INRIA

INSTITUT NATIONAL DE RECHERCHE EN INFORMATIQUE ET EN AUTOMATIQUE

Équipe orpailleur

Systèmes de connaissances et extraction de connaissances dans les bases de donneées

Lorraine

THÈME 3A

Table des matières

1. Composition de l'équipe **1**
2. Présentation et objectifs généraux **1**
3. Fondements scientifiques **2**
3.1. L'extraction de connaissances dans les bases de données 2
3.1.1. ECBD symbolique 3
3.1.2. La classification par treillis et la recherche de motifs fréquents 3
3.1.3. La fouille de données avec des modèles de Markov 3
3.1.3.1. La reconnaissance de successions culturales 5
3.1.4. La fouille de textes 5
3.1.4.1. Une approche de la fouille de textes 5
3.1.4.2. Indexation conceptuelle de textes 6
3.1.4.3. La construction de la synthèse d'un ensemble de textes 6
3.1.4.4. Hybridation de méthodes de classification 6
3.1.5. Les aspects bioinformatiques et la fouille de données en biologie 7
3.1.6. La fouille de bases de données en chimie organique 8
3.2. La représentation et la gestion des connaissances 8
3.2.1. La classification, les systèmes de RCO et les logiques de descriptions 8
3.2.2. La gestion des connaissances 9
3.2.3. Systèmes à bases de connaissances et raisonnement spatial qualitatif 10
3.3. Le Web sémantique et les systèmes intelligents de traitement de l'information 11
3.3.1. La problématique liée au Web sémantique 11
3.3.2. L'accès intelligent à l'information sur le Web 11
3.3.3. Les problèmes de représentation et de manipulation de documents 12
3.3.3.1. Objets et Web sémantique 12
3.3.3.2. La manipulation de documents en fonction de leur contenu 12
5. Logiciels **12**
5.1. Les modèles de Markov pour l'ECBD numérique : CarottAge 12
5.2. Les logiciels pour la fouille de textes 13
5.3. Les logiciels pour l'analyse et la simulation d'organisations spatiales agricoles 13
5.4. Le système KASIMIR 14
5.5. Les systèmes RÉSYN et RÉSYN-ASSISTANT 14
5.6. Le traitement intelligent de l'information et le Web sémantique 14
5.7. DefineCrawler : un crawler paramétrable pour la recherche intelligente sur le Web 15
8. Actions régionales, nationales et internationales **15**
8.1. Actions locales 16
8.1.1. La collaboration URI et Orpailleur 16
8.1.2. La collaboration READ et Orpailleur 16
8.2. Actions nationales 16
8.2.1. L'ARC INRIA GENI : génération et inférence 16
8.2.2. Une collaboration avec l'INRA 17
8.2.2.1. Modélisation de dispersion de transgènes. 17
8.2.3. Une collaboration avec le Cemagref 17
8.2.4. Le projet KASIMIR 18
8.2.5. Le projet KVM 19
8.2.6. Une collaboration sur le thème du RàPC (Université de Lyon 1) 19
8.2.7. Les travaux de recherches post-GDR CNRS 1093 TICCO 19
9. Diffusion des résultats **20**

9.1. Animation de la Communauté scientifique 20
9.2. Enseignement 20
10. Bibliographie **20**

1. Composition de l'équipe

Responsable scientifique
Amedeo Napoli [CR CNRS]

Responsables permanents
Florence Le Ber [CR INRIA - détachement]
Jean Lieber [MdC, Université Henri Poincaré - UHP Nancy 1]
Jean-François Mari [Professeur, Université de Nancy II]
Emmanuel Nauer [MdC, Université de Metz]
Yannick Toussaint [CR INRIA]

Assistante de projet
Antoinette Courrier [Technicienne CNRS]

Chercheurs doctorants
Rim Al Hulou [doctorante, ATER Université de Nancy 2]
Sandra Berasaluce [doctorante avec co-encadrement, bourse MENRT]
Martine Cadot [doctorante, Professeure certifiée sur poste PRAG, Université Henri Poincaré - UHP Nancy 1]
Fairouz Chakkour [doctorante, ATER Université Henri Poincaré - UHP Nancy 1]
Hacène Cherfi [doctorant, ATER Université Henri Poincaré - UHP Nancy 1]
Mathieu D'Aquin [doctorant, bourse MENRT]
Sébastien Hergalant [Doctorant, bourse co-financée INRA-Région]
Sandy Maumus [doctorante avec co-encadrement, bourse co-financée INSERM-Région]
Jean-Luc Metzger [doctorant, bourse co-financée INRA - INRIA]
Laszlo Szathmary [doctorant, bourse financée sur projet ANVAR KVM]

Ingénieur expert
Sébastien Brachais [Ingénieur associé INRIA]

Collaborateur extérieur permanent
Benoît Bresson [Collaboration Orpailleur-CAV Nancy/Oncolor]

2. Présentation et objectifs généraux

L'orpailleur est l'artisan qui recueille par lavage - à travers un tamis - les paillettes d'or dans les fleuves et les terres aurifères. L'or, dans le cadre de la conception de systèmes à bases de connaissances (SBC dans la suite), correspond à la connaissance. Cette connaissance est de plusieurs types et a plusieurs origines : elle peut reposer sur de l'expertise, des expériences, des explications, des stratégies et des façons de faire. Elle peut être donnée de façon explicite - par des spécialistes - ou exister de manière implicite - dans des bases de données de toutes natures. Pour être opérationnelle, cette connaissance doit de plus être représentée dans des formalismes adéquats pour être ensuite manipulée par des procédures de raisonnement.

L'objectif premier du projet Orpailleur est de concevoir des systèmes intelligents pour résoudre des problèmes en exploitant les données et les connaissances relatives à un domaine d'application donné comme l'agronomie, la biologie, la chimie, la médecine, la sidérurgie, ou encore, de façon plus générale, pour traiter l'information scientifique, technique et culturelle, et mettre en œuvre une ingénierie des langues et des documents.

Les travaux de recherche du projet Orpailleur s'articulent autour de trois axes majeurs : l'extraction de connaissances dans les bases de données, la gestion et la représentation de connaissances, et enfin le Web sémantique. Ces axes sont interdépendants et peuvent se schématiser par une boucle qui fait passer « des données aux connaissances », comme l'illustre la figure 1. L'élément commun partagé par ces trois axes est la *classification* qui intervient à tous les niveaux le long de ces trois axes. Au départ, les données brutes sont hétérogènes et semi-structurées car provenant de sources diverses : bases de documents textuels, séquences

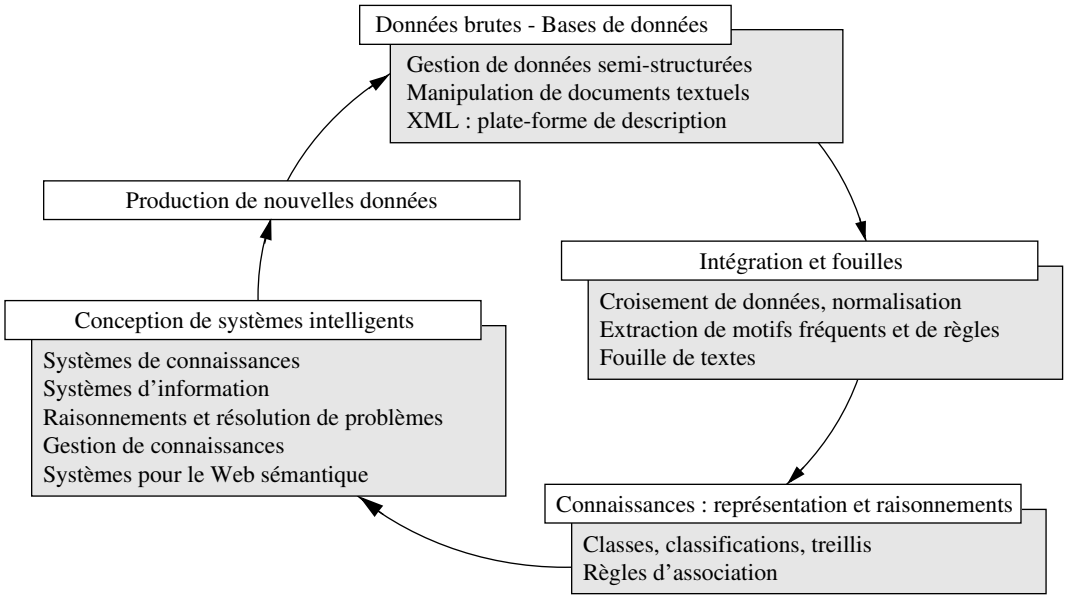

Figure 1. Des données aux connaissances ou l'articulation des recherches dans Orpailleur.

génomiques, dossiers médicaux, bases de réactions chimiques, données de terrain, ... En particulier, pour les documents textuels, le langage XML sert de plate-forme de description intermédiaire entre des données semi-structurées et les éléments de connaissances recherchés. Une fois les données intégrées - filtrées et transformées dans un format adéquat pour être traitées - des processus de fouille de données peuvent être appliqués pour faire émerger des éléments de connaissances potentiellement exploitables. Les processus de fouille s'appuient principalement sur la classification par treillis - avec recherche de motifs fréquents et extraction de règles d'association - et la classification par modèles de Markov cachés, tout en exploitant un modèle du domaine des données. Ce modèle - encore appelé *ontologie* - est un des composants de base d'un système intelligent ou système de connaissances Les éléments de connaissances extraits peuvent venir compléter les connaissances du système intelligent considéré, et être à leur tour manipulés pour résoudre de nouveaux problèmes, participer à la gestion des connaissances dans un domaine donné ou pour le Web sémantique. De nouvelles données sont alors produites, qui alimentent des bases de données, qui peuvent à leur tour être fouillées.

Le processus de transformation entre données et connaissances repose sur la *classification*, qui intervient à tous les niveaux : pour modéliser un domaine, le comprendre, le représenter - sous la forme d'une hiérarchie de concepts - et manipuler les concepts représentés pour résoudre des problèmes. La classification est un outil polymorphe qui a pris une place prépondérante le long des trois axes de référence pour Orpailleur.

Dans la suite, les travaux de recherches du projet Orpailleur sont détaillés, en allant des données aux connaissances, pour finir par la mise en œuvre du Web sémantique.

3. Fondements scientifiques

3.1. L'extraction de connaissances dans les bases de données

Mots clés : *extraction de connaissances dans les bases de données*, *méthodes symboliques pour la fouille de données*, *classification par treillis*, *recherche de motifs fréquents (dans des tableaux de données)*, *extraction de règles*, *modèles de Markov cachés pour la fouille de données*.

