INRIA

INSTITUT NATIONAL DE RECHERCHE EN INFORMATIQUE ET EN AUTOMATIQUE

Action ORPAILLEUR

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

Lorraine

THÈME 3A

Table des matières

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, bourse co-financée Syrie – INRIA)]
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, bourse co-financée Syrie – INRIA)]
Hacène Cherfi [(doctorant, bourse co-financée Région – INRIA)]
Sébastien Hergalant [(DEA/doctorant, bourse co-financée INRA-Région)]
Jean-Luc Metzger [(doctorant, bourse co-financée INRA – INRIA)]

Chercheur post-doctorant

Rafik Taouil [(Collaboration Dyade-Bull-Orpailleur)]

Collaborateur extérieur permanent

Benoît Bresson [(Collaboration Orpailleur-CAV Nancy/Oncolor)]

Stagiaires

Tawfik Labib [(DESS Double Compétence)]
Sandy Maumus [(DESS Double Compétence)]

2 Présentation et objectifs généraux

Les thèmes de recherches de l'avant-projet Orpailleur portent principalement sur l'étude et la conception de systèmes intelligents : systèmes de connaissances, systèmes d'information, et systèmes d'extraction de connaissances dans les bases de données (ou encore systèmes de fouille de données). Ces systèmes intelligents sont multi-formes et sont appelés à fonctionner dans différents domaines d'application, parmi lesquels se trouvent la biologie, la chimie, l'espace, le temps et le Web.

Axes de recherche

- Représentation des connaissances et raisonnements par classification, à partir de cas, et spatio-temporel.
- Gestion des connaissances, systèmes d'information, et Web sémantique.

- Extraction de connaissances dans les bases de données (fouille de données) : méthodes symbolico-numériques, extraction de motifs fréquents, extraction de règles d'association, fouille de textes, modèles de Markov cachés d'ordre supérieur pour la fouille de données, implication statistique et ensembles approximatifs pour la fouille de données.
- Fouille de données pour la bioinformatique et pour la planification en synthèse organique.
- Développement de systèmes associés à ces problématiques.

Relations internationales et industrielles

- Participation à l'ARC INRIA Escrire (avec Exmo, UR Rhône-Alpes et ACACIA, UR Sophia-Antipolis).
- Collaboration suivie avec l'INRA.
- Partenariat avec le CAV (centre de lutte contre le cancer, Nancy) et le laboratoire d'ergonomie du CNAM à Paris dans le projet Kasimir.
- Participation au projet KVM : gestion des connaissances et de mémoires d'entreprises ; partenariat avec ECOO et MAIA au LORIA, et la société Kappa à Paris.
- Collaborations privilégiées avec les laboratoires LIRMM (Informatique Montpellier) et LSIC (informatique chimique Montpellier).
- Collaboration privilégiée avec l'équipe d'études sur le raisonnement à partir de cas du LISI, Lyon.
- Collaboration avec le musée de La Villette (Paris) pour la gestion et la mise en place d'expositions.
- Collaboration privilégiée avec l'équipe URI de l'INIST (recherches sur l'information scientifique et technique, et la bibliométrie).
- Partenariat avec Bull pour le projet Supersélect.
- Relation internationale avec l'université de Debrecen (Hongrie).

3 Fondements scientifiques

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.

Les thèmes de recherches dans l'avant-projet Orpailleur portent principalement sur l'étude et la conception de *systèmes intelligents* : systèmes (à bases) de connaissances, systèmes d'information, et systèmes d'extraction de connaissances dans les bases de données (ECBD ou encore fouille de données). Ces systèmes intelligents sont multi-formes et sont appelés à fonctionner dans différents domaines d'application, parmi lesquels se trouvent le raisonnement spatial et la gestion spatiale du territoire en agronomie, la cancérologie, la bibliométrie, la fouille et l'analyse de textes scientifiques et techniques, la chimie et la synthèse organique, sans oublier le Web,

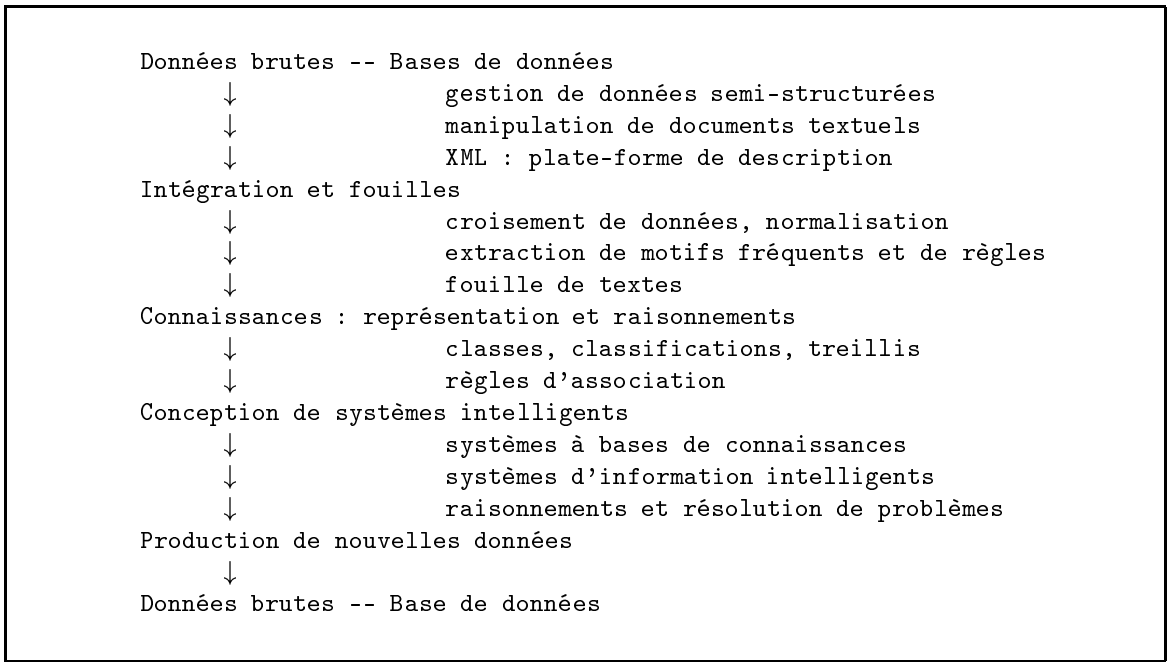

FIG. 1 – Des données aux connaissances ou l'articulation des recherches de l'avant-projet Orpailleur.

et particulièrement le *Web sémantique*. La biologie et le génome — la fouille et l'analyse de séquences génomiques, la bioinformatique plus généralement — deviennent un nouveau terrain d'investigation, qui prend une place de plus en plus importante pour les membres d'Orpailleur.

