Elles organisent la spécialisation de concepts importants du domaine, la relation hiérarchique ayant une sémantique variable et précisée pour chacune (Est-un ou partie-de).. Ces concepts correspondent à autant de critères de consultation des documents. En parcourant la hiérarchie, l'utilisateur est guidé et peut affiner sa recherche : il réduit ainsi les classes de documents en fonction des concepts caractérisant leur contenu.

Les experts du domaine élaborent les hiérarchies manuellement. Ensuite, c'est l'analyse de corpus par un extracteur de termes qui en assure un enrichissement terminologique. Les concepts et le vocabulaire associé servent aussi de langage pour exprimer le besoin en information. L'indexation à l'aide des concepts des hiérarchies est assez immédiate. Elle exploite les termes associés aux concepts et des traitements linguistiques élémentaires comme la lemmatisation. Un environnement de visualisation, OntoExplo, présente plusieurs hiérarchies pour faciliter la focalisation sur des documents particuliers et en assurer la consultation rapide. L'utilisateur choisit des concepts, la collection est réorganisée en fonction de ces concepts, et l'utilisateur peut alors explorer la collection et naviguer entre les documents en suivant les relations entre concepts.

Ce travail est un premier pas prometteur qui illustre une exploitation élémentaire de la notion d'ontologie, avec une annotation sémantique également très simple. L'interface OntoExplo permettrait d'exploiter une plus grande variété de relations, mais la navigation au sein d'ensembles de documents progressivement raffinés aurait moins de sens. Exploiter la richesse des ontologies du point de vue de la navigation au sein de collections documentaires représente un défi stimulant et complexe. Le projet Arkeotek présenté ensuite illustre une approche comparable exploitant plus de relations.

Annotation sémantique par les concepts

Dans le cas du projet Arkeotek, les annotations sont des concepts (des classes) et les relations entre concepts ne servent pas à annoter. En revanche, elles permettent de suggérer des concepts reliés comme annotations. De plus, les fragments de textes à annoter étant reliés entre eux par des relations logiques, ces relations entre fragments sont prises en compte pour juger de la pertinence d'une annotation calculée automatiquement. Un prototype a été développé pour gérer l'ontologie et mettre en œuvre ce mode d'annotation. Il a été finalisé en 2005 et utilisé pour construire une ontologie de l'archéologie des techniques, corrigée en 2007 à partir d'une première évaluation pour l'annotation. Ce prototype est repris dans le projet Dynamo. Il a été présenté dans deux publications à des ateliers spécialisés sur ontologies et patrimoine (Aussenac-Gilles, Roux, Blasco 2006) (Aussenac-Gilles, 2006).

Annotation sémantique par un graphe d'instances de concepts en relation

Dans le cas du projet MODE, les annotations sont des instances de concepts et des instances de relations entre concepts. Les fragments de texte sont indépendants. Les critères de qualité de l'annotation portent sur le fait qu'un maximum de mots de ce fragment soient associés à une instance de concept, et ensuite sur le fait que certains types de concepts, que l'on s'attend à trouver impérativement dans ces fragments, soient présents. Un plug-in de l'éditeur d'ontologies Protégé, TexViz, a été développé pour mettre en œuvre ce processus. Ce système présente à l'utilisateur les textes les plus mal annotés. Il permet de modifier facilement l'ontologie, d'y ajouter ou de retirer ou de modifier des termes, des concepts, des relations ou leurs instances. Le prototype TexViz et le modèle de données associé ont été présentés à la conférence IC 2007 et au workshop OntoLex associé à ISWC2007.

Cet axe de recherche est au cœur du projet Dynamo. Nous l'abordons avec des équipes spécialistes de l'annotation sémantique impliquées dans Dynamo : Ph. Laublet du LALIC et J. Mothe de SIG-EVI.

4.3.4 Ontologies et identification statistique de communautés scientifiques

Ontologies pour mesurer l'intelligibilité de textes

Dans le cadre d'un projet CNES, B. Rothenburger, N. Chikhi et moi-même étudions l'aide que peut apporter une ontologie à composante terminologique pour localiser dans des textes des termes dont le sens aurait glissé dans le temps et qui renverraient à des sens différents donc à des concepts différents chez l'auteur et chez le lecteur. Une variante de ce problème se pose dans le cadre de l'annotation de données scientifiques par des mots-clés qui peuvent, eux aussi, être interprétés différemment dans le temps ou par des communautés scientifiques ou sociales différentes (Rothenburger, 2006).

Dans ce projet, nous avons défini les principes d'une utilisation simultanée de connaissances contenues dans des ontologies et des connaissances issues de textes afin de mesurer l'adéquation des secondes par rapport aux premières. Le but de cette confrontation est de s'assurer que les connaissances contenues dans des documents restent intelligibles lorsque les lecteurs sont culturellement éloignés des rédacteurs. Nous avons proposé un cadre concret pour mettre en œuvre ces moyens : un dispositif d'audit de grandes archives de données scientifiques visant à assurer la pérennité de leur intelligibilité. Nous caractérisons les propriétés requises pour les ontologies afin qu'elles puissent répondre à l'objectif de l'approche. Nous présentons enfin les moyens théoriques, pratiques et méthodologiques que nous utilisons pour assurer la confrontation des connaissances issues des deux sources. Exposé à l'atelier « ontologies et textes » de la conférence TIA 2007.