3.1.1. ECBD symbolique

L'extraction de connaissances dans des bases de données - abrégée en ECBD - est une activité qui consiste à analyser un ensemble de données brutes de façon à en extraire des connaissances exploitables. C'est un expert du domaine relatif aux données - l'« analyste » - qui est chargé de diriger l'extraction. En fonction de ses objectifs, l'analyste va sélectionner des données et utiliser des outils de *fouille de données* pour construire des modèles expliquant les données. L'analyste peut ensuite sélectionner et exploiter les modèles qui représentent un point de vue satisfaisant. Pour mener à bien son activité, l'analyste met à contribution ses connaissances mais aussi un ensemble d'outils regroupés au sein d'un système d'ECBD. Un système d'ECBD s'articule autour de quatre composantes principales :

- les bases de données et leurs systèmes de gestion,
- un système à base de connaissances pour la gestion des connaissances et la résolution de problèmes sur le domaine relatif aux données,
- un système de fouille de données pouvant s'appuyer sur des techniques symboliques ou numériques comme les classifications par treillis et par arbres de décision, l'induction, l'analyse des données ou les statistiques,
- une interface se chargeant des interactions et de la visualisation des résultats.

Un système d'ECBD vise à traiter des bases de données volumineuses et évolutives, et il peut, pour ce faire, s'appuyer sur des connaissances du domaine lors du processus d'extraction des connaissances. L'ECBD peut être ainsi vue comme le processus alimentant un système à base de connaissances : les connaissances extraites sont stockées dans la base pour être réutilisées dans d'autres applications et mises à jour le cas échéant.

3.1.2. La classification par treillis et la recherche de motifs fréquents

La classification par treillis est une technique de fouille de données symbolique qui permet d'analyser une population, d'extraire des corrélations, des motifs et des règles, selon certains points de vue choisis. Elle relève de l'analyse de « tableaux booléens de données » : présence - absence de propriétés mono-valuées. Elle s'appuie sur la correspondance de Galois associée à une relation pour faire émerger un treillis de concepts formels (hiérarchie de concepts), où un concept est un couple (*intension,extension*) et des règles d'association exactes ou partielles. Ces hiérarchies de concepts particulières se construisent en fonction des connaissances disponibles sur le domaine des données et produisent des structures ordonnées interprétables et réutilisables.

Parallèlement à la classification par treillis, l'extraction de *motifs fréquents* correspond à l'extraction d'ensembles de propriétés dont le nombre d'occurrences dans les individus d'une population étudiée est supérieur à un seuil donné. Ces motifs recouvrent les fermés d'une correspondance de Galois. À partir des motifs extraits, des règles d'association qui expriment des corrélations entre les propriétés qui composent les motifs fréquents peuvent également être extraites. La recherche de motifs fréquents et l'extraction de règles d'association reposent sur la construction du treillis de Galois de la relation « l'objet `o` possède la propriété `p` », qui, à partir d'un tableau booléen, fait émerger un treillis de concepts formels, décrits par des ensembles de propriétés et des ensembles d'individus qui s'y rattachent.

Les règles extraites des motifs fréquents mettent en corrélation des propriétés qui composent les motifs fréquents. À ces règles sont associés des coefficients comme le support (la proportion d'individus qui possèdent le motif) et la confiance (la proportion d'individus qui vérifient la règle). Dans le même ordre d'idées, l'*analyse statistique implicative* consiste à faire émerger des règles auxquelles est associée une pondération d'ordre statistique : « *si a alors presque b* ». C'est là un thème dans lequel s'effectue le travail de thèse de Martine Cadot, qui consiste à étudier en détail et de façon conjointe les formalismes que sont l'analyse statistique implicative, la recherche de motifs fréquents et l'extraction de règles d'association [30][7].

3.1.3. La fouille de données avec des modèles de Markov

Une des originalités d'Orpailleur est de réutiliser certains travaux de classification numérique pour procéder à de la fouille de signaux spatiaux-temporels. Pour cela, nous avons appliqué un processus de classification à différents types de données temporelles ou spatiales, telles que des données issues d'un processus industriel de

laminage de tôles par un train à bande, des données agronomiques correspondant à des relevés d'occupations du sol, ou, plus récemment, des séquences nucléotidiques de bactéries lactiques [24][13][14].

Le point commun entre ces trois traitements est l'utilisation de modèles stochastiques - les *modèles de Markov cachés d'ordre 1 ou 2* abrégés en HMM1 et HMM2 pour *Hidden Markov Models* d'ordre 1 ou 2 - développés initialement pour la reconnaissance de la parole et l'identification du locuteur. Confronté à des données où le classement d'un échantillon - une *brame* laminée, une parcelle agricole ou un groupe nucléotidique - dépend autant de ses caractéristiques propres que de sa place dans la séquence, c'est-à-dire des caractéristiques de ses voisins temporels ou spatiaux, l'expert du domaine appréhende la modélisation stochastique et essaie de segmenter en zones stationnaires puis transitoires. En particulier, les modèles de Markov d'ordre 2 qui nous intéressent principalement ici permettent une meilleure prise en compte des durées des suites d'états stationnaires et transitoires.

La modélisation d'un signal temporel ou spatial par HMM est fondée sur deux principes : le signal peut être découpé en segments par une chaîne de Markov et le signal est la réalisation d'un processus stationnaire représenté par une densité de probabilité sur l'espace des observations à l'intérieur d'un segment. La représentation d'un phénomène temporel ou spatial à l'aide d'un HMM peut avoir plusieurs objectifs :

- L'*estimation* ou apprentissage de paramètres, par exemple pour calculer les probabilités des successions de cultures à la lumière des observations faites dans une région.
- La *discrimination* ou reconnaissance, par exemple pour retrouver, une fois le modèle estimé, la suite des états la plus probable du processus qui explique une suite d'observations.
- La *segmentation*, par exemple pour rechercher une date de changement entre deux états donnés.

Un des grands intérêts des HMM est que des algorithmes polynomiaux existent dans chacun de ces trois domaines.

Contrairement aux algorithmes classiques qui fournissent une réponse exacte, les HMM autorisent un apprentissage automatique et capturent la variabilité inhérente aux processus vivants par le truchement d'algorithmes d'estimation statistiques - exacts, eux, par ailleurs - et par une prise en compte de faibles dégénérescences encore inexpliquées, et donc encore attribuées au hasard. Cette modélisation du hasard permet de mesurer l'incertitude que nous avons d'un phénomène et de mieux appréhender son explication.

La mise en œuvre d'une application dans un domaine donné pour l'activité de fouille des données se fait selon le scénario suivant :

- compréhension et prise en charge - par l'informaticien - du domaine et de ses données,
- modélisation stochastique pour la recherche des zones stationnaires et transitoires,
- premiers résultats de classification et interprétation avec les experts du domaine,
- début d'une suite de nouvelles expériences de classification avec des experts de plus en plus aptes à comprendre et utiliser ces méthodes.

Une étude a été menée dans le domaine de l'agronomie sur la recherche de régularités temporelles et spatiales dans des bases de données sur l'occupation du sol. Dans cette étude, il est montré que la probabilité *a posteriori* des régimes cachés d'un modèle de Markov du second ordre (HMM2) est un signal riche et porteur de sens, et qu'elle permet à un agronome de retrouver, de quantifier et de raisonner sur les successions de cultures les plus probables [24].

Une seconde étude s'est faite sur des données relatives au génome, avec deux objectifs : tester la généralité de la méthode en l'appliquant à la bioinformatique et rechercher des répétitions dans le génome d'une bactérie. Ce travail a débouché naturellement sur la recherche de réitérations dans le génome, qui est menée depuis 2001 par Sébastien Hergalant, d'abord dans le cadre de son DEA de génie bioinformatique puis dans le cadre de sa thèse d'Université (co-encadrée par Orpailleur et le laboratoire de Génétique et Microbiologie, UMR - UHP - INRA). Les HMM2 se sont ainsi avérés être des outils intéressants pour un bioinformaticien dans la localisation des réitérations de nucléotides - en tandem, dispersés ou inversés - dans une bactérie [13][14].

3.1.3.1. La reconnaissance de successions culturales

Nous avons utilisé les algorithmes d'apprentissage à base de HMM1 et HMM2 sur des données spatio-temporelles portant sur l'utilisation du territoire, pour étudier les successions culturales pratiquées dans différentes régions françaises : Lorraine, Bassin de la Seine et Midi-Pyrénées. Les HMM permettent de représenter des observations spatio-temporelles comme des successions d'états où les transitions entre états dépendent de l'état courant et des 1 ou 2 états précédents (suivant l'ordre du modèle du HMM). L'extraction de connaissances à partir de telles données est particulièrement importante pour appréhender différents problèmes liés à l'agriculture - comme la gestion des ressources en eau - et pour la compréhension du métier d'agriculteur et de son évolution. Nous avons travaillé en deux temps, en traitant successivement l'information temporelle et l'information spatiale :

- la segmentation temporelle des données permet de mettre à jour les principales successions culturales pratiquées dans une région donnée et leurs évolutions.
- la segmentation spatio-temporelle des données permet de mettre à jour des sous-régions homogènes pour les successions culturales et leurs évolutions.

Les résultats fournis par les algorithmes ont été évalués en relation étroite avec des experts agronomes de l'INRA. Les connaissances ainsi extraites des bases de données sont ensuite confrontées aux connaissances qualitatives des experts de terrain et utilisées dans différents modèles agronomiques [24].

3.1.4. La fouille de textes

3.1.4.1. Une approche de la fouille de textes

Un processus de fouille de textes doit, à partir d'un texte - ou d'un ensemble de textes - décomposé en groupes syntaxiques cohérents, fournir des éléments de synthèse permettant d'appréhender et de manipuler globalement le ou les textes étudiés. Ces éléments synthétiques peuvent comporter un treillis de concepts, un ensemble d'explications associées à ces concepts sous la forme de règles d'association, et un ensemble d'index, dérivant des concepts et des explications. Les explications peuvent être complétées par une ontologie du domaine en liaison avec des thésaurus.

Les perspectives d'utilisation de la fouille de données sur de grandes collections de textes sont importantes. Les textes expriment un spectre très large d'information mais la forme sous laquelle cette information est encodée, en langue naturelle, la rend difficile à déchiffrer. À l'heure actuelle, il y a peu de véritables travaux en fouille de textes, et en tout cas, fréquemment confusion avec les problèmes d'accès à l'information. Ainsi, les textes sont rarement utilisés pour découvrir de nouvelles informations. Il se trouve notamment de nombreux travaux portant sur la classification de documents et permettant d'avoir des visions thématiques de collections de documents. Pour notre part, nous distinguons deux objectifs principaux dans la fouille de textes :

- Le premier vise à partiellement automatiser la construction de ressources linguistiques pour les outils de Traitement Automatique de la Langue.
- Le second est de chercher dans une grande collection de textes, à faire émerger de nouvelles connaissances sous la forme de corrélations entre des faits ou des événements qui sont décrits dans les textes, grâce à une analyse transversale des textes.

C'est plutôt le second de ces objectifs qui retient notre attention. L'originalité de notre approche réside dans l'association des différentes compétences présentes dans le projet Orpailleur : gestion et représentation des connaissances, analyse linguistique robuste et algorithmes de fouilles de données symboliques. Les perspectives de recherche portent sur la qualité de l'extraction de l'information à partir des textes et sur la formalisation du contenu des textes. De plus, la richesse, mais aussi la complexité, des structures que l'on peut extraire d'un texte, comparées à la nature des données généralement disponibles dans des bases de données, posent de nouveaux défis pour la définition d'algorithmes de fouille de textes.

Ci-après, nous décrivons trois thèmes de recherche sur lesquels nous travaillons actuellement et que nous comptons développer davantage dans le futur proche.