La figure 1 montre comment s'articulent entre elles les recherches effectuées dans l'équipe Orpailleur, en particulier comment peut s'effectuer le passage de données brutes à des connaissances exploitables dans un système intelligent. Les données brutes sont hétérogènes et peuvent provenir de sources très diverses : bases de données diverses, documents textuels, Web, capteurs, etc., ce qui nécessite de savoir intégrer, traiter et gérer des données semi-structurées. Pour les documents textuels, le langage XML peut servir de plate-forme de description et permet d'appréhender une bonne partie des irrégularités des données textuelles. Une fois les données intégrées et traduites dans un format adéquat, les processus de fouilles peuvent être appliqués pour faire émerger des éléments de connaissances potentiellement exploitables dans un système intelligent. Les processus de fouille s'appuient principalement sur la recherche de motifs fréquents — groupes de propriétés apparaissant dans les données avec une fréquence supérieure à un seuil donné —, la classification par arbres de décision et par treillis, l'extraction de règles d'association. Les processus de fouille exploitent également un modèle du domaine des données — une ontologie — intégré à la base de connaissances du système intelligent, en l'occurrence un système de représentation des connaissances par objets. Les éléments de connaissances nouveaux sont alors utilisés pour compléter la connaissance du système intelligent. Ils peuvent alors à leur tour être manipulés par le raisonnement par classification ou par le raisonnement à partir de cas, afin d'intervenir dans la résolution de problèmes sur le

domaine des données. De nouvelles données peuvent alors être produites, qui alimentent une base de données, qui peut à son tour être fouillée, et la boucle de traitement est bouclée.

Le ciment qui donne toute sa cohérence à cette colonne vertébrale et qui permet la circulation entre données et connaissances est la *classification*. L'activité de classification intervient à tous les niveaux dans le traitement intelligent des données et des connaissances. Les classes et la classification servent à organiser un domaine, à le comprendre, à représenter et à manipuler les entités de ce domaine pour pouvoir résoudre des problèmes. À l'origine, les classes et la classification étaient essentiellement utilisées en analyse de données pour des besoins plutôt numériques. Depuis, les classes et la classification ont pris une place prépondérante dans le champ de l'intelligence artificielle, et plus particulièrement dans les thèmes étudiés dans l'avant-projet Orpailleur : les membres d'Orpailleur ont beaucoup œuvré pour cela, et ils continuent à le faire.

3.1 Systèmes de connaissances, représentation des connaissances et raisonnements

Mots clés : systèmes de connaissances, représentation des connaissances (par objets), systèmes classificatoires, logiques de descriptions, structures ordonnées, représentation de l'espace, treillis de relations spatiales, raisonnement par classification, raisonnement à partir de cas, manipulation intelligente de documents, Web sémantique.

Participants : Rim Al Hulou, Sandra Berasaluce, Benoît Bresson, Fairouz Chakkour, Hacène Cherfi, Florence Le Ber, Jean Lieber, Jean-François Mari, Jean-Luc Metzger, Amedeo Napoli, Emmanuel Nauer, 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 concepts) instanciables qui sont organisées en hiérarchie(s) par l'intermédiaire d'une relation d'ordre partiel (spécialisation ou 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 — complète 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 vérifie 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 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) 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 liées à un thème qui est développé ci-après, la mémoire organisationnelle, et plus généralement la gestion de connaissances (voir les projets KASIMIR et KVM par exemple). La conception d'une mémoire organisationnelle est vue comme une activité de conception qui met en jeu des connaissances sur le domaine étudié, qui suppose que de telles connaissances préexistent et qu'il est possible de les conserver dans une mémoire. L'objectif de la construction d'une mémoire organisationnelle n'est pas seulement de collecter et d'expliciter les connaissances, mais aussi d'élaborer à partir de ces connaissances une réflexion sur l'activité fonctionnelle liée à ces connaissances, pour les analyser et les faire évoluer.

3.1.2 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. Ces recherches s'appuient sur des applications dans le domaine agricole, en collaboration avec des équipes de l'INRA SAD (département Systèmes Agraires et Développement). Trois projets ont été développés ou sont en cours de développement :
- classification de structures spatiales pour l'interprétation d'images satellitaires,
- simulation d'organisations spatiales,
- raisonnement à partir de cas pour l'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 fin 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. Nous avons travaillé sur la représentation des relations

topologiques dans les systèmes de RCO et sur leur organisation sous forme de treillis. L'année 2001 a permis de poursuivre la synthèse de ce travail par des publications, et d'approfondir certains aspects théoriques, en ce qui concerne le calcul de relations qualitatives sur un domaine discret et l'usage des treillis de relations pour le raisonnement spatial.

Le deuxième projet, simulation d'organisations spatiales, s'oriente vers la réécriture d'un modèle multi-agents d'allocation de surfaces plus général, qui permettrait d'intégrer des informations telles que celles issues de la fouille de données *Ter Uti* (représentation de la dynamique des organisations spatiales). Une collaboration en ce sens est en cours avec MAIA.

Le troisième projet a débuté en 99 dans la cadre du stage de DEA d'E. Kaboré et s'est poursuivi dans le cadre du stage de DEA puis de la thèse de J.-L. Metzger. Cette thèse est co-financée par l'INRA et l'INRIA et a débuté fin 2000. Le sujet s'intitule "Élaboration de formalismes de représentation et de raisonnement pour les systèmes d'informations géographiques". Il s'agit de concevoir un système de raisonnement à partir de cas pour aider à l'analyse et à l'exploitation de résultats d'enquêtes en exploitations agricoles. Pour cela, nous nous sommes intéressés à la représentation sous forme de graphes de structures spatiales et au calcul de similarités entre structures spatiales. Nous avons étudié différents systèmes (RCO, graphes conceptuels, logiques de descriptions) pour représenter ces structures. Une phase importante d'acquisition de connaissances et de formulation de problèmes avec les chercheurs de l'INRA a également eu lieu. En particulier nous avons travaillé ensemble sur la transformation en graphes conceptuels de schémas synthétiques (les chorèmes) représentant l'organisation spatiale d'exploitations agricoles. Une des séances a été filmée et fait l'objet d'une analyse par une équipe de recherche en psychologie cognitive (CODISANT, LPI–GRC, Université Nancy 2).

