INRIA

INSTITUT NATIONAL DE RECHERCHE EN INFORMATIQUE ET EN AUTOMATIQUE

Avant-projet Orpailleur

Représentations, raisonnements, et extraction de connaissances dans les bases de donneées

Nancy

THÈME 3A

Table des matières

ORPAILLEUR est un avant-projet du LORIA (UMR 7503) commun au CNRS, à l'INRIA, à l'Université Henri POINCARÉ *Nancy 1, à l'Université Nancy 2 et à l'Institut National Polytechnique de Lorraine.*

1 Composition de l'équipe

Responsable scientifique

Amedeo Napoli [(CR CNRS)]

Responsables permanents

Jean Lieber [(MdC, Université Henri Poincaré — Nancy 1)]
Florence Le Ber [(CR INRA et LORIA)]
Jean-François Mari [(Professeur, Université de Nancy II)]

Assistante de projet

Jamila Merikhi [(à temps partiel)]

Personnel Inria

Yannick Toussaint [(CR INRIA)]

Chercheurs doctorants

Rim Al Hulou [(doctorante, bourse co-financée Syrie – INRIA)]
Sandra Bérasaluce [(doctorante avec co-encadrement), bourse MENRT)]
Benoît Bresson [(doctorant, thèse CNAM)]
Fairouz Chakkour [(doctorante, bourse co-financée Syrie – INRIA)]
Hacène Cherfi [(doctorant, bourse co-financée Région – INRIA)]
Emmanuel Nauer [(doctorant–ATER, Université de Metz)]
Arnaud Simon [(doctorant–ATER, IUT de Strasbourg-Sud, Illkirch)]

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 être représentée et manipulée de façon adéquate par des procédures de raisonnement.

La philosophie du projet Orpailleur se trouve dans cette introduction. Le but des membres du projet est de concevoir des systèmes intelligents mettant en œuvre des connaissances pour

résoudre des problèmes. Ces systèmes intelligents sont multi-formes et sont appelés à fonctionner dans différents domaines d'application, aux premiers rangs desquels se trouvent l'agronomie, l'analyse de textes scientifiques et techniques, la bibliométrie, la chimie (planification de synthèses organiques), la classification de signaux temporels, la médecine, la muséologie, et la sidérurgie.

3 Fondements scientifiques

3.1 Systèmes à bases de connaissances, représentation des connaissances et raisonnements

Mots clés : systèmes à bases de connaissances, représentation des connaissances, objets, classes, hiérarchies, héritage, composition, systèmes classificatoires, logiques de descriptions, structures ordonnées, graphes, treillis, représentation de l'espace, relations spatiales, raisonnement par classification, raisonnement à partir de cas, généralisation, apprentissage à partir d'échecs.

Participants : Rim Al Hulou, Sandra Bérasaluce, Benoît Bresson, Fairouz Chakkour, Hacène Cherfi, Jean Lieber, Florence Le Ber, Jean-François Mari, Amedeo Napoli, Emmanuel Nauer, Arnaud Simon, Yannick Toussaint.

Résumé : *Dans le cadre de la conception de SBC, nous nous intéressons essentiellement aux systèmes de RCO (représentation de connaissances par objets), aux systèmes classificatoires et aux logiques de descriptions. Ces systèmes s'appuient sur une hiérarchie de classes ou de concepts instanciables qui sont organisées en hiérarchie(s) par l'intermédiaire d'une relation d'ordre partiel (la spécialisation ou la subsomption). La hiérarchie des classes peut être consultée pour résoudre des problèmes par l'intermédiaire de procédures (approche procédurale) ou de mécanismes de raisonnement comme la classification de classes ou d'instances (approche déclarative). Un ensemble d'assertions décrivant les faits dans lesquels interviennent les classes et leurs instances — instanciations de classes et instanciations de relations entre classes — peut compléter la représentation de l'univers étudié.*

3.1.1 Systèmes classificatoires et raisonnements

Appréhender un système de RCO comme un système logique a donné naissance à la théorie des systèmes classificatoires, qui s'appuie sur les développements théoriques réalisés dans le cadre des logiques de descriptions. Les opérations principales qui sont à la base du raisonnement sont :

- le test de subsomption qui consiste à vérifier qu'une classe `C` est plus générale qu'une classe `D`,
- la classification de classes qui consiste à placer une nouvelle classe `X` dans une hiérarchie $\mathcal{H}$; la classification d'instances qui consiste à déterminer les classes dont un objet `x` donné

peut être une instance (en particulier, une classe `C` n'est satisfiable que si elle peut avoir effectivement des instances),
- la recherche de propriétés qui consiste à retrouver les propriétés détenues par une classe ou une instance.

Le raisonnement par classification proprement dit s'appréhende comme une procédure de déduction opérant sur une hiérarchie. Sa mise en œuvre repose sur un cycle comprenant trois étapes :

- initialisation (création d'un nouvel objet `x` à classer),
- classification (recherche de la position de `x` dans la hiérarchie),
- mise en place de `x` dans la hiérarchie et exploitation de cette mise en place (ce qui peut ramener le cycle à sa première étape).

Le RÀPC (raisonnement à partir de cas) 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.

3.1.2 Systèmes à bases de connaissances pour l'étude des organisations spatiales agricoles

Dans ce cadre, le travail de recherche est effectué en collaboration avec des agronomes de l'INRA (Centre de Nancy). Il porte sur l'étude des formes d'organisation spatiale de l'agriculture. Trois projets ont été développés ou sont en cours de développement :

- interprétation d'images satellitaires
- simulation d'organisations spatiales
- exploitation d'une base d'enquêtes (cartes et explications).

Le premier projet, interprétation d'images satellitaires, a fait l'objet de la thèse de L. Mangelinck en 1998. Les principaux résultats ont porté sur la représentation dans un système de RCO de structures spatiales définies comme des ensembles d'entités spatiales reliées entre elles par des relations spatiales qualitatives. Dans le courant de l'année 1999, nous avons progressé sur la représentation des relations topologiques dans les systèmes de RCO : elles sont représentées par des classes, munies de propriétés et de méthodes de calcul sur l'image, et organisées sous forme de treillis. Différents treillis ont été étudiés.

Le deuxième projet, simulation d'organisations spatiales, n'a pas progressé durant l'année 1999, dans l'attente de résultats sur les connaissances agronomiques (cf. fouille de données *Ter Uti*).

Le troisième projet débute. Dans le cadre du DEA d'E. Kaboré, nous nous sommes intéressés à la similarité entre structures spatiales : les structures sont représentées par des graphes, et la similarité s'établit à partir de la recherche des plus grands sous-graphes communs entre deux graphes et d'un coût de généralisation associé à chaque sous-graphe.

3.2 Extraction de connaissances dans les bases de données

Mots clés : systèmes à bases de connaissances, représentation des connaissances, objets, classes, hiérarchies, héritage, composition, systèmes classificatoires, logiques de descriptions, structures ordonnées, graphes, treillis, représentation de l'espace, relations spatiales, raisonnement par classification, raisonnement à partir de cas, généralisation, apprentissage à partir d'échec.