3.1.4.2. Indexation conceptuelle de textes

La fouille de textes passe par une *indexation conceptuelle des textes*, qui consiste à indexer des documents textuels par des structures conceptuelles extraites des textes eux-mêmes. Ce type d'indexation permet de dépasser la simple indexation au niveau du terme ou des mots-clés, pour laquelle nous disposons d'outils satisfaisants. Ces structures conceptuelles sont représentées par des concepts organisés en une hiérarchie, sur lesquels il est possible de faire des calculs de spécialisation ou de généralisation. La complexité des phrases dans les textes traités constitue un premier obstacle à cette phase d'extraction de l'information. Ainsi, nous exploitons les compétences internes au projet Orpailleur, en représentation et en extraction de connaissances, notamment sur les logiques de descriptions et l'extraction de motifs fréquents pour mener à bien ces recherches.

Du point de vue de l'analyse des textes, il paraît difficile de s'appuyer sur une démarche trop proche de la linguistique formelle, car la complexité du vocabulaire et de la structure des phrases est rédhibitoire. En revanche, une méthode mettant en jeu une analyse partielle de la syntaxe et de la sémantique, et exploitant des connaissances du domaine permet d'envisager une solution réaliste et actuelle à la question. Nos travaux partent d'une analyse en constituants des phrases et identifient le rôle syntaxico-sémantique des différents constituants. Notre hypothèse de départ est que, dans un domaine de spécialité pour lequel il est possible de constituer un modèle de connaissances, le fait de disposer de la représentation complète et correcte d'un ensemble de phrases de référence permet de mener à bien l'analyse de nouvelles phrases. Ainsi, l'expérience d'Orpailleur en matière de raisonnement à partir de cas est bien adaptée à cette hypothèse : cette approche du traitement de la langue dans un contexte scientifique et technique est novatrice, et elle constitue le thème de la thèse de Fairouz Chakkour.

3.1.4.3. La construction de la synthèse d'un ensemble de textes

Cette problématique cherche à situer un - ou plusieurs - texte(s) par rapport à un - ou plusieurs - autre(s) texte(s), et expliquer en quoi il s'en rapproche et s'en différencie, tout en exploitant des connaissances sur le domaine des textes étudiés. La majorité des travaux actuels en fouille de textes repose sur des méthodes statistiques ou neuronales de classification. Pour notre part, nous avons adopté une méthode symbolique - l'extraction de motifs fréquents et de règles d'association qui nous permettent d'exploiter des connaissances du domaine lors du processus de fouille, et d'envisager un traitement plus riche de l'information contenue dans les textes.

Ces travaux se situent essentiellement dans le contexte de la veille technologique et de l'analyse de l'information. Leur objectif est de montrer en quoi la méthodologie d'extraction de règles d'association constitue une approche intéressante pour la veille technologique. L'utilisateur final est supposé être un expert du domaine, dont l'objectif est soit de retrouver dans un ensemble de règles d'association des connaissances déjà établies dans le domaine soit de découvrir des informations qui n'ont pas encore le statut de connaissances mais qui sont appelées à le devenir (signaux dits « faibles »). Des algorithmes optimisés d'extraction de règles d'association existent et ont été adaptés pour analyser de très grosses bases de données. L'enjeu théorique est de pouvoir prendre en compte des propriétés et des concepts structurés en entrée des algorithmes de fouille, mais aussi de disposer de critères intelligents d'élagage pour appréhender ou pour trier le volume parfois très important des règles extraites (travail de recherche effectué dans le cadre de la thèse d'Université de Hacène Cherfi [11][8][10][9]).

Nos travaux récents nous ont permis d'acquérir une très bonne expertise sur les différents indices statistiques ou probabilistes existants pouvant être utilisés et combinés pour faciliter la lecture d'un ensemble de règles d'association. Nous avons également confronté les valeurs de ces indices au jugement humain d'un expert du domaine. D'une part, cette démarche a souligné l'intérêt de certains de ces indices ; d'autre part, les limites de ce ces indices purement statistiques nous amènent à travailler actuellement à la définition d'un nouvel indice de classement des règles d'association.

3.1.4.4. Hybridation de méthodes de classification

Le foisonnement des approches classificatoires pour la fouille de textes nous a conduit à nous interroger sur l'intérêt de l'usage d'une méthode plutôt que d'une autre, et sur l'opportunité de mettre en œuvre une

combinaison de plusieurs méthodes de classification. Dans un premier temps, nous nous sommes intéressés à l'étude de la complémentarité des méthodes de classification avec des cartes auto-adaptatives de Kohonen et des méthodes de classification par treillis (avec Jean-Charles Lamirel du projet Cortex au LORIA).

La complémentarité des méthodes de classification est traitée par la définition d'une méthode de projection des classes de Kohonen sur les classes d'un treillis construites à partir d'un même ensemble de données. Différentes heuristiques sont testées et une méthode d'évaluation de la qualité de cette projection reste à définir. Les premiers résultats ont montré que :

- Les classes du treillis, de par leur construction symbolique, sont plus faciles à interpréter que les classes de Kohonen, où chaque propriété est pondérée.
- Il est possible d'utiliser la structure hiérarchique du treillis pour associer à la carte de Kohonen une structure hiérarchique.

Les premiers résultats sont encourageants et sont actuellement en cours d'approfondissement et d'extension [15].

3.1.5. Les aspects bioinformatiques et la fouille de données en biologie

Une partie des membres de l'avant-projet Orpailleur s'intéressent de près à l'étude du génome et à l'application de méthodes propres à l'ECBD pour ce faire. Comme pour l'ECBD, deux approches peuvent être considérées, l'une plutôt numérique et l'autre plutôt symbolique.

Dans le cadre du plan État-Région, nous nous sommes rapprochés du laboratoire de Génétique et Microbiologie UA INRA 952 à l'UHP afin d'utiliser des données génomiques dans un travail de fouille de données.

Ce rapprochement s'est fait dans différentes directions :

- Co-encadrement de la thèse de Sébastien Hergalant ;
- L'enseignement avec des interventions dans le DESS *Ressources Génomiques et Traitements Informatiques* (RGTI) ;
- Les tâches collectives avec une participation à des séminaires communs, et la gestion de ressources communes comme le serveur de la communauté bioinformatique GCC.

Nous développons dans ce travail une méthode de fouille de données génomiques, dans laquelle l'utilisateur analyse un signal élaboré pour la circonstance par un HMM d'ordre deux [13][14]. Ce signal, qui représente une probabilité *a posteriori* de classer un résidu ou un groupe de résidus nucléotidiques dans un certain état, permet la localisation de répétitions dans de grandes séquences d'ADN génomique. Cette étude est le résultat d'une collaboration étroite entre des membres de l'équipe Orpailleur et des biologistes du LGM (Laboratoire de Génétique et Microbiologie de l'UHP-INRA), qui se fait essentiellement dans le cadre de la thèse de Sébastien Hergalant (thèse en co-tutelle). Les génomes des micro-organismes contiennent une importante variété de séquences réitérées s'étendant parfois sur 10% ou plus de leur matériel génétique total. La variabilité, la complexité et la spécificité taxonomique des répétitions chez les procaryotes sont similaires à celles décrites chez les organismes supérieurs. Notre modèle d'étude est *Streptomyces*, un genre bactérien aux multiples applications biotechnologiques. Les sièges de répétitions, qui sont des zones stationnaires du point de vue nucléotidique et qui sont détectées par un HMM, s'avèrent être des régions intéressant les biologistes, qui y voient le substrat d'évènements de recombinaison et des responsables des phénomènes d'instabilité génétique conduisant à des phénomènes de transfert horizontal.

Dans le cadre de l'ECBD symbolique pour la bioinformatique, l'avant-projet Orpailleur est impliqué dans une étude sur les « interactions gène-environnement et maladies cardio-vasculaires », avec l'unité INSERM U 525 (Faculté de Pharmacie, Université Henri Poincaré - Nancy 1). Il s'agit de d'exploiter, avec des méthodes de fouille de données, des données génétiques et biologiques de la Cohorte Stanislas, pour évaluer la part des facteurs génétiques et d'environnement dans la variabilité des phénotypes intermédiaires du risque cardio-vasculaire. Ce travail de recherche doit se faire sur la cohorte Stanislas, qui se compose (à l'origine) de 1006 familles, supposées saines, d'origine homogène (deux générations nées en France) avec au moins deux enfants

biologiques par famille. Cette cohorte permet aussi des études longitudinales du fait qu'elle est suivie pendant 10 ans. En outre, elle fournit des banques d'échantillons sanguins et d'ADN.

L'objectif global de ce travail de recherche est d'optimiser l'exploitation de la masse de données recueillies par les investigations en génétiques et en biologie, grâce aux méthodes symboliques de fouille de données, et en particulier, la recherche de motifs fréquents et l'extraction de règles d'association.

En préliminaire, une étude a été menée sur le rôle que peut jouer l'analyste dans le processus de fouille de données [25] : cette première étude montre des résultats encourageants et doit être continuée avec les données de la Cohorte Stanislas, ce qui constitue le sujet de thèse de Sandy Maumus (thèse en co-tutelle).

3.1.6. La fouille de bases de données en chimie organique

Dans le cadre mixte de la conception de systèmes intelligents et de la fouille de données, le travail de thèse de Sandra Berasaluce (thèse en co-tutelle LSIC-LIRMM, Montpellier et Orpailleur, soutenue en décembre 2002 [1]) porte sur l'extraction de connaissances et l'aide à l'interrogation et à la navigation dans des bases de données de chimie organique [6]. En chimie organique, il existe des bases de données d'un volume très important : plusieurs millions de substances décrites avec leurs propriétés chimiques, physiques et biologiques et plusieurs millions de réactions. L'interrogation de ces bases de données est avant tout conçue pour des besoins de documentation, plus que pour aider le chimiste organicien dans la résolution de problèmes de synthèse. L'idée ici est d'exploiter des techniques d'ECBD essentiellement symboliques pour découvrir des régularités dans les données, et faire émerger des schémas réactionnels génériques à partir de descriptions de réactions spécifiques. Ces schémas peuvent ensuite être réutilisés pour optimiser les requêtes et l'indexation dans les bases de données, en combinaison avec un système de connaissances, mais aussi pour mettre au point des plans de synthèse (voir par exemple plus loin les systèmes RÉSYN). Ce travail de recherche est polyvalent et revêt un ensemble d'intérêts théoriques et pratiques en représentation et gestion de connaissances, en fouille de données et en chimie organique (théorie de la synthèse organique).

3.2. La représentation et la gestion des connaissances

Mots clés : *représentation des connaissances (par objets)*, *gestion de connaissances*, *raisonnement par classification*, *raisonnement à partir de cas*, *logiques de descriptions*, *représentation de l'espace*, *treillis de relations spatiales*.

Un *système intelligent* - dans notre cas un *système de connaissances* - s'appuie sur une base de connaissances et un module de raisonnement pour résoudre des problèmes et gérer des connaissances dans un domaine donné. Les connaissances sont représentées par des formules auxquelles est associée une sémantique. Des mécanismes d'inférences permettent de dériver de nouveaux faits à partir des faits existant, en s'appuyant sur la sémantique du formalisme de représentation.