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

3.1.3 La gestion de connaissances

À l'heure actuelle, la conception de systèmes d'information intelligents (SII) 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'informations 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'informations par le contenu des documents, doit plutôt 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.

Dans le cadre du projet KVM— pour *Knowledge Valorization Matrix*, projet dans lequel interviennent les projets ECOO et MAIA du LORIA — l'objectif des recherches menées par les membres de l'avant-projet Orpailleur est de concevoir et de mettre au point un SII 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.). Plus précisément, il faut s'intéresser à des thèmes comme :

- L'étude d'un serveur de connaissances pour la gestion d'une mémoire d'entreprise, prenant en compte les connaissances propres à une entreprise et des connaissances d'ordre plus général comme des ontologies et des stratégies. Cette étude doit se faire dans l'environnement d'un système de RCO, qui autorise la classification des connaissances, la gestion de points de vue multiples et de référentiels, une personnalisation des connaissances destinée à favoriser et adapter les accès aux connaissances.
- Une approche symbolique pour l'aide à la stratégie et à la prise de décision, sur la base du raisonnement à partir de cas, avec la prise en compte de critères de décision qualitatifs. En particulier, des méta-connaissances, des historiques et des connaissances temporelles doivent être appréhendées et exploitées pour l'aide à la décision.
- L'étude et la mise en place de principes de conception d'un SII, qui doivent intégrer des fonctionnalités combinées de recherche d'information par le contenu et de fouille de données ; l'étude dans l'architecture d'un SII des rapports entre XML, le Web, et le serveur de connaissances, et ainsi distinguer les capacités qui doivent relever de XML et des outils de la galaxie XML, et celles qui doivent relever de la technologie des systèmes de RCO.
- 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.

3.1.4 La manipulation intelligente de documents : un pas vers le Web sémantique

Les travaux dont il est question ici se font dans le cadre de l'ARC INRIA Escrire, pour *Embedded Structured Content Representation In REpositories*. L'ARC INRIA Escrire est une action de recherche coopérative entre trois groupes de recherche de l'INRIA : le projet ACACIA (SOPHIA), l'avant-projet EXMO (Grenoble), et Orpailleur. L'objectif d'Escrire est de comparer *in fine* trois formalismes de représentations de connaissances — graphes conceptuels, représentations des connaissances par objets et logiques de descriptions — du point de vue de la gestion de documents scientifiques et techniques, de la représentation de leur contenu et de leur manipulation dans le cadre d'un SII.

La recherche et l'interrogation d'un site Web en s'appuyant sur le contenu des documents sont devenues des nécessités : les formalismes de représentation des connaissances sont de bons candidats pour représenter ce contenu. La représentation du contenu permet de le manipuler pour faire de la recherche par spécialisation, par similitude, par analogie, etc. Le langage XML permet de décrire les éléments du contenu des documents. Par ailleurs, l'intégration et le traitement d'informations provenant de sources variées et hétérogènes sont des questions préalables à la construction de systèmes intelligents de manipulation et de fouille de données, agissant en particulier sur les documents du Web. Le fait que les données et les documents soient hétérogènes et de structures non régulières nécessite de s'intéresser au traitement de e *données semi-structurées* (DSS) 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 l'avant-projet Orpailleur pour prendre en compte et manipuler des données semi-structurées textuelles consiste tout d'abord à les décrire en XML : utiliser XML comme passerelle entre les données brutes et les données réifiées et structurées. Les caractéristiques et les fonctionnalités de XML le rendent particulièrement bien adapté à la description de DSS. Plus précisément, la transformation des données décrites en XML dans un système de RCO peut être réalisée en s'appuyant sur le DOM (*Document Object Model*), une interface permettant l'accès au contenu et à la structure d'un document XML.

La mise en œuvre de raisonnements sur de telles données transformées — par exemple pour des besoins de résolution de problèmes ou de fouille de données — est ensuite dévolue à un système de RCO, où émerge la notion de classe *polythétique*, une classe qui peut se définir à la fois par des disjonctions et des conjonctions d'attributs, par opposition aux conventionnelles conjonctions d'attributs. Les données semi-structurées sont alors réifiées, c'est-à-dire transformées en « objets semi-structurés » et sont ensuite manipulées par des processus de classification adaptés. Un objet semi-structuré est vu comme un ensemble de propriétés valuées. Ces propriétés représentent les liens entre cet objet et d'autres objets. Un objet semi-structuré peut ne pas être attaché à une classe de référence dès sa création. Une procédure de classification recherche alors les classes dont cet objet est une instance. Une requête sur des DSS réifiées peut être représentée comme un concept qui doit être dans la hiérarchie : les instances de ce concept sont ensuite considérées comme les réponses à cette requête.

Dans le cadre de l'ARC INRIA Escrire, le système de représentation des connaissances choisi est RACER, une logique de descriptions mise au point à l'Université de Hambourg. Le système RACER a été choisi pour son expressivité et pour les possibilités d'intégration à Java. Le travail de recherche a consisté à trouver une méthodologie de traduction dans une base de connaissances d'une ontologie du domaine des données — ici des données biologiques — et des documents décrits en XML, et plus précisément des résumés d'articles de biologie sur des interactions entre gènes, ces documents étant annotés manuellement en utilisant les termes de l'ontologie du domaine. Les connaissances du domaine sont représentées dans l'ontologie par des concepts et des relations. Les concepts sont organisés dans une hiérarchie par une relation de subsomption. Les relations peuvent être binaires et n-aires (ces dernières ne sont pas prises en compte directement par le systèmes RACER). Les propriétés de ces relations comme la réflexivité et la transitivité ont été codées en dehors du système de représentation. La formation des requêtes est guidée par une interface graphique. Les requêtes sont écrites dans un langage de requêtes de type SQL, qui s'appuie sur des variables et des opérateurs (*and*, *or*, *not*, $\forall$, $\exists$, *type*, *equal*, etc). Certaines procédures ont été mises en place pour traduire les requêtes en terme de concepts d'une logique de descriptions, car la transformation des opérateurs en constructeurs de concepts ne se fait pas nécessairement naturellement. Le système RACER est ensuite utilisé pour représenter les connaissances relatives aux contenus des documents et au domaine de la biologie (génome), ainsi que pour raisonner sur les données afin de pouvoir retrouver les documents qui satisfont une certaine requête.