Participants : Rim Al Hulou, Sandra Bérasaluce, Benoît Bresson, Fairouz Chakkour, Hacène Cherfi, Jean Lieber, Florence Le Ber, Jean-François Mari, Amedeo Napoli, Emmanuel Nauer, Arnaud Simon, Yannick Toussaint.

3.2.1 ECBD symbolique

L'extraction de connaissances à partir 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 d'analyse de données symboliques pouvant s'appuyer sur des techniques d'induction, de classification (par treillis), les arbres de décision, éventuellement couplé à un système d'analyse de données numériques et de statistiques,
- une interface se chargeant des interactions et de la visualisation des résultats intermédiaires et finaux.

La classification par treillis et les arbres de décision font partie des techniques utilisées dans le groupe Orpailleur pour la conception d'outils de fouille de données symbolique. Les arbres de décision permettent d'extraire des règles de classification des données. La classification par treillis — sur la base des treillis de Galois — permet de construire des treillis de classes à partir des données. Les classes sont organisées en une structure hiérarchique qui met en évidence une certaine organisation des données. La structure obtenue permet de mettre à jour des règles de classification mais aussi des règles d'association décrivant des corrélations entre les données ainsi qu'entre les propriétés qui décrivent les données.

Il faut faire émerger deux points importants : le premier est qu'un système d'ECBD vise à traiter des bases de données volumineuses et évolutives ; le second est l'utilisation de 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 exploitées telles quelles ou encore être mise à jour lors de l'arrivée de nouvelles données. Pour répondre à ces deux exigences, un système d'ECBD doit assurer la communication entre le système de gestion des bases de données et le système à base de connaissances.

3.2.2 ECBD numérique

Certains travaux de classification numérique en reconnaissance de la parole peuvent être adaptés et réutilisés pour étudier la classification de données temporelles ou spatiales, par exemple pour traiter des données issues d'un processus industriel comme les caractéristiques de tôles laminées par un train à bande ou les successions de cultures sur une parcelle géographique donnée. Dans ces deux domaines d'application, des outils à base de modèles stochastiques — comme les modèles de Markov cachés d'ordre 1 ou 2 — développés initialement pour la reconnaissance de la parole et l'identification du locuteur sont utilisés. Ces recherches en ECBD, d'une nature particulière et originale, visent à accroître le coté générique des outils de reconnaissance en investissant un domaine de recherche plutôt vierge. Elles constituent aussi un bel exemple d'inter-disciplinarité.

L'émergence des techniques stochastiques est principalement due à l'apparition de nouveaux serveurs de calcul puissants. Beaucoup d'hypothèses simplificatrices ont été posées dans les années 1980 pour implanter des algorithmes d'apprentissage et de reconnaissance ; l'utilisation de chaînes de Markov du premier ordre est la plus connue. Nous nous intéressons à l'utilisation de modèles stochastiques d'ordre supérieur comme les modèles de Markov d'ordre 2 qui permettent une meilleure prise en compte des durées des suites d'états stationnaires et transitoires.

Reconnaissance de successions culturales

Nous avons étudié les successions culturales pratiquées en Lorraine afin d'intégrer cette connaissance dans un modèle de simulation d'organisations spatiales agricoles en cours de développement à l'INRA. Pour réaliser cette étude, nous utilisons des données *Ter Uti* qui constituent un relevé de l'utilisation du territoire depuis une vingtaine d'années. Ces données sont traitées avec des algorithmes d'apprentissage développés au LORIA pour la reconnaissance de la parole. Ces algorithmes s'appuient sur les modèles de Markov cachés d'ordre 1 et 2 qui permettent de représenter des observations temporelles comme des successions d'états où les transitions entre états dépendent, suivant l'ordre du modèle, de l'état courant et des n états précédents. Simultanément, différents outils de visualisation ont été développés pour faciliter l'appropriation par les experts des résultats de la fouille.

Application au domaine sidérurgique

Ce travail de recherche s'inscrit dans le domaine de la fouille de données sidérurgiques et, plus particulièrement, de données issues d'un train à bandes. Il est mené en collaboration avec

l'IRSID et a été retenu dans le cadre de l'appel d'offres CNRS « Modélisation et simulation numérique ».

Le laminage est vu comme un processus stochastique dans lequel une suite de mesures effectuées sur le train à bandes va permettre de retrouver la suite d'états la plus probable, donc de prédire l'apparition d'un défaut quand celui-ci est caractérisé par une sous-séquence d'états. Les phénomènes physiques qui régissent la dégradation des cylindres du laminoir ne sont pas connus avec précision et la suite des mesures effectuées périodiquement sur le laminoir est entachée d'erreurs. Les états du laminoir sont cachés et en première approximation constituent une chaîne de Markov à mémoire limitée.

3.3 Fouille de textes et analyse de l'information scientifique et technique

Mots clés : information scientifique et technique, informatique linguistique, terminologie, interprétation, fouille de textes, synthèse de textes, classification, logiques de descriptions, treillis..

La fouille de données textuelles consiste en l'analyse d'un volume important de documents textuels pour fournir à l'utilisateur une vision synthétique et interprétable de leur contenu. Nous orientons nos travaux dans deux directions complémentaires : une approche informatique linguistique du texte dont le but est d'extraire du texte des structures conceptuelles ; une approche basée sur les connaissances et le raisonnement pour réaliser la tâche de fouille de données proprement dite sur les structures extraites.

Une première étape dans la collaboration avec l'INIST (le projet ILC) et dans le projet ILIAD du GIS Sciences de la Cognition nous a permis de mettre en place une plate-forme expérimentale, actuellement opérationnelle, d'analyse de l'information basée sur le repérage de termes et de leurs variations, associée à des méthodes de classifications statistiques (SDOC). Toujours en lien avec l'INIST, l'exploitation de cette plate-forme a fait l'objet d'une thèse [3] portant sur l'étude des groupes nominaux complexes et leurs propriétés, et sur l'observation d'un certain nombre de régularités dans le comportements de couples de prépositions dans les textes scientifiques. Il nous est ainsi devenu possible de mieux cerner les phénomènes syntaxiques sur lesquels nous voulons nous focaliser actuellement.

Cette année, nous avons commencé un travail exploitant des résultats précédents pour extraire des structures prédicatives – le verbe relié à ses arguments par des relations actancielles – dans des textes. L'objectif est de mettre en œuvre le raisonnement à partir de cas pour extraire ces structures. Les données ainsi extraites dans les textes sont plus riches sémantiquement que les listes de termes extraites dans le cadre d'ILIAD et elles seront utilisées en entrée des modules de fouille de textes.