Approches statistiques pour l'identification de communautés scientifiques sur le web

Un nouveau contrat avec le CNES a permis de poursuivre l'étude sur l'évolution de sens de termes en ciblant l'accès à des données scientifiques par des communautés de chercheurs, avec une application à l'astronomie. Les observatoires virtuels sont des sites web où les astronomes mettent en ligne les données relatives à des observations ou à des expériences spatiales. Une première approche statistique a été expérimentée pour repérer des communautés de pratique à partir des liens électroniques existant entre ces observatoires. Puis l'étude a été étendue aux liens de co-citations entre publications de ces communautés. Dans les deux cas, les documents (pages web ou articles scientifiques) et les relations qui existent entre eux (citations ou références) forment un graphe. Ces graphes contiennent des connaissances implicites souvent appelées structures de communautés. Des algorithmes de classification statistique les exploitent pour identifier des classes de documents étroitement liés, dont on fait l'hypothèse qu'ils sont révélateurs d'une « communauté ». Des algorithmes proches permettent de repérer des nœuds attractifs et largement cités, les autorités, dont les informations sont a priori très fiables. D'autres nœuds jouent un rôle de portail, de synthèse de l'état de l'art et renvoient vers un grand nombre de documents, ce sont les hubs.

La thèse de N. Chikhi, que je co-encadre avec B. Rothenburger, porte sur l'application et l'amélioration de ce type d'algorithme. Il a proposé un premier ensemble d'amélioration des algorithmes de fouille de la structure du graphe de sites d'un domaine, et ce par des techniques de réduction de dimensionnalité comme l'analyse en composante principale, l'analyse en composants indépendants, etc. Il a obtenu les meilleurs résultats avec la factorisation en matrice non négative et l'information mutuelle normalisée (Chikhi et al., 2007). L'étude a ensuite porté sur l'analyse des citations bibliographiques (Chikhi et al, KES, 2008). Enfin, un dernier travail combine les deux sources d'information, sites web et publications, pour mieux identifier les communautés scientifiques. Un nouvel algorithme a donc été défini, basé d'une part sur une mesure de similarité originale prenant en compte simultanément citations et références, et d'autre part sur la technique de la factorisation de matrices non négatives (NMF). Deux versions de cet algorithme ont été implantées et comparées à quatre autres algorithmes (co-citation, couplage bibliographique, autorités de HITS et hubs de HITS). Les expérimentations conduites sur deux collections d'articles

scientifiques montrent que notre approche réalise les meilleures performances. Ces travaux ont donné lieu à 4 communications internationales (Chikhi et al., 2008). Courant 2009, l'étude a été approfondie en utilisant une approche probabiliste, car elle permet d'identifier des communautés selon des critères multiples et de faire des classes qui se recouvrent. L'algorithme proposé consiste d'abord à lisser les données (selon des paramètres optimaux calculés en fonction du corpus) pour compenser la rareté des liens, avant d'effectuer un calcul probabiliste (Chikhi et al., 2009).

La suite de ce travail sera de mesurer la couverture des sites des communautés ainsi identifiées par des ontologies de référence, c'est-à-dire de vérifier si la terminologie et les concepts de l'ontologie se retrouvent (ou non) dans les documents représentatifs de ces communautés. Inversement, on évaluera si ces groupes de documents sont de bons candidats pour former un corpus à partir duquel bâtir une ontologie représentative des domaines spécialisés identifiés.