Bibliographie globale

Systèmes de RCO : [15].

RÀPC : [7] [6] [9] [12] [13].

Représentation et manipulation de données spatiales : [5] [3] [20].

Web sémantique : [4].

3.2 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.

Participants : Rim Al Hulou, Sandra Berasaluce, Martine Cadot, Fairouz Chakkour, Hacène Cherfi, Sébastien Hergalant, Florence Le Ber, Jean-François Mari, Sandy Maumus, Amedeo Napoli, Emmanuel Nauer, Rafik Taouil, 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 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 (lors de l'arrivée de nouvelles données par exemple). 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.

La classification par treillis, la recherche de motifs fréquents et l'extraction de règles informatives, et la classification par arbres de décision, font partie des techniques utilisées dans l'avant-projet Orpailleur pour la conception d'outils de fouille de données symbolique. Ces techniques relèvent de l'analyse de « tableaux de données binaires », et elles peuvent être étendues et adaptées à la prise en compte de données « complexes », où les attributs peuvent être

multivalués ou relationnels. Les structures émergeant de la classification sont des hiérarchies de concepts/classes et d'objets, qui relèvent du modèle de RCO. Une telle approche permet de prendre en compte des connaissances du domaine et de produire des structures interprétables. Les treillis de Galois servent aussi à faire émerger des règles d'association exactes ou partielles. En tenant compte des connaissances sur le domaine des données, il est alors possible d'élaguer l'ensemble des règles extraites obtenues — le processus d'extraction est de complexité exponentielle — et de ne conserver que des règles jugées intéressantes.

Un système d'ECBD s'appuyant sur le système de RCO Y3 — appelons-le ORPAILLEUR — a été réalisé dans l'avant-projet Orpailleur. Les développements logiciels ont consisté à étendre Y3 par des fonctionnalités de classification (de classes et d'instances) et de filtrage. Deux modules de classification par treillis et par arbres de décision ont été ainsi implantés, et ont été couplés à des modules de visualisation (hiérarchies d'objets et données). En particulier, le système ORPAILLEUR a été utilisé pour analyser des données médicales et des données textuelles (thèse d'Arnaud Simon, soutenue en septembre 2000).

3.2.2 ECBD et bases de données

L'avant-projet Orpailleur est associé au projet Dyade (INRIA Rocquencourt) et Bull pour une étude qui concerne les machines parallèles NEC et l'exploitation par ces machines de bases de données volumineuses. Actuellement, l'exploitation des bases de données est classique et s'articule essentiellement autour de fonctions d'interrogation standards. Une extension naturelle est d'étudier la mise en œuvre de techniques et d'outils permettant de pratiquer l'extraction de connaissances dans les bases de données, parallèlement aux fonctions d'interrogation. Pour cela, il est nécessaire de bien connaître la gestion des bases de données, mais aussi les méthodes symboliques d'ECBD, comme la classification par treillis et l'extraction de règles d'association dans un treillis.

Dans ce cadre, l'objectif du travail de recherches de Rafik Taouil, qui bénéficie d'une bourse post-doctorale INRIA depuis le 1er octobre 2000 (bourse prise en charge par Dyade), est d'étudier les diverses extensions possibles d'un langage classique d'interrogation de bases de données dans l'optique de la fouille de données. Un langage d'interrogation peut être étendu pour lui-même, pour autoriser des filtrages plus complexes et plus précis, mais aussi être étendu pour faire émerger des motifs fréquents et des règles d'association dans des tableaux de données volumineux.

En particulier, les motifs fréquents correspondent à des ensembles de propriétés dont le nombre d'occurrences dans les individus d'une population étudiée est supérieure à un seuil donné. À partir des motifs extraits, des règles d'association qui expriment des corrélations entre les propriétés des motifs fréquents peuvent également être extraites. La recherche de motifs fréquents et l'extraction de règles d'association entrent dans le cadre de la classification par treillis et la construction de treillis de Galois, qui, à partir d'un tableau booléen (présence – absence de caractéristiques), permettent de faire émerger des concepts formels, décrits par des ensembles de propriétés et des ensembles d'individus qui s'y rattachent (couples *intension* – *extension*), concepts qui s'organisent en un treillis. Ces techniques essentiellement symboliques peuvent être utilisées pour analyser une population, faire émerger des corrélations et des classifications selon certains points de vue dans une telle population.

3.2.3 La fouille de données avec des modèles de Markov

Une des originalités de l'avant-projet Orpailleur est de réutiliser certains travaux de classification numérique en reconnaissance de la parole pour procéder à de la fouille de signaux spatiaux-temporels, plus précisément 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 — les modèles de Markov cachés d'ordre 1 ou 2 (HMM1 et HMM2) — 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. Pour notre part, ce sont les 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 et qui nous intéressent.

D'un point de vue pratique, nous avons utilisé ces algorithmes d'apprentissage à base de HMM1 et HMM2 sur des données spatio-temporelles de l'utilisation du territoire, pour mettre à jour les successions culturales pratiquées dans différentes régions. Les HMM1 et HMM2 permettent de représenter des observations spatio-temporelles comme des successions d'états où les transitions entre états dépendants, suivant l'ordre du modèle, de l'état courant et des n états précédents.

Les résultats fournis par les algorithmes ont été évalués en relation étroite avec des experts agronomes de l'INRA et sur différents jeux de données, représentant des régions où les successions culturales sont différentes.

À un niveau plus théorique, nous nous intéressons à la classification non supervisée de territoires à l'aide de HMM sur des critères spatiaux, et à la segmentation de données spatio-temporelles.