La fouille de données textuelles se base sur une représentation conceptuelle des structures prédicatives et leur classification dans une logique de descriptions. La thèse de Nicolas Capponi [2] propose une méthode de généralisation inductive de structures prédicatives dédiée à l'analyse de l'information. Elle est basée sur la logique de descriptions CLASSIC. Nous souhaitons poursuivre ces travaux en proposant une méthode de construction de synthèse (au sens résumé) conceptuelle de textes qui puisse être exploitée pour comparer deux textes entre eux ou situer un texte par rapport à une synthèse globale d'un ensemble de textes.

Parallèlement, nous expérimentons la classification par treillis pour extraire des règles d'association entre termes. Cette méthode est en cours d'évaluation mais semble très prometteuse comparée avec les méthodes de classification statistiques utilisées dans la plate-forme ILIAD. L'utilisation des treillis rend la démarche incrémentale et il est donc possible d'augmenter le volume de documents traités ou les propriétés qui leurs sont attribuées. Le processus est traçable : il devient possible de prévoir les conséquence d'une modification d'un ensemble de documents ou de leurs propriétés, et d'envisager un « historique » de la classification et de la génération de règles. Enfin, cette nouvelle démarche permet de manipuler des objets structurés au sein d'une base de connaissances et de dépasser la simple notion d'ensemble de termes utilisée pour caractériser les documents dans le projet ILIAD.

3.4 Extraction de connaissances et aide à l'interrogation de bases de données chimiques

Mots clés : systèmes à bases de connaissances, extraction de connaissances, bases de données, objets, classification, recherche d'information, interrogation et navigation dans des bases de données, apprentissage, grandes bases de données, bases de données à objets, chimie organique.

Participants : Sandra Berasaluce, Jean Lieber, Amedeo Napoli.

Un travail de thèse sur l'extraction de connaissances et l'aide à la l'interrogation et la navigation dans des bases de données de chimie organique vient de débuter (thèse de Sandra Bérasaluce, co-dirigée par Claude Laurenço au CCIPE et LIRMM de Montpellier).

L'extraction de connaissances dans des bases de données est une question actuelle qui se pose avec insistance en chimie organique. Il existe des bases de données publiques qui portent sur plus de 18 millions de substances décrites — avec leurs propriétés chimiques, physiques et biologiques — et sur plus de 10 millions de réactions. L'interrogation de ces bases de données est le plus souvent difficile et frustre ; conçue pour répondre à des besoins de documentation plus que pour aider à la résolution de problèmes, elle se fait avec des moyens plutôt limités et sans aucune corrélation avec, par exemple, un système à base de connaissances d'aide à la synthèse.

Ainsi, il est possible actuellement de retrouver — dans les bases de données de réactions qui sont disponibles sur le marché — la plupart des réactions qui forment un composé donné ou qui atteignent un objectif donné, comme celui de construire un cycle à 7 chaînons par exemple. Cependant, ces bases étant des collections de réactions particulières indépendantes les unes des autres, il est très difficile voire impossible d'obtenir des informations sur des méthodes générales de synthèse relatives à une famille chimique donnée. Par exemple, une requête sur la formation des cycles à 7 chaînons fournira plusieurs milliers de réponses, volume trop important pour que soient possibles l'exploitation manuelle et l'interprétation de ces réponses, en vue de répertorier les différentes méthodes générales dont les réactions retrouvées sont des exemples.

L'idée est donc d'utiliser des techniques d'ECBD— comme la classification conceptuelle, la classification par treillis et la construction d'arbres de décision — pour découvrir des régularités dans les données, par exemple faire émerger des schémas réactionnels génériques à partir de descriptions de réactions spécifiques. Ces techniques doivent permettre d'extraire des unités

de connaissances réutilisables dans un système à base de connaissances, mais aussi d'optimiser les requêtes et l'indexation dans les bases de données. À plus long terme, ce travail s'inscrit dans le cadre des recherches sur la combinaison de bases de connaissances et de bases de données, qui est un passage obligé pour la mise au point de systèmes modernes de traitement de l'information, en particulier pour l'aide à la résolution de problèmes de synthèse en chimie organique.

Ce travail de recherche est polyvalent et revêt un ensemble d'intérêts théoriques et pratiques en informatique et en chimie. Parmi ces intérêts, il faut mentionner pour l'informatique : la gestion de bases de données (traitement de requêtes, indexation), l'extraction de connaissances dans les bases de données, la représentation de connaissances et le raisonnement ; pour la chimie : la modélisation de réactions et leur représentation sous différents points de vue.

3.5 Systèmes intelligents de traitement de l'information

Mots clés : systèmes d'information intégrés, données semi-structurées, navigation intelligente sur le Web, grandes bases de connaissances, grandes bases de données, bases de données à objets, bases de données distribuées.

Le but de travail de recherche est d'étudier la combinaison de la technologie des systèmes à bases de connaissances et de la technologie du Web ; ceci afin de fournir une aide efficace lors de la consultation du Web. Notre approche consiste plus précisément à utiliser des données structurées d'un domaine (références bibliographiques, thésaurus, etc.) pour faire émerger des connaissances sur ce domaine (réseaux d'auteurs, vocabulaire employé par tel ou tel auteur, etc.). Ces connaissances sont alors exploitées pour la recherche ou le filtrage d'information sur le Web. L'accès aux données du Web est réalisé via les moteurs de recherche classiques (AltaVista, Excite, etc.) qui sont utilisés comme des outils distants ; les connaissances émergeantes devant servir à guider l'utilisateur (i) en amont des moteurs de recherche pour la détermination de requêtes, pour l'expression du besoin, etc. et (ii) en aval pour l'évaluation et une présentation plus intelligible des documents.

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.). Les moteurs de recherche sont débordés par l'explosion du Web et ne répondent plus aux tâches de recherche d'information. Nos travaux s'inscrivent dans la lignée des agents dits « intelligents » qui ont vu le jour depuis quelques années. Ces agents tentent de pallier les problèmes rencontrés par les moteurs en combinant des techniques issues de différents domaines tels que la représentation de connaissances et le raisonnement, la linguistique, les bases de données, la recherche d'information, les statistiques, etc.

L'intégration et le traitement d'informations provenant de sources variées et hétérogènes sont des problèmes préalables à la construction de systèmes de fouille de données sur le Web (en particulier). À partir de ces données hétérogènes, nous souhaitons modéliser le domaine — déterminer les concepts du domaine, les liens entre les concepts et le vocabulaire attenant aux concepts — en utilisant un système de représentation capable d'effectuer des raisonnements sur ces données. De ce fait, il est indispensable de maîtriser l'hétérogénéité du vocabulaire utilisé et de sa portée sémantique [13]. La construction d'un vocabulaire de base est donc un

point essentiel dans notre approche. Le fait que les données soient hétérogènes et plutôt non régulières nécessite aussi de s'intéresser au traitement de *données semi-structurées* (DSS) [5] et à l'intégration d'informations provenant de sources différentes.

Les données semi-structurées sont des données hétérogènes, non régulières, sans format fixe bien déterminé qui décrive leur structure et leur organisation. De telles caractéristiques rendent difficile voire impossible la manipulation de telles données par des systèmes de gestion de bases de données (SGBD) classiques sans autre extension ou modification.