3.2.1. La classification, les systèmes de RCO et les logiques de descriptions

Dans le cadre de la représentation des connaissances, le projet Orpailleur s'intéresse particulièrement aux systèmes de *représentation de connaissances par objets* (RCO) et aux logiques de descriptions. La fonction d'un système de RCO est de stocker et d'organiser les connaissances autour de la notion d'objet et de fournir des services inférentiels destinés à compléter l'information disponible. Un système de RCO s'appuie sur une hiérarchie de classes liées entre elles par une relation de spécialisation. Une classe a une identité, un état et un comportement, à la manière d'un type abstrait de données. Elle regroupe un ensemble d'instances qui ont chacune une identité et un état propres, et un comportement décrit par la classe. La hiérarchie des classes est exploitée pour résoudre des problèmes par l'intermédiaire de procédures ou de mécanismes de raisonnement comme la *classification de classes* ou la *classification d'instances* (approche déclarative). La classification de classes consiste à placer une nouvelle classe dans une hiérarchie, tandis que la classification d'instances cherche à déterminer les classes dont un objet donné peut être une instance. La classification s'appuie sur le test de spécialisation qui consiste à vérifier qu'une classe donnée est plus générale qu'une autre classe.

Le processus de classification opère sur la hiérarchie des classes et cherche à mettre en évidence les dépendances implicites qui y existent : dépendances classes - classes et dépendances classes - instances.

Le *raisonnement par classification* s'appréhende alors comme une procédure de déduction opérant sur cette hiérarchie en s'appuyant sur les trois étapes initialisation de l'objet à classer, classification et exploitation de la classification.

Le RÀPC (raisonnement à partir de cas) peut se voir comme une extension naturelle du raisonnement par classification en milieu hiérarchique. Ce formalisme de raisonnement se propose de faire correspondre à l'énoncé d'un nouveau problème `P` une solution `Sol(P)` en tirant parti d'un ensemble de cas, qui sont des problèmes déjà résolus accompagnés de leurs solutions. Un cas mémorisé, ou cas source, est la donnée d'un couple énoncé de problème - solution `(P,Sol(P))` et fait partie d'une base de cas. Le processus du RÀPC se décompose en trois opérations principales : la remémoration, l'adaptation et la mémorisation. Étant donné un problème `cible` à résoudre, la remémoration consiste à retrouver dans la base de cas un énoncé de problème `source`, jugé similaire ou analogue à `cible`. Si `source` existe, sa solution `Sol(source)` est adaptée pour produire une solution `Sol(cible)` de `cible`. Une étape de mémorisation peut compléter les deux étapes précédentes. Les recherches sur le RÀPC sont très importantes et liées à bon nombre d'applications dans les recherches menées par Orpailleur. En particulier, Orpailleur mène des études approfondies sur le problème du RÀPC hiérarchique, la classification floue [23], et le problème de l'adaptation en général [22].

Les systèmes de RCO partagent de nombreuses caractéristiques avec les logiques de descriptions. Ces logiques s'appuient sur les notions de concepts (ils correspondent aux classes d'individus), de rôles (relations entre concepts) et d'individus (instances des concepts). Les concepts possèdent une syntaxe et une sémantique, et sont organisés en une hiérarchie par l'intermédiaire d'une relation de *subsomption*. La classification d'instances et la classification de classes sont à la base du raisonnement *terminologique*. Le projet Orpailleur exploite la logique de descriptions RACER dans plusieurs applications importantes [5][4][3][26][27].

3.2.2. La gestion des connaissances

La *gestion de connaissances* s'occupe de points spécifiques comme acquérir, accroître, transmettre et conserver les connaissances, plus particulièrement dans une organisation ou une entreprise de tout type. Il émerge donc des besoins d'acquisition, de diffusion, d'évaluation, d'évolution et de maintenance des connaissances.

Ainsi, la conception des systèmes intelligents actuels nécessite (i) d'exploiter des bases de connaissances et des ontologies, (ii) d'exploiter conjointement des bases de données de natures différentes et de volumes importants - le Web par exemple -, (iii) de traiter des problèmes complexes comme l'intégration, le croisement et la fouille de données hétérogènes, la navigation et la recherche d'information par le contenu. Dans un tel cadre, le langage XML est bien adapté à la description de documents textuels - c'est une de ses raisons d'être - mais la résolution de problèmes nécessitant des raisonnements et de la recherche d'information par le contenu des documents doit faire appel à la technologie des systèmes de représentation des connaissances, et en particulier, à celle des systèmes de RCO. XML a alors un rôle de passerelle à jouer, entre l'univers des données et celui des connaissances.

Pour le projet Orpailleur, les travaux sur la représentation et la gestion des connaissances sont à la base des recherches actuelles sur les systèmes intelligents en général et sur le Web sémantique en particulier.

Dans le cadre plus spécifique de la gestion de connaissances, la notion de *serveur de connaissances multidimensionnel* est de première importance dans les travaux de recherches du projet Orpailleur (qui est notamment impliqué dans le projet KVM - pour *Knowledge Valorization Matrix* - dans lequel interviennent les projets ECOO et MAIA du LORIA). Plus précisément, les éléments suivants sont en cours d'étude et de développement :

- L'architecture d'un serveur de connaissances pour la gestion d'une mémoire d'entreprise, prenant en compte les informations et les connaissances propres à une entreprise, mais aussi les informations disponibles sur le Web, et mettant en œuvre des mécanismes de raisonnement.
- Une approche symbolique pour l'aide à la décision avec des critères qualitatifs et sur la base du raisonnement à partir de cas. En particulier, des méta-connaissances d'ordre stratégique et tactique, des historiques et des connaissances temporelles doivent être appréhendés et exploités.

- Des fonctionnalités combinées de recherche d'information par le contenu et de fouille de données : une recherche d'information guidée par la fouille des données.

3.2.3. Systèmes à bases de connaissances et raisonnement spatial qualitatif

Nous nous intéressons à différentes formes de raisonnement spatial qualitatif et à la représentation de structures spatiales. Un premier projet concerne la classification de structures spatiales pour l'interprétation d'images satellitaires. Dans ce cadre, nous avons étudié :

- le calcul de relations topologiques sur des images satellitaires (données discrètes).
- la réification des relations et la représentation de structures spatiales qualitatives dans un système de RCO.
- l'organisation de relations topologiques sous forme de treillis, en particulier de treillis de Galois reliant les relations à des primitives de calcul.
- la classification de structures spatiales qualitatives dans un système de RCO.

Ce travail, effectué durant la thèse de Ludmila Mangelinck (1995-98) en collaboration entre l'INRA BIA Nancy et le LORIA (équipe RFIA), se poursuit actuellement par un approfondissement de l'étude des propriétés des treillis de relations topologiques et du calcul des relations sur des données discrètes [19][18]. Nous avons également poursuivi l'étude de la représentation de structures dans les systèmes de RCO [2][20]. L'objectif est maintenant d'étendre les treillis de relations topologiques à d'autres relations spatiales qualitatives, telles que les relations d'orientation et de distance.

Un deuxième projet concerne le développement d'un système de raisonnement à partir de cas pour l'interprétation et la comparaison de structures spatiales. Dans ce cadre, nous étudions :

- la modélisation de structures spatiales qualitatives par des graphes.
- la représentation et la classification de graphes dans les logiques de descriptions.
- la définition de chemins de similarité entre graphes modélisant des structures spatiales.
- l'adaptation des explications liées aux graphes.

Ce travail fait l'objet de la thèse de Jean-Luc Metzger (en cours depuis l'automne 2000), dans le cadre d'une collaboration entre l'INRA SAD et Orpailleur. La thèse est co-financée par l'INRA et l'INRIA et s'intitule « Élaboration de formalismes de représentation et de raisonnement pour les systèmes d'informations géographiques ». L'objectif est de construire un module de RÀPC qui permette de comparer des structures spatiales agricoles et d'adapter des explications liées à ces structures. La problématique agronomique est de comparer et généraliser des résultats d'enquêtes de terrain effectuées auprès d'agriculteurs pour comprendre les relations entre le fonctionnement et l'organisation spatiale des territoires agricoles. Une phase importante du projet a porté sur la modélisation du problème et des connaissances agronomiques [16][17]. Sur cette thématique nous collaborons avec des psychologues de la cognition (CODISANT, LPI-GRC, Université Nancy 2) et des linguistes (GRIC UMR 5612 CNRS, Lyon).

Pratiquement, nous utilisons le système RACER, développé à l'université de Hambourg, en Allemagne. Actuellement, nous développons un module externe de manipulation de graphes, qui s'appuie sur le mécanisme de classification de RACER [26][27].

Un troisième projet a démarré en 2002 sur la simulation de structures spatiales, en collaboration avec l'INRA et le laboratoire ESE, UPRESA 8079 CNRS- Université Paris-Sud. Le problème posé est de simuler la dispersion de transgènes dans un paysage agricole décrit par des caractéristiques qualitatives. Du point de vue informatique, il s'agit de combiner une approche qualitative (représenter un espace à partir de connaissances expertes) et une approche numérique (simuler les successions de cultures et la diffusion des gènes).

L'objectif général de ces différents projets est de développer un couplage de bases de données géographiques et de modèles de raisonnement spatial qualitatif.

3.3. Le Web sémantique et les systèmes intelligents de traitement de l'information

Mots clés : *Web sémantique*, *accès intelligent à l'information*, *recherche d'information intelligente*, *couplage fouille de données - recherche d'information*, *données semi-structurées*.

3.3.1. La problématique liée au Web sémantique

Le Web aujourd'hui est exploité par des personnes, qui en général, recherchent une information ou posent des questions via un moteur de recherche, et analysent le résultat elles-mêmes. Demain, le Web sera « sémantique » : il sera exploité en priorité par des machines qui traiteront des problèmes posés par des personnes et qui délivreront les résultats obtenus à ces mêmes personnes. Le *Web sémantique* va devenir ainsi un espace d'échange d'informations entre machines, permettant l'accès à de très grands volumes d'informations, et fournissant les moyens de gérer ces informations. Une machine sera en mesure d'appréhender le volume d'informations disponibles sur le Web et pourra fournir une aide conséquente aux personnes, si on la dote d'une certaine « intelligence ».

Actuellement, un certain nombre de fonctionnalités pas toujours très satisfaisantes ou efficaces existent sur le Web : des moteurs de recherche qui s'appuient sur une indexation primitive des pages Web, des procédures d'extraction et d'analyse de l'information. Il faut faire intervenir sur le Web l'exploitation de connaissances pour une meilleure gestion des informations disponibles : recherche, manipulation et résolution de problèmes. Le Web doit devenir un espace qualitatif aussi bien que quantitatif, un espace d'échange personnalisé et sûr. Parmi les besoins figurent la nécessité de disposer de langages pour exprimer le contenu des documents (XML est bien placé pour cela), d'une sémantique associée à ces langages, et de moteurs d'inférences associés, qui s'appuient sur cette sémantique pour raisonner. Des ressources de plusieurs types sont également nécessaires : ontologies, bases de connaissances (concepts, règles, fonctions, etc.), bases de données (documents, encyclopédies, etc.).