3.2.4 ECBD, recherche de motifs fréquents et implication statistique

L'ECBD s'appuie sur des méthodes de *fouille de données*, parmi lesquelles figurent la recherche d'ensembles fréquents de propriétés — ou motifs fréquents — et l'extraction de règles à partir de ces motifs, les classification par arbres de décision et par treillis. Ces méthodes constituent le noyau de l'approche symbolique de la fouille de données. Une autre approche de l'ECBD, qui dérive de l'analyse de données et des statistiques, consiste à faire émerger des règles auxquelles est associée une pondération d'ordre statistique : « si a alors presque b ». Ces règles sont la base de l'*analyse statistique implicative* et sont implantées dans le logiciel CHIC (voir Régis Gras et collaborateurs, *L'implication statistique, une nouvelle méthode exploratoire des données*, La Pensée Sauvage Éditions, Grenoble, 1996). Par ailleurs, les *ensembles approximatifs* (voir Z. Pawlak, *Rough Sets*, Kluwer Academic Publishers, 1991) se définissent

par l'intermédiaire d'une approximation inférieure et d'une approximation supérieure, approximations qui peuvent être considérées comme la borne inférieure et la borne supérieure d'un ensemble d'éléments. Les ensembles approximatifs peuvent être exploités pour différents besoins, comme la classification approximative et la représentation de classes *polythétiques* (où les individus partagent un certain nombre de propriétés mais n'ont pas tous nécessairement les mêmes propriétés), la mise en valeur de dépendances approximatives entre individus et entre classes d'individus. Ces aspects classificatoires et l'émergence de dépendances entre individus peuvent être considérés comme des méthodes de fouilles de données et exploitées à cette fin.

Un nouveau thème de recherche dans lequel s'est engagé l'avant-projet Orpailleur (pour un travail de thèse) consiste à étudier en détail et de façon conjointe les trois formalismes que sont (i) l'analyse statistique implicative, (ii) la recherche de motifs fréquents et l'extraction de règles d'association, (iii) les ensembles approximatifs. Il s'agit d'expliciter les relations qui existent entre les trois formalismes et de faire émerger, à partir d'une telle étude, les principes d'une fouille de données symbolico-numérique. Ce travail de recherche revêt des aspects théoriques et pratiques, et s'insert naturellement dans les préoccupations actuelles d'Orpailleur. En particulier, les résultats de ce travail de recherche devraient avoir des retombées sur la représentation de connaissances imprécises, la classification approximative, la fouille de données en général, mais aussi sur la représentation et la manipulation de données semi-structurées.

3.3 Fouille de textes

Mots clés : Fouille de textes, information scientifique et technique, informatique linguistique, terminologie, interprétation, synthèse de textes, classification, treillis de Galois, extraction de motifs fréquents, extraction de règles d'association, raisonnement à partir de cas.

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

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. Plusieurs types d'activités peuvent être concernées par la fouille de textes, chacune se différenciant de par sa finalité :

1. l'acquisition de connaissances à partir de textes qui vise à produire des ressources telles que des thésaurus ou des ontologies pouvant être utilisées par d'autres outils informatiques ;
2. la recherche d'information, la veille technologique et l'analyse de l'information pour lesquelles l'utilisateur est expert d'un domaine et c'est lui qui interprétera en final – grâce à ses connaissances du domaine – les résultats de la fouille de textes.
3. la construction de synthèse (au sens résumé « sémantique ») d'un texte ou d'un ensemble de textes, où il s'agit de situer un texte par rapport aux autres et de pouvoir expliquer en quoi il s'en rapproche et en quoi il s'en différencie.

La quasi-totalité des travaux actuels en fouille de textes reposent sur des méthodes numériques, statistiques ou neuronales de classification. Nous avons adopté une méthode symbolique

qui nous permet d'associer au processus de fouille des connaissances du domaine et d'envisager un traitement beaucoup plus fin de l'information dans les textes.

Nos travaux ont porté essentiellement sur la veille technologique ou l'analyse de l'information. Ils se structurent suivant deux dimensions complémentaires. La première est l'approche informatique linguistique, dont le but est d'extraire du texte des structures conceptuelles. Bien que cette problématique ait fait l'objet de nombreuses publications en linguistique informatique, des questions spécifiques à la fouille se posent. La seconde dimension est une approche s'appuyant sur les connaissances et sur le raisonnement pour réaliser la tâche de fouille de données proprement dite sur les structures caractérisant les textes.

Indexation conceptuelle des textes par le raisonnement à partir de cas

Les objectifs que nous nous sommes fixés sont d'indexer des documents textuels par des structures conceptuelles qui nous permettent de dépasser la simple indexation au niveau du terme (ou mot-clé) qui est utilisée jusqu'à présent. Ces structures conceptuelles sont des objets que nous organisons au sein d'une base de connaissances et sur lesquels il devient alors possible de faire des calculs de spécialisation ou de généralisation. La complexité des phrases dans les textes que l'on traite constitue bien sûr le premier obstacle à cette phase d'extraction de l'information. Seule une méthode permettant de prendre en compte une analyse partielle de la syntaxe et de la sémantique et exploitant des connaissances du domaine permet d'envisager une solution à cette question.

Nos travaux partent d'une analyse en constituants des phrases et nous cherchons à identifier le rôle syntaxico-sémantique des différents constituants. Notre hypothèse de départ est que, dans un domaine de spécialité – pour lequel il est possible de se constituer un modèle de connaissances – le fait de disposer de la représentation complète et correcte d'un ensemble de phrases peut nous permettre de déduire l'analyse de nouvelles phrases. Ainsi, le raisonnement à partir de cas est parfaitement adapté à cette hypothèse (et à sa validation) ; son utilisation constitue une approche très novatrice du traitement de la langue dans un contexte scientifique et technique [7]. Toutefois, ce type de traitement est complexe, et nous avons réduit le cadre expérimental de nos travaux (au moins dans un premier temps) à la prise en compte des relations méronymiques dans des textes sur le stress professionnel.

Fouille de textes et extraction de règles d'association

Nous abordons la fouille de textes selon deux dimensions très fortement reliées : la première, présentée dans cette section, repose sur l'extraction et l'exploitation de règles d'association. La seconde, présentée dans la section suivante, repose plus particulièrement sur la classification.

Dans le cadre de la thèse de H. Cherfi, nous abordons la fouille de textes en nous centrant sur l'extraction de règles d'association. Notre objectif est de montrer en quoi la méthodologie d'extraction de règles d'association peut fournir une approche intéressante pour la veille technologique. L'utilisateur final est un expert du domaine dont l'objectif est :

- de retrouver au travers des règles d'association des connaissances déjà établies dans le domaine ;
- de découvrir des informations qui n'ont pas encore le statut de connaissances dans la mesure où ce sont des signaux dit « faibles ».