Une des options prises dans le groupe Orpailleur pour prendre en compte et manipuler des données semi-structurées (essentiellement textuelles) consiste tout d'abord à les décrire avec le langage de description de données textuelles XML [5]. Les caractéristiques et les fonctionnalités de XML le rendent particulièrement bien adapté à la description de DSS. La mise en œuvre de raisonnements sur de telles données — par exemple pour des besoins de résolution de problèmes ou d'ECBD — est ensuite dévolue à un système de RCO, où émerge la notion de classe *polythétique*, classe qui peut se définir à la fois par des disjonctions et des conjonctions d'attributs, par opposition aux conventionnelles conjonctions d'attributs [4]. Les données semi-structurées sont alors transformées en objets semi-structurés et et sont manipulées par des processus de classification adaptés, à la prise en compte de disjonctions entre autres [5].

Un autre enjeu est la production d'*information élaborée* qui donne une idée synthétique d'un ensemble de données et qui puisse être utilisée — à l'image des connaissances — dans un raisonnement. Concrètement, la production d'informations élaborées peut se voir comme un processus d'ECBD qui agit sur des informations scientifiques et techniques (IST) ; elle peut consister à chercher par exemple à dégager les principaux thèmes de recherche sous-jacents à un corpus de références bibliographiques, ou encore les collaborations entre auteurs, l'émergence d'une technique bien particulière, etc. Nous touchons en cela au domaine de la bibliométrie qui fixe les bases d'exploitation de l'IST. Là aussi, une normalisation minimale des données à exploiter est indispensable pour éviter des biais statistiques [11].

Un dernier enjeu concerne le traitement distribué de l'information dans une architecture client-serveur. Les données sont naturellement distribuées — notamment sur le Web — sur des sites distants et il est particulièrement important de pouvoir tenir compte de cette distribution et de la maîtriser. Les problèmes actuels de normalisation, d'échanges de formats et de gestion des accès aux données sont dans une phase préliminaire d'étude [9].

Les travaux menés jusqu'à présent concernent l'intégration des données (locales ou distantes, multi-sources et de natures différentes) et la construction d'informations élaborées à partir de ces données. Un système d'investigation sur le Web a déjà été mis en place. Ce système permet un accès indifférencié aux données locales ou distantes, ainsi que des croisements entre ces données [14] [12]. Concrètement, nous avons développé un ensemble de modules permettant (i) de réaliser des opérations fondamentales d'intégration de données telles la normalisation des données manipulées, la convergence du vocabulaire, la suppression des doublons dans les données, etc. ; (ii) d'accéder à l'ensemble des données par une interface Web (génération dynamique de pages HTML).

D'un point de vue pratique, une collaboration avec des chercheurs de l'INRS (Institut National de Recherche et Sécurité) a orienté notre contexte d'application vers le domaine médical. Ce domaine constitue un terrain particulièrement propice aux expérimentations, du fait qu'il s'avère riche en fonds structurés (bases de données, thésaurus, etc.) et en données en ligne.

L'utilisation et le croisement de données structurées et hétérogènes pour la construction d'un système intelligent de recherche d'informations ont permis :

- La structuration du domaine pour un accès hiérarchisé à l'information : des accès thématiques sont construits automatiquement par des méthodes de classification (à partir de descripteurs de références bibliographiques par exemple).
- La traduction de termes pour un accès multilingue : une traduction automatique du vocabulaire du domaine peut être effectuée via un thésaurus [10] ou encore par des corrélations de descripteurs au travers de références bibliographiques multilingues.
- La génération d'un environnement d'investigation spécialisé (et intégré) sur le Web permettant à l'utilisateur d'être assisté dans l'étape consistant à définir le vocabulaire de la requête à soumettre à un moteur de recherche (pour une recherche d'information sur Internet), ou encore d'obtenir des compléments d'informations (références bibliographiques locales) sur les documents du Web retrouvés.
- Le filtrage d'information sur Internet : à partir des critères sélectionnés par l'utilisateur, une requête est générée automatiquement et est soumise aux moteurs de recherche. Cette requête est précisée par l'ajout d'un contexte de recherche (vocabulaire proche) aux critères sélectionnés.

Il s'agit maintenant de déterminer aussi quelles connaissances vont permettre l'émergence de documents pertinents. Ces connaissances vont servir à favoriser la recherche d'information sur le Web (formulation automatique de requêtes). Elles devront également permettre d'analyser (valider, rejeter, juger, etc.) et de classer les documents, proposés en réponse par les moteurs de recherche. Dans ce but, la classification et le raisonnement ont un rôle essentiel à jouer. Il devient alors nécessaire de prendre en compte ces documents (textuels) hétérogènes, de les coder dans un formalisme de représentation pour être en mesure d'effectuer des raisonnements : par exemple, traiter des requêtes analogues, reconnaître qu'une requête est plus générale qu'une autre, classifier des requêtes, etc. Les résultats de ces travaux peuvent être étendus à la gestion de grandes bases de connaissances et de grandes bases de données.

Ces travaux prennent actuellement part, dans le cadre de l'action Escrire, à des collaborations avec les projets ACACIA et SHERPA de l'INRIA. Le but de cette action est d'étudier différents types de représentation des connaissances pour la gestion de documents scientifiques et techniques.

4 Logiciels

4.1 Le logiciel RÉSYN

Participants : Sandra Bérasaluce, Jean Lieber [correspondant], Amedeo Napoli.

Résumé : *Le système Y3 est écrit en Le-Lisp et nous sert de plate-forme d'expérimentation pour la plupart des développements réalisés dans le cadre des systèmes de RCO. En particulier, le système* RÉSYN/RÀPC *est développé en Y3 dans le cadre du* GDR CNRS *1093 « Traitement Informatique de la Connaissance*

en Chimie Organique » et a pour objet la planification de synthèses en chimie organique. La planification de synthèses s'appuie sur le raisonnement par classification et le RÀPC.

Le système RESYN *sert également de base dans le développement de méthodes d'*ECBD *pour l'aide à l'interrogation et la navigation dans des bases de données de chimie organique (bases de réactions et de produits). Le système* RESYN *est utilisé pour les connaissances qu'il renferme mais aussi pour ses capacités de résolution de problèmes de synthèse de molécules, en complément des systèmes de gestion des bases de données chimiques.*

4.2 PALETUVIER et CASIMIRPROTOCOLAIRE : représentation hiérarchique de connaissances

Participants : Benoît Bresson [correspondant], Jean Lieber, Amedeo Napoli.