En tant que thème de recherche, le Web sémantique constitue un cadre fédérateur pour une variété de travaux de recherche qu'il faut combiner, parmi lesquels se trouvent la représentation et la gestion de connaissances et de documents, la fouille et l'analyse de textes, l'extraction et la recherche d'informations.

3.3.2. L'accès intelligent à l'information sur le Web

Le besoin en information est primordial dans de nombreux domaines, comme celui de la recherche ou celui de la veille scientifique et technique. Les données relatives à un domaine sont de plus en plus facilement accessibles ; toutefois, cette quantité croissante de données disponibles nécessite de mettre en œuvre des moyens particuliers pour les exploiter. Ainsi, la maîtrise de l'accès à l'information dans un fonds volumineux et hétérogène tel que le Web représente un enjeu majeur pour les consommateurs d'information (chercheurs, entreprises, etc.). Un objectif est de fournir aux chercheurs et aux spécialistes de l'information scientifique un environnement dans lequel ils puissent exploiter les données de leur domaine, pour des besoins de recherches bibliographiques ou d'analyses du domaine.

Les moteurs de recherche sont débordés par l'explosion du Web et ne répondent plus aux tâches de recherche d'information. Une des préoccupations du projet Orpailleur est de favoriser un accès intelligent aux données du Web en exploitant des connaissances relatives au domaine des données traitées. Dans ce cadre est proposée une approche générale qui couple l'exploitation de connaissances (extraites par des techniques de fouille de données) à un système de recherche d'information [28][29]. L'exploitation de connaissances est une technique classique pour favoriser la recherche d'information. L'originalité de l'approche introduite ici est d'utiliser un système de fouille de données - le système IntoBIB - qui fournit les moyens d'exploiter les données structurées d'un domaine (références bibliographiques en particulier) pour faire émerger des connaissances sur un domaine, comme des réseaux d'auteurs, le vocabulaire employé par tel ou tel auteur, etc.

L'idée maîtresse est que la fouille de données et la recherche d'information sont deux approches complémentaires pour appréhender des données structurées ou non : la fouille de données permet de guider la recherche d'information à partir des connaissances extraites des données, et, inversement, la recherche

d'information permet de guider la fouille de données par l'exploitation des connaissances issues de la fouille de données elle-même.

3.3.3. Les problèmes de représentation et de manipulation de documents

3.3.3.1. Objets et Web sémantique

L'omniprésence du Web modifie la manière d'envisager la recherche et l'échange de documents de toutes natures. Les contraintes de fonctionnement du Web privilégient la modularité et l'autonomie des documents manipulés. De ce point de vue, les technologies à base d'objets sont souvent utilisées pour répondre à ces contraintes. Qu'ils se trouvent mobilisés pour coder les documents, pour représenter leur contenu ou pour implanter des serveurs, les objets sont présents au cœur du Web, et leur rôle est appelé à se renforcer.

Par ailleurs, la nécessité de contrôler les informations, documents ou données, par l'intermédiaire d'une sémantique, renvoie aux problématiques de représentation des connaissances en général et par objets en particulier. C'est pourquoi, alors que se déploient d'importants projets sur le Web sémantique, les objets se font de plus en plus indispensables, à la fois en tant qu'éléments de représentation des connaissances, en tant que support de programmation et d'échange et en tant qu'unité de déploiement modulaire de services. En plus, l'utilisation de XML sous toutes ses facettes comme passerelle entre documents et objets trouve dans l'idée de Web sémantique une justification naturelle [5][4][3].

3.3.3.2. La manipulation de documents en fonction de leur contenu

L'utilisation des technologies de l'Internet permet de partager des documents et des connaissances. Les documents numériques et numérisés peuvent être rendus accessibles de manière standard et transparente auprès des utilisateurs concernés. Une ambition, à terme, est de réaliser de véritables serveurs de connaissances, permettant la recherche et la manipulation de ressources. Cependant, l'organisation des sites Internet se révèle une tâche coûteuse et la recherche de documents pertinents peu efficace ; la recherche et l'interrogation d'un site en s'appuyant sur le contenu des documents sont devenues des nécessités : les formalismes de représentation des connaissances sont les formalismes adéquats pour représenter ce contenu. La représentation du contenu d'un document permet de manipuler ce document pour faire de la recherche par spécialisation, par similitude, par analogie, etc.

Le langage XML permet de décrire les éléments du contenu des documents, mais il est aussi bien adapté à la description de données semi-structurées. Une des approches sur lesquelles un travail de recherche a été entrepris dans le projet Orpailleur consiste à décrire les données semi-structurées en XML, puis à combiner les fonctionnalités de XML et celles des systèmes de RCO pour exploiter au mieux les données semi-structurées, pour la résolution de problèmes, la recherche d'information, la fouille de données, la navigation, la visualisation, etc. [5][4][3].

5. Logiciels

5.1. Les modèles de Markov pour l'ECBD numérique : CarottAge

Participants : Florence Le Ber, Jean-François Mari [correspondant].

CarottAge est un acronyme crée à partir du mot *carotte* qui se dit *markov* en russe et du mot âge. C'est aussi un procédé d'analyse de la constitution des sols. CarottAge s'appuie sur la théorie des chaînes de Markov cachés pour calculer et afficher un signal dont l'analyse permet l'extraction de régularités temporelles et spatiales.

Ce système permet de représenter des observations spatio-temporelles comme des successions d'états, où les transitions entre états dépendent, suivant l'ordre n du modèle, de l'état courant et des n états précédents. La segmentation en n états peut s'effectuer selon une dimension (segmentation temporelle) ou deux dimensions (segmentation spatiale). Un emboîtement de HMM pour effectuer des segmentations spatio-temporelles a également été étudié.

Ce système est opérationnel et a été appliqué à l'extraction de connaissances sur les successions de cultures, leur évolution temporelle et leur répartition spatiale dans les bases de données TerUti, qui sont des bases de

données statistiques du ministère de l'agriculture. Il est aussi utilisé en bioinformatique pour la compréhension du phénomène de transfert horizontal chez les bactéries lactiques.

CarottAge est écrit en C++ et nécessite UNIX et X11R6. Il possède une licence GNU et il est la propriété de l'INRIA, de l'INRA et de l'Université de Nancy 2. Il est déposé à l'association de protection des logiciels. Ce logiciel est actuellement utilisé dans différentes unités de recherches INRA, en lien avec des problématiques de gestion ou de protection de l'eau dans les territoires agricoles.

5.2. Les logiciels pour la fouille de textes

Participants : Fairouz Chakkour, Hacène Cherfi, Yannick Toussaint [correspondant].

Les ressources, outils et environnements utilisés dans le cadre de la fouille de textes compte un étiqueteur, un lemmatiseur du français, et un système de d'extraction de motifs fréquents et d'extraction de règles.

L'étiqueteur de Brill attribue aux mots d'un texte une fonction grammaticale. Cet outil, initialement prévu pour travailler sur l'anglais a été adapté au français et au traitement de thésaurus par l'INALF et les membres d'Orpailleur. Il met en œuvre des techniques d'apprentissage statistiques et probabilistes pour construire des règles lexicales et contextuelles utilisées ensuite pour l'étiquetage. Le lemmatiseur du français qui produit le lemme d'une forme fléchie est développé par Fiammetta Namer à l'Université de Nancy 2, avec qui nous collaborons de façon permanente.

L'extraction de motifs fréquents et l'extraction de règles d'association se font dans la cadre de la veille technologique pour l'information scientifique et technique. Le travail d'extraction se fait sur des résumés d'articles scientifiques issus de notices bibliographiques (portant sur les phénomènes de mutation génétique de bactéries en résistance aux antibiotiques). Ces textes ont été préalablement indexés automatiquement par des termes issus du thésaurus PASCAL (la base bibliographique et le thésaurus PASCAL viennent de l'INIST).

Pour faciliter l'interprétation des résultats de ces expériences par un expert de l'INIST, nous avons développé deux environnements :

- Un environnement intégré de construction d'un treillis (à partir d'un ensemble de données booléennes) et de navigation hypertextuelle dans le treillis résultant. Il faut souligner qu'une structure de treillis construite à partir de données du monde réel est généralement très complexe à appréhender. Une réflexion est en cours pour adjoindre à cet environnement des possibilités graphiques pour en augmenter ainsi la convivialité.
- Un environnement d'analyse d'un ensemble de règles d'association qui est écrit en JAVA et qui permet de classer les règles suivant différents indices statistiques, ainsi que d'effectuer des recherches sur l'ensemble des règles en fonction du contenu des règles.

5.3. Les logiciels pour l'analyse et la simulation d'organisations spatiales agricoles

Participants : Florence Le Ber [correspondant], Amedeo Napoli.

Un système de reconnaissance de modèles d'organisations territoriales agricoles à partir d'images satellitaires a été réalisé en Y3 (pas de développements nouveaux depuis 1998). Ce système est destiné à aider les agronomes à interpréter les images dans un but de diagnostic et de prévision de l'évolution des territoires. La reconnaissance de modèles s'exprime comme une classification de structures, où les structures sont des ensembles d'objets reliés entre eux. Le système produit une reconnaissance cartographiée, c'est-à-dire qu'il produit une image finale où sont représentées par une même couleur les parties de l'image initiale associées à un même modèle.

Parallèlement, ont été développés des logiciels de simulation : à partir des données d'un territoire et d'un système de production agricole, il s'agit d'organiser l'occupation de l'espace comme pourrait le faire un agriculteur et de produire des cartes possibles d'occupation du sol. Trois modèles ont été implantés : un modèle à base de règles, un modèle multi-agents et un modèle de recuit simulé. Ces trois systèmes sont

utilisables pour des objectifs distincts (pas de développements nouveaux depuis 1999). Ce travail doit être repris prochainement et combiné avec les modèles de successions de cultures obtenus par HMM.

Un prototype est en cours de développement pour l'analyse et la comparaison de données d'enquêtes en exploitations agricoles. Ces données sont de différents types : cartes, données textuelles, synthèses graphiques. Une base de cas a été constituée sur des exploitations des Causses et de Lorraine, ainsi qu'une base de connaissances sur le domaine.

5.4. Le système KASIMIR

Participants : Mathieu d'Aquin [correspondant], Sébastien Brachais, Benoît Bresson, Jean Lieber, Amedeo Napoli.

Le système KASIMIR est développé dans le cadre du projet Kasimir et est dédié à l'aide au traitement du cancer du sein au stade locorégional. La nouvelle version de KASIMIR comprend plusieurs composants reliés par des médiateurs. Le composant prénommé PAULETTE est un système de RCO dédié à la résolution de problèmes. L'interface homme-machine permet de saisir un problème et d'afficher une solution, avec une mise à jour événementielle de la solution liée aux modifications du problème. Le composant PALÉTUVIER permet d'afficher la hiérarchie des classes manipulées par PAULETTE ou une sous-hiérarchie. KASIMIR est destiné à être étendu à d'autres localisations cancéreuses voire à d'autres types de problèmes, en cancérologie, ou ailleurs. D'ores et déjà, une version pour l'aide au diagnostic et au traitement du cancer de la prostate ainsi qu'une version pour la prise en charge des neutropénies ont été développées. Ces extensions envisagées ont induit une volonté de généricité de l'implantation de KASIMIR. Ainsi, KASIMIR est paramétré par des fichiers XML : l'interface est paramétrée par des fichiers décrivant notamment les attributs des problèmes à saisir et leurs types, PAULETTE est paramétrée par des fichiers décrivant les concepts à manipuler (concepts atomiques et définis, problèmes, solutions). Ainsi, tant que le formalisme de RCO implanté le permet, KASIMIR permet de construire des applications de résolution de problèmes dans un domaine quelconque, en implantant uniquement les fichiers de descriptions XML, sans modifier le programme JAVA.