Nous travaillons dans ce cadre sur des résumés d'articles scientifiques issus de notices bibliographiques 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 (Base bibliographique et thésaurus PASCAL de l'INIST). Une première indexation basée sur 3300 termes différents produit 1200 règles, alors qu'un nettoyage de l'indexation ne gardant que 630 termes en produit 128. Ces règles ont été confiées à un expert pour qu'il les interprète par rapport à ses connaissances du domaine. Même si la première réaction est de dire que toutes les règles présentées sont cohérentes du point de vue du domaine, certaines sont plus « interprétables » que d'autres, par exemple quand tous les termes nécessaires à l'expert pour commenter la règle sont bien présents dans cette règle. Nous travaillons également à l'explicitation d'une telle notion d'interprétabilité.

Nous avons choisi de travailler sur l'évaluation des règles d'association par rapport aux besoins de l'expert notamment sur les critères de sélection de celles qui sont plus pertinentes que les autres par rapport au domaine ou par rapport aux besoins de l'expert. Nous souhaitons aussi exploiter davantage les propriétés logiques des règles pour que l'expert puisse faire une lecture plus fine des règles et qu'il puisse y trouver des connaissances qui ne soient pas déjà largement attestées. Nous expérimentons également un certain nombre d'indices statistiques déjà existants pour ordonner les règles avant de les présenter à l'expert [8].

Hybridation de méthodes de classification

Les approches de la classification pour la fouille de données sont nombreuses et reposent sur des méthodes plutôt numériques comme les réseaux de neurones, les statistiques ou, comme nous les développons dans l'avant-projet Orpailleur, sur des méthodes symboliques comme la classification par treillis et la recherche de motifs fréquents.

Le foisonnement de ces approches nous ont bien sûr amenés à nous interroger sur l'intérêt de l'usage d'une méthode plutôt qu'une autre. Notre réflexion nous a conduit à formuler les points suivants :

1. l'utilisation conjointe de plusieurs méthodes de classifications peut s'avérer, aux yeux des utilisateurs et des experts, plus riche que l'utilisation d'une seule méthode, à condition de pouvoir identifier l'enrichissement effectif ;
2. les méthodes de classification se situant dans des paradigmes très différents, une comparaison purement formelle ou mathématique de ces méthodes ne nous semble pas constituer une réponse au point ci-dessus. Nous adoptons donc une démarche expérimentale, qui est évaluée par rapport à une tâche particulière de fouille de données ;
3. les systèmes interactifs, comme l'outil MicroNomad (développé dans le projet Cortex au LORIA), permettant de naviguer entre différents points de vue d'un même ensemble de données permettent à l'utilisateur de faire des déductions intéressantes sur les données. Il est opportun de confronter cette démarche interactive à des méthodes formelles et symboliques comme l'extraction de règles d'association.
4. dans un premier temps, nous nous associons à Jean-Charles Lamirel du projet Cortex pour étudier la complémentarité des méthodes de classification comme celle des cartes auto-adaptatives de Kohonen avec nos méthodes de classification par treillis.

Il est possible résumer les intérêts et inconvénients des cartes auto-adaptatives de Kohonen ou des treillis de Galois en quelques points :

- les cartes de Kohonen :
  - le schéma de pondération des classes est complexe et donc difficile à interpréter par un utilisateur ;
  - la topographie des cartes de Kohonen confère au système un bon pouvoir illustratif et de bonnes capacités à synthétiser l'information d'une grande base de données ;
  - le schéma de pondération permet de prendre en compte la non-linéarité et donc de valoriser les signaux dits « faibles ».
- les treillis de Galois :
  - les treillis possèdent généralement un très grand nombre de classes, ce qui en fait des structures difficiles à interpréter ;
  - la construction d'un treillis peut être incrémentale, ce qui peut s'avérer extrêmement utile lors du traitement de volumes importants de données ;
  - des règles d'association peuvent être extraites et associées à des classes du treillis.

Nous avons abordé la complémentarité des méthodes de classification par la définition d'une méthode de projection des classes de Kohonen sur les classes d'un treillis construits à partir d'un même ensemble de données. Différentes heuristiques ont été testées et nous avons défini une méthode d'évaluation de la qualité de cette projection. Les premiers résultats ont montré [18, 11, 10] :

- 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. À ce titre, les classes symboliques peuvent être perçues comme une explication ;
- qu'il est possible d'utiliser la structure hiérarchique du treillis pour associer à la carte de Kohonen une structure hiérarchique ;
- qu'avec certaines heuristiques de projection, la structure hiérarchique calculée préserve la connexité des classes de Kohonen.

Cette première approche de l'hybridation de méthodes de classification a suscité un grand intérêt, certainement en réponse à la multiplication des algorithmes qui sont proposés dans les conférences sur la fouille de données. Nous souhaitons en premier lieu conforter nos premiers résultats par des expérimentations sur d'autres types de données et étudier comment les règles d'association peuvent être introduites dans cette démarche.

Bibliographie globale sur l'ECBD

ECBD symbolique et recherche de motifs fréquents : [16, 17].

ECBD numérique : [1].

Fouille de textes : [8, 7, 18, 11, 10].

3.4 Systèmes intelligents de traitement de l'information et Web sémantique

Mots clés : accès intelligent à l'information, système hypertexte, extraction de connaissances à partir de données, recherche d'information intelligente, Web sémantique,

données semi-structurées, grandes bases de données, bases de données à objets.

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

Accès intelligent à l'information

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.

Le but de nos travaux est de développer des environnements dans lesquels exploiter des DSS (références bibliographiques et Web en particulier). Nous souhaitons notamment fournir aux chercheurs, ou 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. Nous proposons dans ce cadre une approche générale qui couple l'exploitation de connaissances (extraites par des techniques de fouille de données) à un système de recherche d'information (RI) [2] ; cette approche a permis de mettre en place un système opérationnel pour fouiller des données bibliographiques et accéder au Web. Ce système, dénommé IntoBIB, repose sur l'utilisation du principe de l'hypertexte pour accéder de façon exploratoire aux données, dans le but identifier celles qui répondent à un certain besoin. Ces données peuvent alors être exploitées par des fonctionnalités de fouille (dénombrements, classifications, extractions de règles, etc.), déclenchées à la demande dans le but d'extraire de nouvelles connaissances capables de guider la recherche d'information. . L'idée est que la fouille de données et la recherche d'information sont deux approches extrêmement complémentaires pour appréhender des données : 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.