5 REFERENCES

- AUGER A., BARRIERE C., « Pattern based approaches to semantic relation extraction: a state-of-the-art », *Terminology*, Auger A. and Barriere C (Eds.), special issue on “Pattern-based approaches to semantic relation extraction”, Amsterdam/Philadelphia, John Benjamins Publishing Company, 14-1, p. 1-19, 2008.
- AUSSENAC N., How to combine data abstraction and model refinement: a methodological contribution in MACAO. *A future for Knowledge Acquisition, Proc. of EKAW'94, 8th European Knowledge Acquisition Workshop*. Berlin: Springer Verlag. Series Lecture Notes in AI, N°867. 262-282. 1994.
- AUSSENAC-GILLES N. GEDITERM : un logiciel pour gérer des bases de connaissances terminologiques. *Terminologies Nouvelles*, 19 : 111-123. 1999.
- AUSSENAC-GILLES N., TERMINAE, an experimental contribution. *EON2002, workshop on Evaluation of Ontology-based Tools associated to EKAW2002, Sigüenza, Spain. Oct. 2002*. 112-128. <http://www.CEUR-WS.org/Vol-62>. 2002
- AUSSENAC-GILLES N., Supervised Text Analysis for Ontology and Terminology Engineering. *Proceedings of the Dagstuhl Seminar 05071 on “Machine Learning for the Semantic Web”*. Dagstuhl (Germany). 13-18 Feb. 2005.
- AUSSENAC-GILLES N. Ontology or meta-model for retrieving scientific reasoning in documents: the Arkeotek project, Workshop on Exploring the limits of global models for integration and use of historical and scientific information, Héraklion (Greece), 23-24 oct. 2006, M. Doerr, A. Renear (Eds.), 2006. Accès: http://cidoc.ics.forth.gr/workshops/heraklion_october_2006/Aussenac.pdf
- AUSSENAC-GILLES N., Le web sémantique, quel renouvellement pour la recherche d'information ?, *Recherche d'information : état des lieux et perspectives*. M. Boughanem, J. Savoy (Eds.), Paris : [Hermès](#), p 97 - 132, Collection Recherche d'information et web, 2008.
- AUSSENAC-GILLES N., BAZIZ M., HERNANDEZ N., Ontologies pour la recherche d'information : importance de la dimension terminologique. *Terminologie et accès à l'information spécialisée*. W. Mustapha El Hadi (Ed.), [Hermès](#), Techniques et traités des sciences et techniques de l'information, 1-24, 2006.
- AUSSENAC-GILLES N., BIÉBOW B. and SZULMAN S., Revisiting Ontology Design: A method based on corpus analysis, Dieng, R. and O. Corby (eds.), *Knowledge Engineering and Knowledge Management: Methods, models and tools*. Lecture Notes in Artificial Intelligence 1937. 172–188. Berlin: Springer Verlag. 2000.
- AUSSENAC-GILLES N., BOURIGAULT D., The Th(IC)2 Initiative : Corpus-Based Thesaurus Construction for Indexing WWW Documents. in *Proc. of the EKAW'2000 workshop « Ontologies and texts »*. Juan-Les-Pins (F). Oct. 2, 2000. 71-78 <http://www.ceur-ws.org/Vol-51/>
- AUSSENAC-GILLES N., BOURIGAULT D., CONDAMINES A., GROS C.. How can knowledge acquisition benefit from terminology? *Proceedings of the 9th Knowledge Acquisition Workshop. Banff, Univ. of Calgary (CA). Feb. 1995*. H-1/H-14. 1995.
- AUSSENAC-GILLES N., BOURIGAULT D., TEULIER R., Analyse comparative de corpus : cas de l'ingénierie des connaissances. *Actes des 14^e journées Francophones d'Ingénierie des Connaissances (IC2003)*. R. Dieng-Kuntz (Ed.). Laval (F), 1-3 Juillet 2003. Presses Universitaires de Grenoble. 67-84. 2003.
- 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. Spiliopoulou (Eds.), *proceedings of the ECCAI workshop OntoLex08 - From Text to Knowledge: The Lexicon/Ontology Interface*, p. 1-6, 2008.
- AUSSENAC-GILLES N., CONDAMINES A., Bases de connaissances Terminologiques: enjeux pour la consultation documentaire. In *Actes des 1ères journées du Chapitre Français de l'ISKO : Organisation des connaissances en vue de leur intégration dans les systèmes de représentation et de recherche d'information, oct. 1997*, J. Maniez et W. Mustapha El Hadi (eds.), Villeneuve d'Asq : Presses de l'Université Charles de Gaulle (UL3 travaux et Recherches), 71-88, 1997
- AUSSENAC-GILLES N., CONDAMINES A., Entre textes et ontologies formelles : les bases de connaissances terminologiques. *Ingénierie et capitalisation des connaissances*. (Eds.) M. Zacklad, M. Grundstein, Paris : Hermès, Traité IC2, 153-177, 2001.
- AUSSENAC-GILLES N., CONDAMINES A. Documents électroniques et constitution de ressources terminologiques ou ontologiques. *Revue Information, Interaction, Intelligence I3*. Numéro spécial *Document numérique*. Ed. Charlet J. et Salaün J.-M. 4 (1):75-94. 2004.
- AUSSENAC-GILLES N., CONDAMINES A., SEDES F., Evolution et maintenance des ressources termino-ontologiques : une question à approfondir. Hors-série de la revue *Information - Interaction - Intelligence (I3)*, [Cépaduès Editions](#), 7-14, 2006.