Résumé : PALETUVIER *est un outil développé en* JAVA *pour la gestion de hiérarchies. Cet outil permet de créer et de mettre à jour une hiérarchie de concepts dans le cadre d'une application donnée, à partir de la description des concepts de cette application et de la relation d'ordre qui lie ces concepts. Des fonctionnalités de création de concepts primitifs ont aussi été développés, et un outil convivial pour visualiser des hiérarchies et interagir avec elles a été mis au point.*

CASIMIRPROTOCOLAIRE *est un système d'aide au traitement du cancer du sein qui s'appuie sur* PALETUVIER. *Ce système repose sur une représentation hiérarchique d'un protocole de traitement de ce type de cancers. Un concept de cette hiérarchie représente une catégorie de tumeurs. À certains de ces concepts sont associés des traitements. Classer une tumeur donnée dans cette hiérarchie permet ainsi d'indiquer les traitements de cette tumeur qui sont donnés par le protocole.*

4.3 Un logiciel pour l'interprétation des paysages

Participant : Florence Le Ber [correspondant].

Résumé : *Nous avons réalisé en Y3 un système de reconnaissance de modèles d'organisations territoriales agricoles à partir d'images satellites. 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 modèle 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.

4.4 Logiciel pour l'ECBD symbolique

Participants : Amedeo Napoli, Arnaud Simon [correspondant].

Résumé : *Une application médicale d'*ECBD *est menée conjointement avec le Registre Lorrain du Cancer de l'Enfant (*RLCE*, Hôpital d'Enfants de Nancy-Brabois). Cette application consiste à analyser le* RLCE*, en interaction avec un médecin spécialiste du domaine, dans le but d'avoir une meilleure compréhension des données du registre. Ces données ont surtout un caractère géographique, familial, et administratif. À terme, cette analyse doit conduire à une meilleure connaissance des malades et de leur environnement. Cette connaissance doit servir à améliorer la prise en charge et le suivi des patients, mais aussi l'accueil et le soutien aux familles.*

*Le système d'*ECBD *développé comprend principalement trois modules de fouille et deux modules de visualisation :*

- *Module de construction d'arbre de décision : il repose sur l'algorithme AlFReDO, pour « Algorithme de Fouille dans une Représentation des Données par Objets », qui utilise les techniques classiques de construction d'arbres de décision, ainsi que les principes de l'apprentissage par généralisation, dans le cadre de données représentées dans un système de RCO.*
- *Module de classification par treillis : il repose sur un algorithme incrémental de construction de points de vues s'appuyant sur la théorie des treillis de Galois, qui a la particularité d'exploiter les connaissances du domaine lors de la construction des points de vues. À un point de vue est associé un treillis dont la structure est utilisée afin d'extraire un ensemble de règles qui expliquent les données. Les points de vues sont ensuite élagués pour être stockés dans la base de connaissances du système lorsqu'ils sont validés par l'analyste.*
- *Module d'extraction de règles : il repose sur un algorithme d'extraction de règles s'appuyant sur la théorie des « ensembles approximatifs » (*rough sets*). Cet algorithme peut être utilisé seul pour extraire des règles qui expliquent les données ou couplé avec AlFReDO. Dans le second cas, des techniques propres aux ensembles approximatifs sont utilisées afin de réduire le nombre de descripteurs utilisés, ce qui permet d'améliorer les performances d'AlFReDO.*
- *Un module de visualisation permet de visualiser l'organisation des données ainsi que les résultats des différents algorithmes de fouille.*
- *Un module de cartographie adaptable à tout type de cartes est appliqué pour visualiser le point de vue géographique des données. Cet outil est utilisé pour mettre en évidence la répartition géographique des facteurs étudiés (essentiellement dans l'application associée au* RLCE*). La répartition obtenue est ensuite*

comparée à des cartes de répartition existantes pour faire apparaître d'éventuelles corrélations.

4.5 Logiciel pour l'ECBD numérique

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

Résumé : *Dans le cadre de la fouille de données dans les systèmes d'information géographique, nous avons développé des méthodes d'apprentissage et d'inférence fondées sur une modélisation à l'aide des chaînes de Markov d'ordre 2.*

*Les données, fournies par la Direction de la Recherche Agricole et Forestière (*DRAF*) lorraine, ont été traitées à l'aide d'outils provenant du monde de la recherche en reconnaissance de la parole.*

*En collaboration avec des experts agronomes de l'*INRA *(station* SAD *de Mirecourt), nous avons écrit un progiciel permettant la construction de modèles d'ordre quelconque, leurs apprentissages sur des données temporelles et spatiales. Nous avons particulièrement mis l'accent sur les outils de visualisation qui permettent aux experts agronomes d'évaluer les résultats de la modélisation et de s'approprier la connaissance mise en lumière.*

4.6 Les logiciels pour la fouille de texte

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

Résumé : *Outre AlFReDO présenté précédemment, les ressources, outils et environnements utilisés sont les suivants :*

- *Étiqueteur de Brill : 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 par notre équipe. Il met en œuvre des techniques d'apprentissage statistiques et probabilistes pour construire des règles lexicales et contextuelles utilisées ensuite pour l'étiquetage.*
- *Lemmatiseur du français : le lemmatiseur du français produit le lemme d'une forme fléchie (développé en collaboration de Fiammetta Namer, Université de Nancy 2).*
- *Classification statistique : issu des recherches menées à l'*INIST, SDOC *est un outil de classification statistique s'appuyant sur la méthode des mots associés et utilisant l'*indice d'équivalence.
- CLASSIC *: dans le cadre de la généralisation inductive, nous avons installé et exploité la logique de descriptions* CLASSIC.
- *Des corpus : le projet* ILIAD *nous a amené à constituer un corpus de textes. Ce sont environ 10 000 résumés d'articles scientifiques en provenance de la base Pascal de l'*INIST *qui ont ainsi été collectés.*

4.7 Les logiciels pour le traitement de l'information

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

Résumé : *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 « Moteur de Recherche, Filtrage et Classification d'Informations sur Internet », 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, actuellement en cours de développement, 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 dans le cadre de l'intégration de bases de données et la résolution de problèmes dans le domaine des données. Les données sont essentiellement des documents textuels, pour certains décrits en XML *(ou devant l'être). Dans un tel cadre, le langage* XML *sert de support à la description des documents tandis que la logique de descriptions* FACT *permet de mettre en œuvre des raisonnements (par classification) et d'exploiter des connaissances du domaine, ceci afin de résoudre des problèmes de natures variées : aide à la navigation et recherche d'informations dans des bases de données hétérogènes, résolution de problèmes se posant sur les requêtes elles-mêmes, comme la classification de requêtes et le traitement de requêtes analogues.*

5 Actions régionales, nationales et internationales

5.1 Actions régionales

5.1.1 Collaboration URI et Orpailleur

Participants : Rim Al Hulou, Dominique Besagni [INIST], Benoît Bresson, Fairouz Chakkour, Hacène Cherfi, Claire François [INIST], Luc Grivel [INIST], Jean Lieber, Florence Le Ber, Jean-François Mari, Bernard Maudinas [INIST], Amedeo Napoli, Emmanuel Nauer, Xavier Polanco [INIST], Ivana Roche [INIST], Jean Royauté [INIST], Arnaud Simon, Yannick

Toussaint.

La collaboration entre l'équipe URI (Unité de recherches et d'innovation) de l'INIST et le groupe Orpailleur a pour ambition de 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 rédaction commune d'articles scientifiques et la mise en œuvre opérationnelle des recherches dans le contexte de l'INIST.