Un enjeu de ce système est également la production de connaissances qui donnent une idée synthétique d'un ensemble de données et qui puissent être exploitées dans un raisonnement. Concrètement, cette production de connaissances peut se voir comme un processus d'ECBD qui agit sur des informations scientifiques et techniques. Elle peut consister à chercher 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'information scientifique et technique. Là aussi, une normalisation minimale des données à exploiter est indispensable pour éviter des biais statistiques.

La mise en place d'un ces environnements nécessite d'exploiter les travaux issus de différents domaines de recherche : fouille de données, recherche d'information, gestion de bases de données. D'un point de vue technique, il est nécessaire de faire collaborer les fonctionnalités offertes dans chacun de ces domaines. Pour cela, nous avons adopté une approche modulaire qui repose sur l'utilisation de XML pour la représentation et l'échange des données manipulées, ainsi que sur un traitement par flux de données.

Accès à l'information sur le Web – Web sémantique

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. Le but de ce travail de recherche est d'étudier l'exploitation de connaissances dans le cadre d'un accès intelligent aux données du Web.

Le système IntoBIB décrit précédemment fournit les moyens d'exploiter les données structurées d'un domaine (références bibliographiques en particulier), et de faire émerger des connaissances sur ce domaine, comme des réseaux d'auteurs, le vocabulaire employé par tel ou tel auteur, etc. Ces connaissances sont alors exploitées pour la recherche d'information sur le Web. L'accès aux données du Web est réalisé via les moteurs de recherche classiques (AltaVista, Google, etc.) qui sont utilisés comme des outils distants. IntoBib fournit par conséquent un cadre général pour guider l'accès aux données du Web, ceci en combinant l'accès par navigation hypertextuelle (par exemple dans des thèmes de plus en plus spécialisés) à l'interrogation des moteurs de recherche, utilisés comme passerelles entre IntoBIB et le Web. L'utilisation de techniques d'ECBD permet dans ce contexte de déterminer automatiquement des requêtes ou encore d'assister l'utilisateur dans l'expression de son besoin ; cette aide à la formulation par l'apport de connaissances extraites de données permet d'améliorer considérablement l'efficacité de la recherche d'information sur le Web [2].

Cependant, une aide à la formulation à partir de connaissances extraites de données du domaine (pour interroger les moteurs de recherche) s'avère insuffisante dans un processus complet — notamment itératif — de recherche d'information. C'est pourquoi, un agent de recherche d'information sur le Web a été également développé. Cet agent exploite des connaissances extraites des données par des techniques de fouille pour guider la recherche d'information. Il parcoure le Web en suivant les liens hypertextes et exploite des connaissances (contexte, règles associatives, etc.) pour évaluer les documents rencontrés selon des critères propres ; il se soustrait ainsi aux méthodes de classement quelque peu mystérieuses des moteurs de recherche.

Les nombreuses possibilités d'exploiter des connaissances du domaine dans le cadre d'une recherche d'information intelligente nous ont amenés à développer un agent complètement paramétrable. Il s'agit désormais d'identifier plus précisément les paramètres (selon les types de connaissances exploités en particulier) qui permettent de mener une recherche d'information efficace sur le Web.

Expérimentations et réalisations : vers la veille technologique

Les travaux menés jusqu'à présent concernent l'intégration des données (locales ou distantes, multi-sources et de natures différentes) et l'extraction de connaissances à partir de ces données. Un système d'investigation sur le Web a été mis en place, qui permet un accès indifférencié aux données locales ou distantes, ainsi que des croisements entre ces données. 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'information 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 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 le Web : 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), mais également permettre d'analyser (valider, rejeter, juger, etc.) et de classer les documents, rencontrés lors d'un parcours du Web par notre agent. 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.

Bibliographie générale : [4, 2, 14]

3.5 Bioinformatique et fouille de données en biologie

Mots clés : génome, classification, recherche de motifs fréquents, extraction de règles d'association, rôle de l'analyste en fouille de données, modèles de Markov.

Participants : Bertrand Aigle [Laboratoire de Génétique et Microbiologie, UHP Nancy 1], Bernard Decaris [Laboratoire de Génétique et Microbiologie, UHP Nancy 1], Sébastien Hergalant, Tawfik Labib, Florence Le Ber, Pierre Leblond [Laboratoire de Génétique et Microbiologie, UHP Nancy 1], Jean-François Mari, Sandy Maumus, Amedeo

Napoli, Rafik Taouil, Sophie Visvikis [Unité INSERM U525, Nancy].

Certains 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 Etat-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 :

- La recherche avec le co-encadrement d'un étudiant en DEA d'ingénierie génétique ; ce DEA s'est poursuivi à la rentrée 2001 par une thèse toujours en co-tutelle financée par l'INRA et la Région Lorraine.
- 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 bio-informatique `GCC`.

Le travail de DEA [19] a porté sur le développement et l'utilisation de modèles de Markov cachés (HMM) adaptés à l'étude de grandes séquences d'ADN génomique bactérien. Une panoplie d'outils permettant la segmentation de telles séquences par des HMM d'ordre deux a été conçue dans cette optique. L'analyse des résultats obtenus a été réalisée par visualisation graphique de ces segmentations, puis par recherche systématique de corrélations entre l'allure du graphe et la structure des données dans le génome.

A la base, de tels modèles ont été constitués de façon à n'émettre aucune hypothèse préalable sur la nature du matériel génétique à analyser. Le processus d'apprentissage des séquences via ces HMM reste donc vierge de tout a priori sur les données ; ce qui le rend capable d'accepter et de traiter n'importe quel arrangement séquentiel.

Sur le chromosome de la bactérie *Streptomyces coelicolor* – modèle d'étude au laboratoire de Génétique et Microbiologie – ce travail a permis de démontrer la capacité de ces HMM à détecter plusieurs grands types de motifs réitérés, très différents entre eux tant au niveau physique (organisation topologique sur la séquence) que fonctionnel. L'étude détaillée de ces motifs montre que certains d'entre eux pourraient être la cause ou la conséquence des phénomènes d'instabilité génétique observés chez cette bactérie. Ces phénomènes se traduisent par des réorganisations génomiques, des amplification, des pertes ou des transferts de matériel génétique.