- AUSSENAC-GILLES N., CONDAMINES A., Corpus et terminologie. *La redocumentarisation du monde*. (Ed.) R.T. Pédauque. Toulouse : Cépaduès Editions. 131-147, 2007.
- AUSSENAC-GILLES N., CONDAMINES A., Variations syntaxiques et contextuelles dans la mise au point de patrons de relations conceptuelles, *Filtrage sémantique dans les textes : Approches symboliques*. J.-L. Minel (Ed.), Paris : [Hermès](#) Sciences Publications, 4, 109 - 149, Collection Recherche d'information et web, 2009.
- AUSSENAC-GILLES N., A. CONDAMINES, SZULMAN S., Prise en compte de l'application dans la constitution de produits terminologiques. *Actes des 2^e Assises Nationales du GDR I3, Nancy (F), Déc. 2002*. Toulouse : Cépaduès Editions. 289-302. 2002.
- AUSSENAC-GILLES N., DESPRES S., SZULMAN S., The TERMINAE Method and Platform for Ontology Engineering from texts, *Bridging the Gap between Text and Knowledge - Selected Contributions to Ontology Learning and Population from Text*. P. Buitelaar, P. Cimiano (Eds.), [IOS Press](#), p. 199-223, 2008.
- AUSSENAC N., DIENG R., Models of Problem Solving for Knowledge Acquisition: comparison of MACAO and 3DKAT, *Proceedings of the 5th EKAW 91, SISYPHUS Project, part II, Crieff (Scotland), may 1991*.
- AUSSENAC-GILLES N., HERNANDEZ N., Du linguistique au conceptuel : étapes de l'identification de relations conceptuelles à partir de textes. *Atelier "Acquisition et modélisation de relations sémantiques" associé à la conférence TIA 2009, Toulouse, 20/11/2009*, S. Despres, N. Grabar (Eds.).
- AUSSENAC-GILLES N., JACQUES M.-P., Designing and Evaluating patterns for Ontology Enrichment from texts, Staab S., Svatek, V. (Eds.), *International Conference on Knowledge Engineering and Knowledge Management EKAW 2006, Prague, oct. 2006*, Lecture Notes in Artificial Intelligence 4248, 158-165. 2006. [Springer-Verlag](#), Access: <ftp://ftp.irit.fr/IRIT/CSC/EKAW2006defin-LNCS4248-0158.pdf>
- AUSSENAC-GILLES N., JACQUES M.-P., Designing and Evaluating Patterns for Relation Acquisition from Texts with CAMÉLÉON, Auger A. and Barriere C. (Eds.), *Terminology 14-1, Pattern-based approaches to semantic relation extraction*, Amsterdam/Philadelphia, John Benjamins Publishing Company, 14-1, p. 45-73, 2008.
- AUSSENAC-GILLES, KAMEL M., HERNANDEZ N., Towards a platform for supervised relation extraction from text. *Interdisciplinary Laboratory on Interactive Knowledge Systems (ILIKS) 2008 annual meeting, Toulouse (F), 01-02/12/2008*, Laure Vieu (Eds.).
- AUSSENAC-GILLES N., MATTA N., Making a method of problem solving explicit with MACAO, *International Journal of Human-Computer Studies*, New York: Academic Press. 40 : 193-219. 1994.
- AUSSENAC-GILLES N., MATTA N., Expliciter une méthode de résolution de problèmes avec MACAO : problèmes méthodologiques. *L'acquisition des connaissances : tendances actuelles*, Toulouse : (Eds.) N. Aussenac-Gilles, P. Laublet, C. Reynaud. Toulouse : Cépaduès. 1996. 29-48. 1996.
- AUSSENAC-GILLES N., ROUX V., BLASCO P., The Arkeotek project: structuring scientific reasoning and documents to manage scientific knowledge, *ICSH 2006 : Atelier international sur l'Indexation des Connaissances en Sciences Humaines, Nantes (F), 26-27 juin 2006*, S. Calabretto, N. Aussenac-Gilles (Eds.), [Université de Nantes](#), Actes de la Semaine de la Connaissance, V.3, (CD/ROM), 2006. Accès: <ftp://ftp.irit.fr/IRIT/CSC/Aussenac-ICSH2006.doc>
- AUSSENAC-GILLES N., SEQUELA P., Les relations sémantiques : du linguistique au formel. *Cahiers de grammaire*, Numéro spécial sur la linguistique de corpus. A. Condamines (Ed.). Toulouse : Presse de l'UTM. 25 175-198. 2000.
- AUSSENAC-GILLES, N. and SOERGEL D., 2005. Text Analysis for Ontology and Terminology Engineering, *Applied Ontology*. 1(1): 35-46.
- AUSSENAC N., SOUBIE J-L., FRONTIN J., A knowledge Acquisition Tool for Expertise Transfer, *Proceedings of EKAW'88, GMD Studien Nr 143, 1988*, pp 8.1-8.12.
- AUSSENAC N., SOUBIE J-L, FRONTIN J., RIVIÈRE M-H., A mediating representation to assist knowledge acquisition, *Proceedings of the 3rd European Knowledge Acquisition Workshop (EKAW 89)*, Paris (F), July 1989. 516-529.
- AUSSENAC N., SOUBIE J-L, FRONTIN J., RIVIÈRE M-H., A mediating representation to assist knowledge acquisition, *Proceedings of the 4th Knowledge Acquisition Workshop (KAW 89)*, Calgary (CA), Oct 1989.
- AUSSENAC N., SOUBIE J-L., Place d'un outil d'acquisition de connaissances dans la conception des systèmes intelligents, *Actes des Journées d'Acquisition des Connaissances JAC'90*, Lannion (F). 115-129. 1990.
- AUSSENAC-GILLES N., KAMEL M., Ontology Learning by Analyzing XML Document Structure and Content (short paper). *International Conference on Knowledge Engineering and Ontology Development (KEOD 2009), Madeira (Portugal), 06-08/10/2009*, J. Dietz (Eds.), [INSTICC - Institute for Systems and Technologies of Information, Control and Communication](#), p. 159-165, 2009.
- BAZIZ M., AUSSENAC-GILLES N., BOUGHANEM M., Exploitation des Liens Sémantiques pour l'Expansion de Requêtes dans un Système de Recherche d'Information. *XXI^e Congrès INFORSID 2003*. Nancy, 3-6 Juin 2003. INFORSID. Inforsid, 20 rue Axel Duboul - 31000 Toulouse. 124-131. 2003