Un séminaire commun sur la base d'une demi-journée mensuelle a été mis en place depuis l'automne 1998. 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 comme points particuliers l'extraction et l'analyse de structures prédicatives — groupes nominaux et verbaux — dans des textes scientifiques et techniques et l'utilisation du système NEURODOC dans le contexte de 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.

5.2 Actions nationales

5.2.1 GDR TICCO 1093 CNRS

Participants : Sandra Bérasaluce, Jean Lieber, Amedeo Napoli.

Le GDR CNRS 1093 TICCO — *Traitement informatique de la connaissance en chimie organique* — réunit des chercheurs en chimie organique du CCIPE à Montpellier, des chercheurs en informatique du LORIA et du LIRMM, et des industriels de l'industrie pharmaceutique comme Sanofi-chimie, Roussel-Uclaf, et l'Institut de Recherches Servier. L'objectif du GDR est l'étude et la mise en œuvre de systèmes de RCO, du raisonnement par classification et du RÀPC pour construire des systèmes d'aide à la planification de synthèses de molécules. Ce travail de recherches nécessite de travailler sur une représentation des objets de la chimie organique, une représentation des plans de synthèses de molécules, une modélisation et une représentation des raisonnements élémentaires et des stratégies de synthèse employés par les chimistes pour résoudre un problème de synthèse.

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

5.2.2 Projet GIS Casimir

Participants : Benoît Bresson, Jean Lieber, Amedeo Napoli.

Le projet GIS Casimir (Conception continue d'un savoir casuel) vise à élaborer un système qui fournisse une aide à la décision thérapeutique pour la prise en charge de malades souffrant d'un cancer du sein ainsi qu'une aide au suivi de l'évolution des règles d'actions prises pour soigner les malades. Ce projet s'articule autour de deux champs de recherches d'actualité : le raisonnement à partir de cas et la mémoire organisationnelle. La conception de ce type de mémoire est vue ici comme une activité de conception portant sur le savoir mis en œuvre, donc ici les protocoles de traitement. Cela suppose que le savoir préexiste et qu'il est conservé dans une mémoire. Un des objectifs premiers du projet Casimir est de collecter ce savoir puis de le représenter sous une forme informatique réutilisable (dans un système à base de connaissances par exemple). L'objectif de la construction d'une mémoire organisationnelle n'est pas seulement de collecter et d'expliciter les savoirs, mais aussi d'élaborer à partir de ces savoirs une réflexion sur l'activité fonctionnelle liée à ces savoirs, pour les analyser et les faire évoluer.

Pour l'instant, deux phases de ce travail ont eu lieu. La première est une étude théorique sur l'apprentissage à partir d'échecs, pour engendrer des explications devant servir dans les mises à jour des règles d'actions. Le protocole de traitement évolue : son utilisation à un instant donné peut conduire à des décisions erronées, car obsolètes, ou encore à des impasses, avec obligation d'adapter le protocole, compte tenu de l'état actuel des connaissances en cancérologie du sein. L'apprentissage à partir d'échecs, à travers une analyse de la décision erronée, permet de faire évoluer le protocole.

La seconde phase de travail est applicative : pour pouvoir raisonner avec le protocole et le faire évoluer, il faut le connaître et le représenter informatiquement. C'est l'application CASIMIRPROTOCOLAIRE, décrite précédemment et en cours de développement, qui permet de le faire ([6] donne une idée de l'état actuel du projet et son évolution).

5.2.3 Projet GIS Traces

Participants : Jean Lieber, Amedeo Napoli.

La thématique de la traçabilité des processus de conception et de réutilisation effective de traces de raisonnement est actuellement développée dans deux domaines différents : d'une part, en intelligence artificielle, et plus précisément dans le cadre du raisonnement à partir de cas, et d'autre part, en sciences cognitives. Dans ces deux domaines, les recherches portent sur le recueil, la modélisation et la gestion de connaissances de conception réutilisables. Ces recherches sont généralement menées en parallèle, sans réelle interaction entre les deux domaines. L'objectif et l'originalité du projet TRACES est de confronter des modèles formels et cognitifs développés dans les deux domaines afin d'engendrer les spécifications d'un outil d'aide à la réutilisation de processus de conceptions. Ces spécifications nécessitent l'étude de traces de raisonnement (du point de vue de l'intelligence artificielle et du RÀPC en particulier) mais aussi l'étude de modèles cognitifs des utilisateurs (ici l'utilisateur est un concepteur qui agit dans une situation de conception industrielle). Un but important du projet est de garantir une compatibilité entre les modèles de raisonnement formels du RÀPC, centrés sur la réutilisation,

et les modèles d'ergonomie cognitive. Cette compatibilité est en effet indispensable pour assurer une meilleure utilisation de systèmes visant à fournir à des concepteurs une aide effective dans leur activité.

5.2.4 Collaboration avec Béatrice Fuchs et Alain Mille

Participants : Jean Lieber, 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. Des modèles ont été proposés parallèlement dans l'équipe d'Orpailleur et dans l'équipe d'Alain Mille (équipe « Raisonnement à partir de cas », au LISA-CPE, Université de Lyon 1). Une collaboration s'est engagée avec Alain Mille et Béatrice Fuchs (maître de conférences à l'Université de Lyon 3), pour confronter ces modèles de l'adaptation et les enrichir par nos expériences respectives. Ce travail a conduit à un premier modèle [8, 7] et se précise actuellement. Le modèle présenté 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 ce même cadre, un séminaire spécialisé (*workshop*) portant sur sur la formalisation de l'adaptation en RÀPC, a été organisé par les membres de la collaboration, en liaison avec la conférence internationale sur le RÀPC [1].

5.2.5 Collaboration avec le Musée de La Villette

Participant : Arnaud Simon.

La Cité des Sciences et de l'Industrie de La Villette possède des collections muséologiques très riches d'objets relatifs à l'histoire des sciences et de l'industrie. Les objets des collections sont utilisés pour l'organisation d'expositions par la Cité des Sciences mais aussi par d'autres musées nationaux dans le cadre d'expositions temporaires. Une partie de ces objets est inventoriée dans une base de données relationnelle dans laquelle les possibilités d'accès aux objets se limitent à leurs numéros d'ordre. Les informations stockées dans la base concernent uniquement le suivi des restaurations et des prêts.

Le projet repose sur deux constats. D'une part les responsables d'expositions doivent pouvoir accéder à l'information des collections. D'autre part les collections sont une « mine » de connaissances relatives à l'histoire des sciences et de l'industrie. Ainsi, l'objectif de ce projet est de compléter la base de données existante par les descriptions détaillées des objets, de représenter les connaissances du conservateur, expert en histoire des sciences, grâce à un système de RCO. Le couplage entre la base de données et le système de RCO devrait assurer une meilleure gestion des objets de la base, en particulier pour le filtrage d'informations et la classification par points de vue, et permettre aussi une meilleure appréhension par les responsables d'expositions de l'organisation de la base et des concepts auxquels se réfèrent les objets. De plus, le couplage permettra d'utiliser le système d'ECBD développé chez Orpailleur pour découvrir des

corrélations entre les objets et mieux comprendre l'évolution des objets au cours du temps, ainsi que pour, éventuellement, faire émerger de nouveaux thèmes d'exposition.