Des développements sont en cours pour améliorer et étendre le système KASIMIR. En particulier, une extension à la classification hiérarchique floue a été effectuée. Une extension plus générale à la classification élastique, selon une approche RÀPC, est à l'étude. Enfin, la mise à disposition sur le Web des services de KASIMIR selon les principes du Web sémantique est également étudiée.

5.5. Les systèmes RÉSYN et RÉSYN-ASSISTANT

Participants : Sandra Berasaluce [correspondante], Claude Laurenço [LSIC et LIRMM Montpellier], Jean Lieber, Amedeo Napoli.

À l'origine, le système RÉSYN a été développé en Y3 dans le cadre du GDR CNRS 1093 « Traitement Informatique de la Connaissance en Chimie Organique ». Le système RÉSYN a pour objet la planification de synthèses en chimie organique. Une extension de RÉSYN, appelée RÉSYN/RÀPC a été développée par Jean Lieber, pour intégrer le raisonnement à partir de cas (RÀPC) dans RÉSYN et ainsi compléter le seul raisonnement par classification utilisé dans RÉSYN. Actuellement, c'est le prototype RÉSYN-ASSISTANT qui a pris la relève : le système est écrit en JAVA et reprend une bonne partie de développements effectués sur RÉSYN : l'objectif est de proposer une aide à la compréhension des problèmes de synthèse organique. Pour cela, des outils de perception par blocs des molécules ont été développés (conduisant à une représentation des molécules sous plusieurs points de vue). Les développements actuels oriente RÉSYN-ASSISTANT vers l'extraction de connaissances dans des bases de données de réactions.

5.6. Le traitement intelligent de l'information et le Web sémantique

Participants : Rim Al Hulou, Amedeo Napoli, Emmanuel Nauer [correspondant].

Deux systèmes principaux de traitement de l'information sont actuellement en cours de développement. Un système générique de traitement d'informations et de données brutes - en fait une boîte à outils composée d'un

ensemble de modules - est actuellement en cours de développement. Le système baptisé « IntoBib », dont la finalité est l'aide à la navigation et à la recherche d'information sur le Web, repose sur un choix particulier d'assemblage de modules. Les modules proviennent de différents horizons. La boîte à outils DILIB, qui est une plate-forme dédiée au traitement de l'information reposant sur le format SGML, a fourni un certain nombre de modules. D'autres modules nécessaires à des traitements spécifiques ont été développés de façon *ad hoc* : un module de mise en corrélation de descripteurs de langues différentes dans des notices multilingues, un module de classification par treillis de documents suivant un treillis de concepts, un module de normalisation des auteurs, et un module de normalisation des descripteurs dans un contexte multi-bases. D'autres modules encore proviennent du réseau - lemmatiseur, grapheur - ou sont directement utilisables sur le réseau (moteurs de recherche, service de traduction, etc.).

Un système dont la finalité est la prise en compte et la manipulation de données semi-structurées est développé pour manipuler des documents par leur contenu, dans le cadre d'un système de connaissances. Les données sont essentiellement des documents textuels, décrits en XML. Dans un tel cadre, le langage XML sert de support à la description des documents tandis que la logique de descriptions RACER permet de mettre en œuvre des raisonnements par classification et d'exploiter des connaissances du domaine, pour la recherche d'informations par le contenu, la classification de requêtes et le traitement de requêtes analogues.

5.7. DefineCrawler : un crawler paramétrable pour la recherche intelligente sur le Web

Participants : Emmanuel Nauer [correspondant], Amedeo Napoli.

Les nombreuses possibilités d'exploiter des connaissances d'un domaine pour guider la recherche d'information sur le Web nous ont amenés à développer un système générique, nommé DefineCrawler, capable de parcourir le Web en suivant les liens hypertextes et dont le comportement peut être facilement paramétré, par le biais d'un document XML, pour définir des systèmes de recherche d'information particuliers. Un ensemble de paramètres génériques a été défini à partir de l'étude d'outils de recherche d'information sur le Web (moteurs de recherche, agents parcourant le Web en suivant les liens hypertextes, etc.), des paradigmes qui régissent les systèmes de recherche documentaires classiques, et des problèmes particuliers liés à la recherche d'information sur le Web. Nous avons retenu trois types de paramètres :

- les paramètres de départ, qui définissent le comportement global (profondeur maximale de parcours, ensemble d'URL de départ, répertoire dans lequel sera stocké l'ensemble des données recueillies lors du parcours, nombre de processus parallèles pour parcourir le Web, temps maximal de parcours, ...).
- les paramètres de validation qui contiennent un ensemble de conditions (reliées par des opérateurs booléens) que doivent satisfaire les documents. L'objectif est ici d'éliminer les documents inintéressants ou inutiles pour l'utilisateur.
- les paramètres d'évaluation qui permettent de définir un ensemble de conditions d'évaluations supplémentaires dont les résultats peuvent être combinés, via l'expression d'une formule mathématique, pour définir la méthode de calcul du score associé à chaque document. Ce score servira à établir le classement global des documents ainsi que l'ordre de parcours des liens hypertextes.

Pour permettre un maximum d'interopérabilité, chaque condition de validation ou d'évaluation est définie par une commande externe à l'agent dont la seule contrainte est de renvoyer une valeur numérique. Le choix de faire appel à une commande externe à l'agent laisse envisager l'utilisation de tout type de commande ou système.

8. Actions régionales, nationales et internationales

8.1. Actions locales

8.1.1. La collaboration URI et Orpailleur

Participants : Dominique Besagni [INIST], Claire François [INIST], Bernard Maudinas [INIST], Xavier Polanco [INIST], Ivana Roche [INIST], et tout Orpailleur.

La collaboration entre l'équipe URI (Unité de recherches et d'innovation) de l'INIST et le groupe Orpailleur cherche à mettre à profit la spécificité et les contextes propres aux deux équipes pour faire avancer les recherches et le développement de logiciels dans le cadre de l'analyse de l'information scientifique et technique. Les finalités et la valorisation de la collaboration portent essentiellement sur la mise en œuvre de recherches et de projets communs. Des contacts permanents existent entre les deux équipes, globalement et individuellement. Parmi les thèmes principaux qui intéressent cette collaboration se trouvent l'ECBD et plus particulièrement la fouille de textes. Plus précisément, des travaux sont en cours de développement sur un certain nombre de points dont :

- L'étude des stratégies d'interrogation de grandes bases de données textuelles et l'élaboration d'une typologie de requêtes.
- La prise en compte de données semi-structurées provenant de bases de données textuelles hétérogènes.
- L'étude et la mise en œuvre d'une méthodologie pour la fouille de textes, avec l'extraction et l'analyse de structures prédicatives et l'utilisation du système NEURODOC pour l'ECBD.
- L'étude de XML comme une plate-forme intermédiaire pour la description de documents textuels (scientifiques et techniques), en vue d'une manipulation intelligente de ces documents dans l'environnement d'un système de RCO.

Par ailleurs, un projet commun est en cours de développement, qui concerne la réalisation d'une plate-forme expérimentale d'analyse de l'information textuelle.

8.1.2. La collaboration READ et Orpailleur

Participants : Abdel Belaïd [READ, LORIA], Fairouz Chakkour, Hacène Cherfi, Yannick Toussaint.

Le projet *Citations* vise à traiter automatiquement les références bibliographiques des articles scientifiques qui ont été numérisées. L'objectif est donc d'utiliser conjointement des méthodes linguistiques simples et des règles d'agglomération pour aider à la segmentation des références et retrouver les champs bibliographiques de chacune des entrées.

8.2. Actions nationales

8.2.1. L'ARC INRIA GENI : génération et inférence

Participants : Fairouz Chakkour, Hacène Cherfi, Amedeo Napoli, Yannick Toussaint.

La génération de textes de bonne qualité passe (entre autre) par une interaction étroite entre génération, représentation des connaissances et inférence. L'ARC INRIA GENI se propose d'examiner les relations existant entre les formalismes de représentation des connaissances utilisés en linguistique et les logiques de descriptions. L'ARC INRIA GENI réunit les projets ATOLL (INRIA Rocquencourt), LANGUE & DIALOGUE (LORIA), les laboratoires ILPL-IRIT (Toulouse) et LATTICE (Paris 7), et enfin Orpailleur.

Peu de travaux existent sur les rapports entre les logiques de descriptions et le traitement du langage naturel, aussi bien sur un plan pratique que théorique. Un premier travail de recherche sur l'analyse de résumés de textes biologiques avec des logiques de descriptions a été mis en œuvre dans le cadre de l'ARC INRIA ESCRIRE auquel a participé le projet Orpailleur. Un des objectifs de l'ARC GENI est de poursuivre cette initiative et d'examiner jusqu'à quel point les logiques de descriptions peuvent être utilisées pour implanter les ontologies

lexicales et encyclopédiques sous-jacentes au traitement automatique de la langue et représenter le sens d'un texte.

L'ARC GENI sera aussi l'occasion d'aborder des questions sur la forme appropriée des éléments dans une terminologie et, en parallèle, sur l'expressivité du langage de représentation, ici la logique de descriptions sous-jacente. Il ne faut pas oublier que plus le langage de représentation est riche, plus le problème des inférences dans le cadre logique est complexe.

De plus, dans l'ARC GENI, la représentation du sens d'un texte à l'aide d'une logique de descriptions passe également par la définition d'une grammaire associant à une expression $\mathtt{F}$ de la langue une représentation sémantique qui correspond à une formule ϕ de la logique de descriptions, l'interprétation de l'expression $\mathtt{F}$ étant alors un modèle $\phi^{\mathtt{I}}$ de la formule ϕ.

8.2.2. Une collaboration avec l'INRA

Participants : Florence Le Ber, Jean-François Mari, Jean-Luc Metzger, Amedeo Napoli.

Cette collaboration déjà ancienne s'exprime actuellement dans deux projets principaux, l'un concernant les systèmes à bases de connaissances, l'autre la fouille de données.

Dans le cadre du premier projet, nous avons travaillé en 2002 avec les chercheurs agronomes de l'INRA SAD (unités de Mirecourt, Montpellier, Toulouse) à la formalisation de données d'enquêtes en exploitations agricoles dans le but de développer un système à bases de connaissances. Nous participons également à un groupe de recherche inter-unités INRA, le groupe FORTE (pour *Formes d'organisations territoriales des activités agricoles à finalité environnementales*).

Dans le cadre du second projet, une application des modèles de Markov d'ordre 1 et 2 a été mise en œuvre pour la reconnaissance de successions culturales en collaboration avec l'unité INRA SAD de Mirecourt et l'unité INRA Agronomie de Toulouse. Les modèles de Markov ont été utilisés sur des données de différentes régions (Sud-ouest, Lorraine, Bassin de la Seine) et dans des cadres applicatifs distincts.