- BAZIZ M., BOUGHANEM M., AUSSENAC-GILLES N., Semantic representation of Documents by Ontology-Document Mapping. In Proceedings of the 2nd ACM SIGIR Workshop on *Semantic Web and Information Retrieval (SWIR 2004)*. Sheffield (UK), July 25-29th. 2004.
- BAZIZ M., BOUGHANEM M., AUSSENAC-GILLES N., IRIT at CLEF 2004: The English GIRT task, *Cross Language Evaluation Forum CLEF'2004 Workshop, Bath, UK, 15-17 sept. 2004.*, C. Peters, P. Clough, J. Gonzalo, G. J. F. Jones (Eds.), LNCS 3491. Springer-Verlag., 283-291, 2005.
- BAZIZ, M. BOUGHANEM M., AUSSENAC-GILLES N., Semantic Networks for a Conceptual Indexing of Documents in IR. *ISPS'2005 Seventh International Symposium on Programming and Systems*, Algiers, Algeria. 9-11 mai 2005.
- 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, *Proceedings of COLIS 2005 (5th International Conference on Conception 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
- BAZIZ M., BOUGHANEM M., AUSSENAC-GILLES N., Evaluating a Conceptual Indexing Method by Utilizing WordNet, *Accessing Multilingual Information Repositories: 6th Workshop of the Cross-Language Evaluation Forum, CLEF 2005, Revised Selected Papers, Vienna, Austria, 21-23 sept. 2005*, C. Peters, F. C. Gey, J. Gonzalo, G. J.F. Jones (Eds.), Lecture Notes in Computer Science, Vol. 4022, 2005.
- BAZIZ M., BOUGHANEM M., AUSSENAC-GILLES N., A Conceptual Indexing Approach for the TREC Robust Task, *The 14th Text REtrieval Conference Proceedings (TREC 2005)*, Gaithersburg, Maryland, 15-18 nov.2005, E. M. Voorhees , Lori P. Buckland (Eds.), NIST, 2005.
- BEAUBEAU D., AUSSENAC-GILLES N., TCHOUNIKINE P., *Mona au pays des rôles : opérationnalisation de modèles conceptuels MONA en ZOLA*. Rapport Interne 96-23-R, IRIT, Juillet 1996.
- BOURIGAULT, D., Upery : un outil d'analyse distributionnelle étendue pour la construction d'ontologies à partir de corpus, *Actes de Traitement Automatique des Langues Naturelles (TALN 2002)*. 75–84, Nancy France. 2002.
- BOURIGAULT D., AUSSENAC-GILLES N., CHARLET J. Construction de ressources terminologiques ou ontologiques à partir de textes : un cadre unificateur pour trois études de cas. *Revue d'Intelligence Artificielle (RIA)*. Numéro spécial *Techniques Informatiques et Structuration de Terminologies*. Pierrel J.M. et Slodzian M. (Ed.). Paris : Hermès. 18 (1) : 87–110. 2004.
- BREUKER J., VAN DE VELDE W., The CommonKADS library for expertise modelling: reusable problem solving components, Amsterdam : IOS Press. 1994.
- BUITELAAR P., CIMIANO P., MAGNINI B., *Ontology Learning From Text: Methods, Evaluation and Applications*, IOS Press, 2005.
- CHARLET J., BACHIMONT B., BOUAUD J., ZWEIGENBAUM P., Ontologie et réutilisabilité : expérience et discussion, N. Aussenac-Gilles, P. Laublet, C. Reynaud (eds) : *Acquisition et Ingénierie des Connaissances*, Toulouse : Cépaduès-Editions, p.69-88, 1996.
- CHARLET J., *L'ingénierie des connaissances : développements, résultats et perspectives pour la gestion des connaissances médicales*. Mémoire d'habilitation à diriger des recherches en Informatique de l'université de Pierre et Marie Curie. Décembre 2002.
- CHARLET J., AUSSENAC-GILLES N., PIERRA G., NADAH N., SZULMAN S., TEGUIAK H.V., DAFOE : Une plateforme multi-méthodes et multi-modèles pour le développement d'ontologies de domaine, *Journées Francophones sur les Ontologies (JFO 2008)*, Lyon (F), 1-3 déc. 2008, D. Benslimane, C. Roche, S. Spaccapietra (Eds.), [ACM](#), 1-12, 2008.
- CHARLET J., SZULMAN S., AUSSENAC-GILLES N., NAZARENKO A., HERNANDEZ N., NADA N., SARDET E., DELAHOUSSE J., PIERRA G., Apport des outils de TAL à la construction d'ontologies : propositions au sein de la plateforme DaFOE (poster). *Journées Francophones d'Ingénierie des Connaissances (IC 2009)*, Hammamet (Tunisie), 25-29/05/2009, F. Gandon (Eds.), [INRIA](#), (en ligne), 2009a.
- CHARLET J., SZULMAN S., AUSSENAC-GILLES N., NAZARENKO A., HERNANDEZ N., NADA N., SARDET E., DELAHOUSSE J., PIERRA G., Apport des outils de TAL à la construction d'ontologies : propositions au sein de la plateforme DAFOE (démonstration). *Traitement Automatique des Langues Naturelles (TALN 2009)*, Senlys (France), 24-26/06/2009, Tome 2, A. Nazarenko, T. Poibeau (Eds.), [LIPN - Université Paris 13](#), p. 461-463, 2009b.
- CHARLET J., SZULMAN S., AUSSENAC-GILLES N., NAZARENKO A., HERNANDEZ N., NADA N., SARDET E., DELAHOUSSE J., TEGUIAK V., Dafoe : an Ontology Building Platform From Texts or Thesauri (poster). *Conférence Internationale sur la Terminologie et l'Intelligence Artificielle (TIA 2009)*, Toulouse (France), 18-19/11/2009, M.-C. L'Homme, S. Szulman (Eds.), [Université Paul Sabatier - Toulouse](#), (support électronique), 2009c.