À l'heure actuelle, le conservateur des collections à proposé une première étude portant sur une description détaillée de trente robots jouets. Nous avons représenté les différents concepts qui s'y rattachent. L'utilisation du système d'ECBD a permis, dans un premier temps, de mettre en évidence l'influence cinématographique et culturelle sur la morphologie et les fonctionnalités des robots jouets. De plus, d'après les utilisateurs potentiels de la base de données, les points de vue obtenus avec la classification par treillis mettent en évidence des concepts potentiellement utilisables comme thèmes d'exposition. Devant ces résultats prometteurs, la suite naturelle de ce projet est de compléter les informations relatives à l'ensemble des objets des collections et la représentation des connaissances liées au domaine.

5.2.6 Fouille de données – Application au domaine de la santé

Participants : Amedeo Napoli, Arnaud Simon.

L'objectif de ce projet est de concevoir un système interactif d'aide à l'analyse de données relatives aux cancers des enfants en Lorraine. Le projet s'appuie sur le fait qu'une exploitation des données du Registre Lorrain du Cancer de l'Enfant avec des méthodes d'ECBD (arbres de décision, classification par treillis dans le cadre d'un système de RCO) peut conduire à une meilleure compréhension des données et à la découverte de connaissances, implicites dans ces données. Ces connaissances peuvent ensuite être utilisées pour aider à une prise en charge et un suivi plus satisfaisants des malades.

5.2.7 ECBD et HMM

Participants : Florence Le Ber, Jean-François Mari.

Une application des modèles de Markov d'ordre 1 et 2 au domaine sidérurgique, plus précisément pour fouiller des données issues d'un train à bandes a été mise en œuvre. Cette application a été menée en collaboration avec l'IRSID. Cette application a été retenue dans l'appel d'offres CNRS « Modélisation et simulation numérique ».

Une autre application des modèles de Markov d'ordre 1 et 2 a été mise en œuvre pour la reconnaissance de successions culturales. Cette étude se propose d'étudier les successions culturales pratiquées en Lorraine depuis une dizaine d'années, afin d'intégrer cette connaissance dans un modèle d'organisation spatiale de territoires agricoles en cours de développement à l'INRA. Ce projet, après avoir été retenu en 1998 dans l'appel d'offres « Programme de systèmes d'information géographique » initié par le CNRS et l'IGN dans le cadre du GDR Cassini, s'est vu prolonger en 1999 dans le second appel d'offres du même programme.

5.2.8 ARC A3-ILEC de l'AUF (AUPELF-UREF)

Participants : Fairouz Chakkour, Hacène Cherfi, Yannick Toussaint.

Nous participons à la seconde phase de l'Action de Recherche Concertée ARC A3 sur la construction automatique de terminologies et de relations sémantiques entre termes à partir de

corpus. Notre objectif est de participer à l'évaluation des outils qui ont été développés dans le cadre du projet ILIAD, en comparaison avec d'autres approches. La première phase a concerné l'évaluation des extracteurs terminologiques, et la seconde, qui concerne les comparaison, vient tout juste de débuter.

5.2.9 ARC INRIA Escrire

Participants : Rim Al Hulou, Hacène Cherfi, Amedeo Napoli, Emmanuel Nauer, Yannick Toussaint.

Cet ARC INRIA se fait en collaboration avec le projet ACACIA (Rose Dieng, INRIA SOPHIA-ANTIPOLIS) et Jérôme Euzenat du projet SHERPA (INRIA RHÔNE-ALPES).

Un intranet, et plus généralement, l'utilisation des technologies de l'Internet, sont des opportunités pour les entreprises, d'accéder et de partager la connaissance bien souvent difficilement accessible sous forme documentaire. Les documents numériques et numérisés peuvent être rendus accessibles de manière standard et transparente auprès de tous les utilisateurs concernés. L'ambition, à terme, est de réaliser de véritables serveurs de connaissances permettant la recherche et la manipulation des ressources de l'entreprise.

Cependant, les limites de cette approche apparaissent rapidement : l'organisation des sites se révèle une tâche coûteuse et la recherche plein texte peu efficace. La recherche et l'interrogation d'un site en s'appuyant sur le contenu des documents est une nécessité et les formalismes de représentation de connaissances sont de bons candidats pour représenter ce contenu. La représentation du contenu peut permettre de manipuler ce contenu pour faire de la recherche par analogie, par spécialisation, par similitude, etc.

Cependant, il existe différents formalismes de représentation de connaissances et nul ne connaît exactement leurs qualités respectives. Le but de l'ARC Escrire consiste donc à comparer trois types de représentations de connaissances — graphes conceptuels (GC), représentations de connaissances par objets (RCO) et logiques de descriptions (LD) — du point de vue de la représentation du contenu de documents et de leur manipulation. Pour cela, l'action s'appuie sur les compétences dans chacune des représentations des projets ACACIA (GC), SHERPA (RCO) et Orpailleur (LD) respectivement. L'objectif de l'action consiste à comparer les apports de chacun des types de représentation pour la représentation du contenu dans les serveurs de connaissances.

La mise à l'épreuve de ces différents formalismes pour le traitement d'un jeu de documents (devant être fourni sans doute par un partenaire industriel) nécessite de mener une réflexion méthodologique sur le passage des textes à leur représentation formelle (de façon suffisamment indépendante des formalismes employés) en lien avec le type d'accès que l'on veut avoir sur ces documents. Cette représentation formelle sera définie conjointement et introduite dans un format XML (pour " eXtensible Markup Language "). Un ensemble de requêtes définies de manière coordonnée sera évaluée dans chacun des contextes.

À l'issue de ce travail, les différents formalismes seront comparés entre eux (mais aussi à la recherche plein-texte) selon le protocole prédéfini. Celui-ci devra apprécier des critères tant qualitatifs (expressivité des requêtes, accessibilité/lisibilité des informations, etc.) que quantitatifs (temps de réponse à une requête, taux de précision/rappel des réponses, etc.).

Cette évaluation proposera une grille d'analyse des avantages et inconvénients d'un langage de représentation formel vis-à-vis de la recherche d'informations sur le Web, et tentera de déterminer les contextes favorables à l'exploitation de chacune de ces représentations.

5.3 Actions internationales

5.3.1 Action Intégrée ECOS-CONICYT avec le Chili

Participants : Xavier Polanco [INIST], Jean Royauté [INIST], Yannick Toussaint.