8.2.2.1. Modélisation de dispersion de transgènes.

Ce projet est récent et implique différentes équipes d'agronomes, de généticiens et de biométriciens. L'objet est d'étudier la façon dont des plantes modifiées génétiquement (OGM) peuvent (se) diffuser dans un paysage agricole « réel ». Dans ce cadre, notre objectif est de développer des méthodes de simulation de l'organisation spatio-temporelle d'un territoire en combinant des connaissances qualitatives et numériques. En particulier nous joindrons nos approches qualitatives sur l'espace avec les approches à base de HMM utilisées en ECBD numérique (voir 3.1.3.1).

Projets :

- MENRT, Programme Impact des OGM : Modélisation de dispersion de transgènes à l'échelle de paysages agricoles (2002-2004, responsable C. Lavigne, UPRESA 8079 CNRS).
- CNRS (STIC, SHS), Programme Société de l'information : Usage raisonné des représentations spatiales comme objets intermédiaires dans des projets de développement participatif (2002-2003, responsable S. Lardon, INRA/ENGREF).

8.2.3. Une collaboration avec le Cemagref

Participants : Florence Le Ber, Jean-François Mari, Amedeo Napoli.

Nous avons entamé une collaboration avec le laboratoire LISC du Cemagref à Clermont Ferrand, qui est spécialisé en modélisation des écosystèmes (approches statistiques et individu centrées). La collaboration porte sur les points suivants :

- utiliser des méthodes de fouille de données (HMM et treillis de Galois) pour analyser les résultats des modèles individu centrés.
- comparer ces méthodes avec les approches statistiques développées au LISC.
- combiner les approches qualitatives de l'espace avec les modèles individu centrés pour la modélisation d'écosystèmes spatialisés.

L'objectif de cette collaboration est de développer des outils pertinents et des méthodes efficientes pour l'aide à la gestion et à la protection de l'environnement.

Projet :

CNRS (STIC) - IGN - Cemagref, Programme GETM : Modélisation, comparaison et interprétation d'organisations spatiales agricoles. Aspects techniques, sociaux et cognitifs de la mobilisation de représentation de l'espace (2002-2004, responsable F. Le Ber, LORIA).

8.2.4. Le projet KASIMIR

Participants : Mathieu d'Aquin, Sébastien Brachais, Benoît Bresson, Jean Lieber, Amedeo Napoli.

Le projet Kasimir a pour objectif la gestion des connaissances en oncologie [22][21]. Il réunit Orpailleur, l'association ONCOLOR (réseau de soins en cancérologie de la région Lorraine), le centre régional de lutte contre le cancer de Lorraine Alexis Vautrin (Vandœuvre-lès-Nancy) et le laboratoire d'ergonomie du CNAM (Paris), dont les membres concernés font également partie du projet INRIA Eiffel. La gestion des connaissances visée dans le projet Kasimir se fonde sur une approche originale issue des travaux en ergonomie de Pierre Falzon et Catherine Sauvagnac sur les activités méta-fonctionnelles. Elle s'appuie sur une confrontation de faits « du monde réel » (les cas médicaux) avec la base de connaissances : celle-ci doit évoluer pour tenir compte de ces faits.

Pour chaque localisation cancéreuse, le traitement s'appuie sur un *référentiel* qui est un protocole de décision, issu d'études statistiques et utilisable de manière littérale pour environ 70% des cas de cancer. Le référentiel peut être vu comme un ensemble de règles dont les conclusions sont des traitements. Le logiciel KASIMIR/RÉFÉRENTIEL a été implanté et permet d'effectuer cette prise de décision pour les cancers du sein au stade locorégional (non métastatique). Il ne propose de traitement que pour les cas couverts par le référentiel. Son implantation s'appuie sur une représentation des connaissances par objets. Sa base de connaissances est une représentation du référentiel et a nécessité une mise en évidence de nombreuses connaissances implicites. La dernière version en date de KASIMIR/RÉFÉRENTIEL se veut « générique », ce qui permettra de réutiliser le système pour d'autres localisations cancéreuses : une version pour le cancer de la prostate a d'ores et déjà été réalisée.

Actuellement, KASIMIR/RÉFÉRENTIEL évolue vers un serveur de connaissances selon les principes du Web sémantique : un premier travail vise à rendre opérationnel le système KASIMIR/RÉFÉRENTIEL pour une utilisation par des médecins répartis sur la région Lorraine. Les cas hors référentiel pour le traitement du cancer du sein constituent 30% des cas à traiter : les cancérologues tentent alors d'*adapter* le référentiel, lors de réunions du « comité de concertation pluridisciplinaire ». Dans ce contexte, la conception et le développement du système KASIMIR/HORS RÉFÉRENTIEL sont envisagés. Ce système doit s'appuyer sur les principes du raisonnement à partir de cas (RÀPC) : il doit effectuer une sélection de la règle à adapter puis une adaptation du traitement associé à cette règle. La mise en œuvre algorithmique de l'adaptation se fera selon les principes de la classification élastique, qui ont été développés pour la planification de synthèse en chimie organique dans RÉSYN/RÀPC. Un travail d'acquisition et de modélisation des connaissances d'adaptation est nécessaire, et s'appuie notamment sur des discussions avec les experts sur des comptes-rendus du comité de concertation pluridisciplinaire qui est chargé d'examiner les cas hors référentiel, nécessitant une adaptation. Cette acquisition des connaissances a permis de mettre en évidence différents schémas d'adaptation effectivement réalisées par les cancérologues lors de réunions du comité de concertation pluridisciplinaire. Ce travail se poursuit par une acquisition des connaissances pour l'instanciation de ces schémas. À titre d'exemples, les adaptations se font en cas de contre-indications ou en cas de caractéristiques trop imprécises du problème courant pour la prise de décision. Par ailleurs, cette étude a montré la nécessité de prendre en compte l'imprécision sur les seuils utilisés dans la décision. Par exemple, 4 cm est un seuil de taille de tumeur : selon que la taille de la tumeur du patient est inférieure ou supérieure à ce seuil, la proposition thérapeutique va changer. Comme ce seuil est imprécis, la décision prise pour une taille de tumeur de $3,9$ cm est sujette à caution. Une première version de KASIMIR/HORS RÉFÉRENTIEL s'appuyant sur la classification hiérarchique floue est en cours de finition et permet de prendre en compte ce problème de seuil. La représentation des connaissances pour le raisonnement à partir de cas dans un tel cadre nécessite par ailleurs l'utilisation de

points de vue [12] (voir aussi M. d'Aquin, Besoins en représentation des connaissances et représentation par objets multi-points de vue pour l'adaptation en raisonnement à partir de cas, Mémoire de DEA, Université Henri Poincaré Nancy 1, 2002).

En parallèle, de nouvelles fonctionnalités de KASIMIR/RÉFÉRENTIEL vont être développées. Un éditeur de connaissances facilitera la description de nouvelles bases de connaissances et la maintenance des anciennes. Une gestion intelligente des données sur les patients doit également être mise en œuvre. Cette gestion doit être couplée au moteur de raisonnement de KASIMIR/RÉFÉRENTIEL (les données accessibles sont les données présentes de façon explicite dans la base et celles qui peuvent être déduites).

8.2.5. Le projet KVM

Participants : Gérôme Canals [ECOO, LORIA], François Charpillet [MAIA, LORIA], Vincent Chevrier [MAIA, LORIA], Claude Godard [ECOO, LORIA], Abdelhalim Larhlimi [ECOO, LORIA], Amedeo Napoli, Emmanuel Nauer, Laszlo Szathmary.

Dans le cadre du projet KVM, l'avant-projet Orpailleur collabore avec les projets ECOO et MAIA du LORIA. L'objectif de cette collaboration est de construire un système générique pour la gestion des connaissances, et plus particulièrement la gestion d'une mémoire d'entreprise. Pour leur part, les membres de l'avant-projet Orpailleur qui sont impliqués travaillent sur la mise au point d'un système capable de gérer un référentiel multidimensionnel de connaissances, autour duquel vont graviter les éléments d'information circulant dans une entreprise : données, connaissances, et informations de toutes natures (messages, notes, notices, documents, modes d'emploi, etc.). Dans ce cadre, il faut s'intéresser à plusieurs thèmes principaux, parmi lesquels : (i) l'étude d'un serveur de connaissances pour la gestion d'une mémoire d'entreprise, (ii) une approche symbolique pour l'aide à la stratégie et à la prise de décision, sur la base du RÀPC, (iii) l'étude et la mise en place de principes de conception d'un SII pour l'entreprise.

Sur le plan pratique, ce travail de recherche doit déboucher sur l'implantation d'un SII pour la gestion des connaissances dans une entreprise, qui intègre l'ensemble des fonctionnalités décrites ci-dessus.

8.2.6. Une collaboration sur le thème du RàPC (Université de Lyon 1)

Participants : Béatrice Fuchs [LISI, Université de Lyon 1], Jean Lieber, Alain Mille [LISI, Université de Lyon 1], Amedeo Napoli.

Dans le cadre du RÀPC, l'étape d'adaptation joue un rôle central. C'est malheureusement une étape très peu modélisée dans la littérature, malgré l'organisation de journées sur l'adaptation en RÀPC dans des contextes internationaux. Des modèles ont été proposés parallèlement dans l'avant-projet Orpailleur et dans l'équipe dirigée par Alain Mille (professeur des Universités, au LISI à l'Université Claude Bernard Lyon 1). Une collaboration s'est engagée avec Alain Mille et Béatrice Fuchs (maître de conférences à l'Université Lyon 3, LISI), pour confronter ces modèles de l'adaptation et les enrichir par nos expériences respectives. Ce travail a conduit à un premier modèle qui s'appuie sur deux idées principales. La première est le fait de considérer un cas comme un chemin dans un espace de recherches, ce qui permet de bénéficier des recherches en planification à partir de cas. La seconde est de décomposer la relation entre le problème à résoudre et le problème dont on connaît une solution, de façon à décomposer la tâche complexe de l'adaptation en sous-tâches plus simples.

Dans la continuité de ces recherches, un algorithme d'adaptation générique a été proposé. Il s'appuie sur les notions d'appariement entre problèmes et de dépendance entre un problème et la solution qui lui est associée. Bien que cet algorithme repose sur une représentation très simple des problèmes et des solutions (par des n-uplets de réels), il peut se généraliser à des représentations plus complexes. L'étude de telles généralisations dans une double optique théorique et applicative constitue une perspective de cette collaboration.

8.2.7. Les travaux de recherches post-GDR CNRS 1093 TICCO

Participants : Sandra Berasaluce, Claude Laurenço [LSIC et LIRMM Montpellier], Jean Lieber, Amedeo Napoli.