- CHARLET J., SZULMAN S., AUSSENAC-GILLES N., NAZARENKO A., HERNANDEZ N., NADA N., SARDET E., DELAHOUSSE J., TEGUIAK V., BANEYX A., DAFOE : une plateforme pour construire des ontologies à partir de textes et de thésaurus (démonstration). *Journées Francophones Extraction et Gestion de Connaissances (EGC 2010), Hammamet (Tunisie), 26-29-JAN-10*, J.-M. Petit, M. Zaki (Eds.), [Hermès Science Publications](#), p. 1-3, 2010 (à paraître).
- CHIKHI, ROTHENBURGER B., AUSSENAC-GILLES N., A Comparison of Dimensionality reduction techniques for Web Structure N. F. Mining, *IEEE/WIC/ACM INTERNATIONAL Conference on Web Intelligence, Silicon Valley (USA), 2-5 nov. 2007*, Li Tsau Young (T.Y.) (Eds.), [IEEE Computer Society](#), 116-119, 2007.
- CHIKHI N. F., ROTHENBURGER B., AUSSENAC-GILLES N., A Nonnegative Factor Model for Authoritative Documents Identification. *IEEE International Conference on Information Reuse and Integration (IRI 2008), Las Vegas (USA), 13-15 july. 2008*, R. Alhajj and K. Zhang (Eds.), [IEEE](#), 262-267, 2008.
- CHIKHI N. F., ROTHENBURGER B., AUSSENAC-GILLES N., A New Algorithm for Community Identification in Linked Data. *International Conference on Knowledge-Based and Intelligent Information & Engineering Systems (KES 2008), Zagreb (Croatia), 3-5 sept. 2008*, [Springer-Verlag](#), 2008.
- CHIKHI N. F., ROTHENBURGER B., AUSSENAC-GILLES N., Combining Link and Content Information for Scientific Topics Discovery, *IEEE International Conference on Tools with Artificial Intelligence (ICTAI 2008), Dayton, Ohio (USA), 3-5 nov. 2008*, [IEEE Computer Society](#), 211-214, 2008.
- CHIKHI N. F., ROTHENBURGER B., AUSSENAC-GILLES N., Community Structure Identification: A Probabilistic Approach (regular paper). *International Conference on Machine Learning and Applications, Miami (USA), 13/12/2009-15/12/2009*, [IEEE](#), 2009. (Best paper award)
- CIMIANO P., *Ontology Learning and Population from Text. Algorithms, Evaluation and Applications*. Berlin: Springer. 2007.
- CONDAMINES A., L'interprétation en sémantique de corpus : le cas de la construction de terminologies, *Revue Française de Linguistique Appliquée, Corpus : état des lieux et perspectives. Vol.XII-1*. p. 39-52, 2007.
- CONDAMINES A., Taking *genre* into account for analyzing conceptual relation patterns, à paraître dans *Corpora*. 2009.
- FOURNIER D., 1998, *Étude et conception d'un système de gestion d'une Base de Connaissances Terminologiques*, Mémoire d'ingénieur CNAM, Toulouse. Mars 1998.
- GARCIA D., N. AUSSENAC-GILLES, AND A. COURCELLE. Exploitation, pour la modélisation, des connaissances causales détectées par COATIS dans les textes. In *Journées Françaises d'Acquisition des Connaissances JAC'96*, mai 1996. LIRMM, Université de Montpellier. 123-136. 1996.
- GARCIA D., AUSSENAC-GILLES N., Exploitation, pour la modélisation, des connaissances causales détectées par COATIS dans les textes. *Ingénierie des connaissances*, Eds. D. Bourigault, J. Charlet, G. Kassel, M. Zacklad. Paris : Eyrolles. janvier 2000 : 257-274.
- GROS C., H. ASSADI, N. AUSSENAC-GILLES, AND A. COURCELLE. Task Models for Technical Documentation Accessing. In *European Knowledge Acquisition Workshop, EKAW'96 Poster session, Complement to the proceedings*, mai 1996. University of Nottingham, Nottingham, GB.
- HANDSCHUH S., STAAB S. (Eds.), *Annotation for the Semantic Web*, Volume 96 "Frontiers in Artificial Intelligence and Applications", IOS Press, 2003.
- HEARST, M.A., Automatic acquisition of hyponyms from large text corpora. *Proceedings of the 14th conference on Computational linguistics*, Morristown, NJ, USA, Association for Computational Linguistics, p. 539-545, 1992.
- HERNANDEZ N. *Ontologies pour l'aide à une activité de veille ou d'exploration d'un domaine*. Université Toulouse 3, spécialité Informatique, école doctorale EDIT, déc. 2005. Directeur de thèse : J. Mothe.
- 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, 10/07/2007-12/07/2007*, [Cépaduès](#), 2007.
- KAMEL M., AUSSENAC-GILLES N., Construction automatique d'ontologies à partir de spécifications de bases de données. *Journées Francophones d'Ingénierie des Connaissances (IC 2009), Hammamet (Tunisie), 25-29/05/2009*, F. Gandon (Eds.), [Univ. Hassan II](#), p. 85-96, 2009.
- KAMEL M., AUSSENAC-GILLES N., How can document structure improve ontology learning? (regular paper). *Semantic Authoring, Annotation and Knowledge Markup Workshop - collocated with K-CAP 2009 (SAAKM 2009), Redondo Beach, California (USA), 01/09/2009*, S. Handschuh, M. Sintek (Eds.), [CEUR Workshop Proceedings](#), p. 1-8, 2009.
- KERGOSIEN E., KAMEL M., SALLABERRY C., BESSAGNET M.-N., AUSSENAC-GILLES N., GAI0 M., Construction et enrichissement automatique d'ontologies à partir de ressources externes (regular paper). *Journées Francophones sur*