En association avec l'équipe URI de l'INIST, nous avons proposé, en 1998, puis mis en place une Action Intégrée (PAI) avec deux universités chiliennes, la *Universidad de Concepción* (contacts : J. Atkinson et A. Ferreira) et la *Universidad de Chile* (contact : A. Bassi). La première année a été consacrée à la constitution de ressources sur l'espagnol : corpus de textes et lexiques. En 1999, nous en sommes à la deuxième année de coopération, ce qui nous a permis d'adapter à l'espagnol les modules linguistiques d'étiquetage et de lemmatisation de la plate-forme ILC, plate-forme d'analyse de l'information basée sur l'extraction de termes à partir de textes et sur la classification développée par les partenaires français en 1997–1998. La troisième année sera consacrée à l'intégration des outils et amorcera une discussion sur le type de sémantique pouvant être mise en œuvre dans la suite de nos travaux communs.

5.3.2 Action intégrée Balaton

Participants : Florence Le Ber, Amedeo Napoli.

Un système à objets pour la représentation et la manipulation de structures

Ce projet — PAI balaton — se fait en collaboration avec l'Université Kossuth Lajos à Debrecen (Hongrie), où notre contact est Katalin Bognar, enseignant-chercheur à l'institut de mathématiques et d'informatique.

Les objectifs scientifiques de ce projet sont de concevoir un système de RCO adapté à la représentation et à la manipulation de structures. Un tel système peut être utilisé pour la représentation de structures spatiales ou moléculaires par exemple. Les structures sont vues comme des objets composites, des objets ayant des composants sont eux mêmes des objets, les composants étant liés entre eux par des relations vérifiant certaines contraintes. À l'heure actuelle, il n'existe pas de langage de référence pour la représentation et la manipulation de structures. En partant de notre expérience sur les systèmes de RCO et la représentation et la manipulation de structures spatiales et moléculaires, notre but est de concevoir un système de représentation et de manipulation de structures. Une structure est considérée comme un graphe étiquetée dont les sommets et les arêtes sont représentées par des classes (munies d'attributs et de méthodes) dans l'univers d'une RCO. À chaque classe est associé un ensemble de relations, qui modélisent des contraintes inter-attributs. En particulier, la prise en compte et la mise en œuvre d'un tel ensemble de relations est une extension du modèle classique des représentations de connaissances par objets. La manipulation de telles structures peut se faire en utilisant le raisonnement par classification et un certain nombre de variantes de ce mode de raisonnement, qui doivent être étudiées en détail dans le cadre de cette action intégrée Balaton.

6 Diffusion de résultats

6.1 Animation de la Communauté scientifique

- Action intégrée avec le Chili et la Hongrie.
- Participation à des groupes de travail (GDR, PRC).
- Participation à des comités de lecture de revues, à la mise en œuvre de numéros spéciaux de revues et à l'édition d'ouvrages de recherche.
- Organisation de colloques et participation à des comités de programme.

6.2 Enseignement

- Enseignements et organisation scientifique de cours (en France et à l'étranger).
- Encadrements de thèses, DEA, stages DESS et d'IUT.
- Participation à des jurys de thèse en France et à l'étranger.

7 Bibliographie

Livres et monographies

[1] J. Lieber, E. Melis, A. Mille, A. Napoli (éditeurs), *Formalisation of Adaptation in Case-Based Reasoning*, Third International Conference on Case-Based Reasoning Workshop, ICCBR-99 Workshop number 3, S. Schmitt and I. Vollrath (volume editor), LSA, University of Kaiserslautern, 1999.

Thèses et habilitations à diriger des recherches

[2] N. Capponi, *Généralisation de structures prédicatives. Application à l'analyse de l'information*, Thèse d'informatique, Université Henri Poincaré (Nancy 1), 1999.

[3] J. Royauté, *Les groupes nominaux complexes et leurs propriétés : application à l'analyse de l'information*, Thèse d'informatique, Université Henri Poincaré (Nancy 1), 1999.

Communications à des congrès, colloques, etc.

[4] R. Al Hulou, A. Napoli, E. Nauer, « Objets semi-structurés, classes polythétiques et classification », *in : Actes des Septièmes journées de la Société Francophone de Classification, SFC'99*, F. Le Ber, J. Mari, A. Napoli, A. Simon (éditeurs), LORIA, p. 299–306, Nancy, septembre 1999.

[5] R. Al Hulou, A. Napoli, E. Nauer, « XML : un formalisme de représentation intermédiaire entre données semi-structurées et représentations par objets », *in : Actes de la confe'rence LMO'2K, Montreal*, C. Dony (éditeur), Hermès, 2000.

[6] B. Bresson and J. Lieber, « Classification pour l'aide au traitement du cancer du sein », *in : Septième journées de la Société Francophone de Classification, SFC'99*, F. Le Ber, J. Mari, A. Napoli, A. Simon (éditeurs), Unité de recherche INRIA Lorraine, p. 53–59, Nancy, septembre 1999.

[7] B. Fuchs, J. Lieber, A. Mille, A. Napoli, « Towards a Unified Theory of Adaptation in Case-Based Reasoning », *in : Case-Based Reasoning Reasearch and Development — Third International*

Conference on Case-Based Reasoning (ICCBR-99), K.-D. Althoff, R. Bergmann, L. K. Branting (éditeurs), *Lecture Notes in Artificial Intelligence 1650*, Springer, Berlin, 1999.

[8] B. Fuchs, J. Lieber, A. Mille, A. Napoli, « Vers une théorie unifiée de l'adaptation en raisonnement à partir de cas », *in : Actes des journées ingénierie des connaissances*, 1999.

[9] S. Jolibois, E. Nauer, D. Chouanière, J. Ducloy, F. Grandjean, M. Mouzé-Amady, « Adaptation des normes et formats documentaires à la gestion informatisée de corpus bibliographiques », *in : Bulletin des Bibliothèques de France*, 2000. À paraître.

[10] S. Jolibois, E. Nauer, D. Chouanière, J. Ducloy, F. Grandjean, M. Mouzé-Amady, « L'Unified Medical Language System (UMLS) : une base de connaissances multilingue dans le domaine biomédical », *in : Documentaliste - Sciences de l'information*, 2000. À paraître.

[11] E. Nauer, « De l'importance de la normalisation en bibliométrie », *in : Les systèmes d'information élaborée*, Société Française de Bibliométrie Appliquée, Revue Française de Bibliométrie, Ile Rousse, 1999.

[12] E. Nauer, « IciWeb : Système d'Intégration et de Croisement d'Information sur le Web », *in : Hypertextes, Hypermédias et Internet*, J.-P. Balpe, S. Natkin, A. Lelu, I. Saleh (éditeurs), Hermès, p. 333–337, Paris, 1999.

[13] E. Nauer, « Les problèmes de variations terminologiques dans l'indexation de références bibliographiques », *in : Journées Internationales de Linguistique Appliquée (JILA'99)*, LILLA - Université de Nice, juin 1999.

[14] E. Nauer, « Croisement de multiples sources d'information : utilisation du web comme source et interface », *in : Hypertextes et Hypermédias*, J.-P. Balpe, I. Saleh, M. Nanard (éditeurs), Hermes, Paris, 2000.