Le GDR CNRS 1093 TICCO - *Traitement informatique de la connaissance en chimie organique* - est actuellement terminé, mais un certain nombre de travaux de recherche en découlent directement. Le GDR CNRS 1093

TICCO a réuni des chercheurs en chimie organique du LSIC à Montpellier, des chercheurs en informatique du LIRMM (Montpellier) et du LORIA et des chercheurs de l'industrie pharmaceutique (Sanofi-chimie, Roussel-Uclaf, et Institut de Recherches Servier entre autres). L'objectif du GDR a été l'étude et la mise en œuvre de systèmes d'aide à la planification de synthèses de molécules avec comme base informatique les systèmes de RCO, le raisonnement par classification et le RÀPC. Ce travail a nécessité et nécessite encore des recherches sur une représentation des objets de la chimie organique, une représentation des plans de synthèses de molécules, une modélisation des raisonnements élémentaires et des stratégies de synthèse employés par les chimistes pour résoudre un problème de synthèse.

Le travail de thèse de Sandra Berasaluce, co-encadré par Claude Laurenço (LSIC et LIRMM Montpellier), entre dans le cadre du GDR TICCO. À ce titre, Sandra Berasaluce peut bénéficier de la double expertise chimie et informatique, et le GDR TICCO offre un environnement idéal pour ce travail de recherches bidisciplinaire.

9. Diffusion des résultats

9.1. Animation de la Communauté scientifique

- Participation à des groupes de travail nationaux (GDR).
- Participation et responsabilité dans l'ACI Contraintes spatiales et temporelles pour les systèmes d'information géographique (voir http://www.cmi.univ-mrs.fr/~jeansoul/SOLEIL/).
- Participation à des comités de lecture de revues, à l'organisation de numéros spéciaux de revues et à l'édition d'ouvrages de recherche.
- Organisation de colloques et participation à des comités de programme.

9.2. Enseignement

- Enseignements et organisation scientifique de cours à tous les niveaux universitaires.
- Encadrements de thèses, enseignements et stages de DEA, de DESS, stages d'étudiants en écoles d'ingénieurs et à l'IUT.
- Participation à des jurys de thèse et de HDR.

10. Bibliographie

Thèses et habilitations à diriger des recherche

[1] S. BERASALUCE. *Fouille de données at acquisition de connaissances à partir de bases de données de réactions chimiques.* Thèse de chimie informatique et théorique, Université Henri Poincaré Nancy 1, 2002.

Articles et chapitres de livre

[2] F. LE BER, A. NAPOLI. *The design of an object-based system for representing and classifying spatial structures and relations.* in « Journal of Universal Computer Science », septembre, 2002.

Communications à des congrès, colloques, etc.

[3] R. AL-HULOU, O. CORBY, R. DIENG-KUNTZ, J. EUZENAT, C. MEDINA RAMIREZ, A. NAPOLI, R. TRONCY. *Une plate-forme XML pour représenter des documents et leur contenu pour la mise en oeuvre du Web sémantique.* in « Journées scientifiques Web sémantique, Ivry, France », éditeurs J. CHARLET, P. LAUBLET, C. REYNAUD., octobre, 2002, http://www.lalic.paris4.sorbonne.fr/stic/octobre/programme0209.html.

[4] R. AL-HULOU, O. CORBY, R. DIENG-KUNTZ, J. EUZENAT, C. RAMIREZ, A. NAPOLI, R. TRONCY. *Three knowledge representation formalisms for content-based manipulation of documents.* in « Workshop on Semantic Web - SemWeb@KR2002, Toulouse, France », éditeurs M. CRISTANI., avril, 2002, Publication CD-ROM.

[5] R. AL-HULOU, A. NAPOLI. *Combining XML and DL for describing and querying documents* . in « International Workshop on Description Logics - DL'2002 , Toulouse, France », éditeurs I. HORROCKS, S. TESSARIS., avril, 2002.

[6] S. BERASALUCE, C. LAURENÇO, A. NAPOLI. *Extraction de connaissances à partir de bases de données de réactions en chimie organique.* in « Treizième journées francophones d'ingénierie des connaissances - IC'2002, Rouen, France », éditeurs B. BACHIMONT., pages 151-162, juin, 2002.

[7] M. CADOT, A. NAPOLI. *Description et comparaison de deux techniques d'extraction automatique de règles dans une base de données.* in « 9ièmes Rencontres de la Société Francophone de Classification, Toulouse, France », septembre, 2002.

[8] H. CHERFI, Y. TOUSSAINT. *Adéquation d'indices statistiques à l'interprétation de règles d'association.* in « 6èmes Journées internationales d'Analyse statistique des Données Textuelles - JADT 2002, Saint-Malo, France », éditeurs P. S. A. MORIN., pages 233-244, mars, 2002.

[9] H. CHERFI, Y. TOUSSAINT. *Fouille de textes par combinaison de règles d'association et d'indices statistiques.* in « 1er Colloque International sur la Fouille de Textes - CIFT'2002, Hammamet, Tunisie », pages 67-80, septembre, 2002.

[10] H. CHERFI, Y. TOUSSAINT. *How Far Association Rules and Statistical Indices help Structure Terminology ?.* in « Workshop of ECAI2002 : Natural Language Processing and Machine Learning for Ontology Engineering OLT'02, Lyon, France », éditeurs A. M. N. AUSSENAC-GILLES., pages 5-9, juillet, 2002.

[11] H. CHERFI, Y. TOUSSAINT. *Interprétation des règles d'association extraites par un processus de fouille de textes.* in « 13ème Congrès francophone AFRIF-AFIA de Reconnaissance des Formes et d'intelligence Artificielle - RFIA'02, Angers, France », volume 3, pages 975-983, janvier, 2002.

[12] M. D'AQUIN, J. LIEBER, A. NAPOLI. *Représentation multi-points de vue des connaissances pour l'adaptation.* in « 10ème séminaire français de raisonnement à partir de cas - RàPC'2002, Paris, France », éditeurs M.-C. JAULENT, C. L. BOZEC, E. ZAPLETAL., pages 23-31, mai, 2002.

[13] S. HERGALANT, B. AIGLE, B. DECARIS, P. LEBLOND, J.-F. MARI. *Fouille de données à l'aide de HMM : Application à la détection de réitérations intragénomiques.* in « Journées Ouvertes Biologie Informatique Mathématiques - JOBIM'02, Saint Malo, France », pages 269-273, juin, 2002.

[14] S. HERGALANT, B. AIGLE, B. DECARIS, P. LEBLOND, J.-F. MARI. *Intragenomic reiterations detection using hidden Markov models.* in « 10th International Conference on Intelligent Systems for Molecular Biology - ISMB 2002, Edmonton, Canada », août, 2002.

[15] J.-C. LAMIREL, Y. TOUSSAINT. *Association de méthodes symboliques et numériques pour l'analyse du contenu de bases de données.* in « 13ème Congrès Francophone AFRIF-AFIA de Reconnaissance des Formes

et Intelligence Artificielle - RFIA'2002, Angers, France », janvier, 2002.

[16] F. LE BER, C. BRASSAC, J.-L. METZGER. *Analyse de l'interaction experts - informaticiens lors de la modélisation de connaissances spatiales.* in « Journées francophones sur l'ingénierie des connaissances - IC'2002, Rouen, France », mai, 2002.

[17] F. LE BER, J.-L. METZGER, A. NAPOLI. *Modelling and comparing maps with graphs.* in « ECAI Workshop on Spatial and Temporal Reasoning, Lyon, France », juillet, 2002.

[18] F. LE BER, A. NAPOLI. *A Galois lattice for qualitative spatial reasoning and representation.* in « ECAI Workshop on Advances in Formal Concept Analysis for Knowledge Discovery in Databases - FCAKDD'2002, Lyon, France », éditeurs M. LIQUIÈRE., juillet, 2002.

[19] F. LE BER, A. NAPOLI. *Design and comparison of lattices of topological relations based on Galois lattice theory.* in « Principles of Knowledge Representation and Reasoning - KR'2002, Toulouse, France », Morgan Kaufmann Publishers, éditeurs D. FENSEL, F. GIUNCHIGLIA, D. MCGUINNESS, M.-A. WILLIAMS., pages 37-46, avril, 2002.

[20] F. LE BER, A. NAPOLI. *Object-based Representation and Classification of Spatial Structures and Relations.* in « International Conference on Tools with Artificial Intelligence - ICTAI'2002, Washington DC », IEEE, novembre, 2002.

[21] J. LIEBER, M. D'AQUIN, P. BEY, B. BRESSON, O. CROISSANT, P. FALZON, A. LESUR, J. LÉVÊQUE, V. MOLLO, A. NAPOLI, M. RIOS, C. SAUVAGNAC. *The Kasimir Project : Knowledge Management in Cancerology.* in « 4th International Workshop on Enterprise Networking and Computing in Health Care Industry - HealthComm 2002, Nancy, France », pages 125-127, juin, 2002.

[22] J. LIEBER. *Recopier c'est déjà adapter : six types d'adaptation par copie.* in « 10ème séminaire français de raisonnement à partir de cas - RàPC'2002, Paris, France », éditeurs M.-C. JAULENT, C. L. BOZEC, E. ZAPLETAL., pages 11-21, mai, 2002.

[23] J. LIEBER. *Strong, Fuzzy and Smooth Hierarchical Classification for Case-Based Problem Solving.* in « 15th European Conference on Artificial Intelligence - ECAI'02, Lyon, France », IOS Press, Amsterdam, éditeurs F. VAN HARMELEN., pages 81-85, juillet, 2002.

[24] J.-F. MARI, F. LE BER, M. BENOÎT. *Segmentation temporelle et spatiale de données agricoles.* in « Actes des 6èmes Journées Cassini 2002, Presqu'Île de Crozon, France », pages 251-272, septembre, 2002.

[25] S. MAUMUS, A. NAPOLI, R. TAOUIL, S. VISVIKIS. *A first study of the central role of the analyst in the knowledge discovery process in biology.* in « Poster session at the 10th International Conference on Intelligent Systems for Molecular Biology - ISMB'02, Edmonton, Canada », août, 2002, Poster.

[26] J.-L. METZGER, F. LE BER, A. NAPOLI. *Using DL for a Case-Based Explanation System.* in « International Workshop on Description Logics - DL'2002, Toulouse, France », éditeurs I. HORROCKS, S. TESSARIS., pages 203-210, avril, 2002.

[27] J.-L. METZGER, F. LE BER, A. NAPOLI. *Utilisation des logiques de descriptions pour la représentation des structures spatiales.* in « Journées CASSINI 2002, Presqu'île de Crozon, France », septembre, 2002.

[28] E. NAUER. *Complémentarité entre fouille de données et recherche d'information dans le cadre d'analyses bibliométriques.* in « 13ème Congrès francophone AFRIF-AFIA de Reconnaissances des Formes et d'intelligence Artificielle, Angers, France », volume 3, pages 965-974, janvier, 2002.

[29] E. NAUER. *DefineCrawler : un crawler paramétrable pour la recherche d'information intelligente sur le Web.* in « Journées scientifiques Web sémantique, Paris, France », éditeurs J. CHARLET, P. LAUBLET, C. REYNAUD., octobre, 2002, http://www.lalic.paris4.sorbonne.fr/stic/octobre/programme0209.html.

Rapports de recherche et publications internes

[30] M. CADOT, A. NAPOLI. *Une optimisation de l'extraction d'un jeu de règles s'appuyant sur les caractéristiques statistiques des données.* Rapport de recherche, numéro A02-R-162, LORIA, octobre, 2002.