- les Ontologies (JFO 2009), Poitiers (France), 03-04/12/2009*, Bellatreche L., Kassel G., Thiran P. (Eds.), [ACM SIGAPPFR](#), p. 11-20, 2009.
- LECORGNE E., Étude et conception d'une maquette de consultation de base de connaissances terminologiques. Rapport de stage d'ingénieur CNAM, Toulouse. Déc. 1998
- LEPINE P., AUSSENAC-GILLES N., Modélisation de la résolution de problèmes : comparaison expérimentale de KADS et MACAO. *L'acquisition des connaissances : tendances actuelles* (Eds.) N. Aussenac-Gilles, P. Laublet, C. Reynaud. Toulouse : Cépaduès, 131-150, 1996.
- MAEDCHE A., *Ontology learning for the Semantic Web*, volume 665. Kluwer Academic Publisher, 2002.
- MEYER I., Extracting Knowledge-rich Contexts for Terminography : A Conceptual and methodological Framework, D. Bourigault, M.C. L'homme, C.Jacquemin (eds) : *Recent Advances in Computational Terminology*, John Benjamins. p. 279-302, 2001.
- S. MUSTIERE, N. ABADIE, N. AUSSENAC-GILLES, M.-N. BESSAGNET, M. KAMEL, E. KERGOSIEN, C. REYNAUD, B SAFAR. GéOnto : Enrichissement d'une taxonomie de concepts topographiques (regular paper). *International Conference on Spatial Analysis and Geomatic (SAGEO 2009), Paris (France), 25-27/11/2009*, M. Gaio, D. Josselin (Eds.), [Hermès](#), p. 1-17, 2009.
- OTTENS K., AUSSENAC-GILLES N. Un algorithme multi-agent de classification pour la construction d'ontologies dynamiques, *7e Journées Francophone Extraction et Gestion des Connaissances (EGC 2007), Namur (B), 25-27 janvier 2007*, G. Venturini (Eds.), RNTI, Vol E-9, Cépaduès-Éditions, 647-658, 2007.
- OTTENS K., AUSSENAC-GILLES N., GLEIZES M.P., CAMPS V., Dynamic Ontology Co-Evolution from Texts: Principles and case Study. *International Workshop on Emergent Semantics and Ontology Evolution at ISWC 2007 (ESOE 2007), Busan (South Korea), 12/11/2007*, L. Chen, P. Haase, A. Hotho, E. Ong (Eds.), 2007. <http://km.aifb.uni-karlsruhe.de/ws/esoe2007/>
- OTTENS K, HERNANDEZ N., GLEIZES M.-P., AUSSENAC-GILLES N., A Multi-Agent System for Dynamic Ontologies *Journal of Logic and Computation*, [Oxford University Press](#), *Special Issue on Ontology Dynamics*, Vol. 19 : 1-28, 2008.
- REYMONET A., AUSSENAC-GILLES N., THOMAS J., Tâche, Domaine et Application : Influences sur le processus de modélisation de connaissances, *17èmes Journées Francophones d'Ingénierie des Connaissances (IC 2006). Nantes (F), 28-30 Juin 2006*, [Univ. de Nantes](#), Semaine de la Connaissance, M. Lewkowicz (Ed.), Vol.1, 71-80, 2006.
- 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.
- REYMONET A., THOMAS J., AUSSENAC-GILLES N.. Ontology Based Information Retrieval: an application to automotive diagnosis. *International Workshop on Principles of Diagnosis (DX 2009), Stockholm, 14-17/06/2009* (conférencier invite : A. Reymonet), M. Nyberg, E. Frisk, M. Krisander, J. Aslund (Eds.), [Linköping University. Institut of Technology](#), p. 9-14, 2009.
- REYNAUD C., N. AUSSENAC-GILLES, P. TCHOUNIKINE, AND F. TRICHET. The notion of role in conceptual modelling. *In Proceedings of EKAW97 - European Knowledge Acquisition Workshop - R. Benjamins & E. Plaza Eds. Lecture Notes in Artificial Intelligence*. Springer Verlag, Heidelberg, 221-236, 1997.
- REYNAUD C., N. AUSSENAC-GILLES, F. TORT. A support to domain knowledge modelling: a case study. In H. Kangassalo and P.J Charrel (eds), *Information Modelling and Knowledge Bases IX*, vol 45 of *Frontiers in AI and Applications*, 35-50. IOS Press, Amsterdam, 1998.
- 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.
- SCHREIBER G A.Th., WIELINGA B.J., (eds.); *KADS, a principled approach to knowledge based system development*. London: Academic Press. 1993
- SEGUELA P., AUSSENAC-GILLES N., Un modèle de base de connaissances terminologiques. *2e rencontres 'Terminologie et Intelligence Artificielle' TIA'97*, avril 1997. Equipe de Recherche en Syntaxe et Sémantique (ERSS), Université Toulouse-Le Mirail. 71-98

- SEGUELA P., AUSSENAC-GILLES N., Extraction de relations sémantiques entre termes et enrichissement de modèles du domaine. *Actes de la conférence d'Ingénierie des Connaissances IC'99 - Plate-forme AFIA, Palaiseau (F), Juin 1999*. 79-88. 1999.
- SEGUELA P., Construction de modèles de connaissances par analyse linguistique de relations lexicales dans les documents techniques. Thèse de doctorat en Informatique, Université Toulouse III, mars 2001.
- SELLAMI Z., GLEIZES M.-P., AUSSENAC-GILLES N., ROUGEMAILLE S., Dynamic ontology co-construction based on adaptive multi-agent technology (regular paper). *International Conference on Knowledge Engineering and Ontology Development (KEOD 2009), Madeira (Portugal), 06-08/10/2009*, Dietz J. (Eds.), [INSTICC - Institute for Systems and Technologies of Information, Control and Communication](#), p. 56-63, 2009.
- SELLAMI Z., AUSSENAC-GILLES N., GLEIZES M.-P., ROUGEMAILLE S., MBARKI M.. Vers un outil de co-construction d'ontologies à partir de textes à l'aide d'un système multi-agent adaptatif (regular paper). *Journées Francophones sur les Ontologies (JFO 2009), Poitiers, France, 03-04/12/2009*, Bellatreche L., Kassel G., Thiran P. (Eds.), [ACM](#), p. 3-10, 2009.
- Z. SELLAMI, M.-P. GLEIZES, N. AUSSENAC-GILLES, S. ROUGEMAILLE. Dynamic ontology co-construction based on adaptive multi-agent technology. *European Workshop on Multi-Agent Systems (EUMAS 2009), Ayia Napa, Cyprus, 17-18/12/2009*.
- 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.
- SZULMAN S., BIEBOW B., AUSSENAC-GILLES N., Vers un environnement intégré pour la structuration de terminologies : TERMINAE. *Journée de l'ATALA*. Paris, Mars 2001.
- SZULMAN, S., BIEBOW B., AUSSENAC-GILLES N., Structuration de Terminologies à l'aide d'outils d'analyse de textes avec TERMINAE. *Traitement Automatique de la Langue (TAL)*. Numéro spécial « Structuration de Terminologie ». Eds A. Nazarenko, T. Hammon. Hermès : Paris. 43 (1) : 103-128. 2002.
- SZULMAN, S., BIEBOW B., OWL et TERMINAE, *actes de la 15^e conférence d'Ingénierie des Connaissances IC 2004, Lyon (F)*, mai 2004, Presses Universitaires de Grenoble, 2004, 41-52.
- S. SZULMAN, J CHARLET, N. AUSSENAC-GILLES, A. NAZARENKO, E. SARDET, H.V. TEGUIAK. DAFOE: an Ontology Building Platform From Text or Thesauri (poster). *International Conference on Knowledge Engineering and Ontology Development (KEOD 2009), Madeira (Portugal), 06-08/10/2009*, J. Dietz (Eds.), [INSTICC - Institute for Systems and Technologies of Information, Control and Communication](#), p. 1-4. 2009.
- TISSAOUI A., Typologie de changements et leurs effets sur l'évolution de Ressources Termino-Ontologiques (poster). *IC 2009 : Posters des 20es Journées Francophones d'Ingénierie des Connaissances, Hammamet (Tunisie), 25/05/2009-29/05/2009*, Fabien Gandon (Eds.).
- VÖLKER J., HITZLER P., CIMIANO P.: Acquisition of OWL DL Axioms from Lexical Resources. *ESWC 2007*: 670-685. 2007.