La compréhension et l'interprétation des mécanismes impliqués débutent ici par l'analyse au niveau des séquences et présentent un réel intérêt en biologie. Il devient par conséquent nécessaire de généraliser la recherche des motifs d'ADN répétés en automatisant les méthodes de détection présentées ci-avant. D'autres génomes bactériens, comme *Streptomyces ambofaciens*, sont en cours de séquençage et fourniront un ensemble de données originales pour la poursuite de cette étude dans le cadre d'une thèse.

L'ensemble de ces travaux est effectué sur des génomes bactériens à forte valeur industrielle (industrie du médicament, industrie lactique) puisque des organismes tels que *Streptomyces ambofaciens* ou *Streptomyces coelicolor* sont, entre autre, capables de produire plus des trois quarts des antibiotiques connus à ce jour.

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, qui est associée à l'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, Sandy Maumus a étudié un autre aspect important, qui est le rôle que peut jouer l'analyste dans le processus de fouille de données : cette première étude, qui montre des résultats plutôt encourageants, a été menée dans le cadre d'un stage de DESS, et elle doit être continuée dans le cadre d'une thèse en bioinformatique, qui débutera en 2002, avec les données de la cohorte Stanislas cette fois.

Bibliographie : [19].

Sandy Maumus, *Extraction de connaissances : une application à la mycologie — le rôle de l'expert dans le processus de fouille de données*, Rapport de DESS Double-Compétence, Université Henri Poincaré (Nancy 1), Novembre 2001.

3.6 ECBD et exploitation de bases de données en chimie organique

Mots clés : systèmes de représentation de connaissances par objets, extraction de connaissances dans des bases de données chimiques, classification, recherche d'information, interrogation et navigation dans des bases de données, perception, synthèse en chimie organique, recherche de motifs fréquents, extraction de règles d'association.

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

Le travail de thèse de Sandra Berasaluce est co-dirigé par Claude Laurenço (CCIPE et LIRMM de Montpellier) et Amedeo Napoli. Il porte sur l'extraction de connaissances et l'aide à l'interrogation et à la navigation dans des bases de données de chimie organique, qui sont des questions très actuelles en synthèse organique.

Pour résoudre les problèmes de synthèse qui leur sont posés, les chimistes organiciens utilisent de très nombreuses connaissances, informations et données. Outre leurs propres acquis et savoirs, ils sont désormais aidés dans leur tâche par des bases de données qui, bien que devenues indispensables, restent encore très imparfaites. Les bases de données chimiques commerciales existantes essaient de rassembler une grande partie des données issues des travaux de recherche

(environ 18 millions de substances et plus de 10 millions de réactions décrites). Mais ces bases de données telles qu'elles sont conçues sont un recueil de données non structurées (collection d'exemples particuliers de réactions pour les bases de données de réactions) et, de ce fait, ne peuvent répondre qu'à des besoins documentaires ou encyclopédiques.

Un des problèmes majeurs des bases de données est qu'elles n'utilisent pas ou très peu les connaissances du domaine. Or la connaissance est nécessaire au raisonnement. Cette connaissance peut provenir d'expert du domaine mais peut être aussi extraite des bases de données existantes par un processus d'ECBD. C'est ce que nous nous proposons de faire dans ce projet de recherche. Pour cela, nous nous appuyons sur des travaux antérieurs effectués au sein du GDR 1093 du CNRS sur le logiciel Resyn-Assistant. Ce logiciel est dédié à la représentation et à la manipulation des molécules, et à leur perception selon divers points de vue : topologie, fonctionnalité et stéréochimie.

Tout comme il est nécessaire de représenter les molécules à différents niveaux d'abstraction, nous avons défini une façon originale de représenter les réactions. Une réaction est un objet complexe et dynamique qui rend compte du passage d'un ensemble de molécules (les réactants) en un autre ensemble de molécules (les produits) suivant certaines conditions physico-chimiques (les réactions conditionnelles). Nous nous servons des résultats de la perception pour décrire la réaction selon chacun des points de vue précédemment cités.

Vu la diversité des besoins en informations des chimistes, nous avons orienté nos recherches selon deux axes correspondant à deux points de vue majeurs en chimie organique : la fonctionnalité et la topologie. La présence de fonctions dans les molécules est à l'origine de la plupart de leurs propriétés réactives, c'est à dire du fait que des molécules mises en présence dans certaines conditions vont réagir pour donner de nouvelles molécules. Nous avons décidé d'appliquer des techniques de fouille de données sur les descriptions des réactions que nous sommes capables de faire en nous focalisant sur le point de vue de la fonctionnalité. Nous obtenons des motifs que nous pouvons interpréter en terme de chimie : "pour faire tel groupe fonctionnel, on peut combiner telles fonctions entre elles tandis que telles fonctions restent inchangées". Les motifs fréquents dégagés permettent de donner les possibilités envisageables par rapport à un problème donné tandis que les règles d'association déduites permettent de classer les propositions énoncées. Ce sont des informations importantes sur la réactivité des fonctions qui répondent à une demande des chimistes.

Une première expérimentation sur une base de petite taille (base ORGSYN version 2000, environ 6000 réactions) a été effectuée. Elle a conduit à retrouver des résultats déjà observés dans des articles dont les auteurs avaient étudié le contenu des bases de données de réactions. De nombreux motifs ou règles peuvent être validés par l'état des connaissances en chimie. Ainsi, certaines fonctions connues pour leur réactivité se retrouvent dans nombre de motifs avec le type fonction créée, d'autres connues pour leur stabilité sont la majeure partie du temps inchangées. Forts de ces premiers résultats, nous sommes en train de procéder à une deuxième expérimentation sur une base de données de réactions plus grande (base JSM version 1999, plus de 60000 réactions).

La deuxième étude que nous menons concerne le point de vue topologique, et propose une modélisation des réactions de construction de cycles, avec un modèle conceptuel de ce type de réactions. Nous envisageons plusieurs utilisations de ce travail : une indexation des réactions de construction de cycles, et la recherche de toutes les réactions connues pour construire un

cycle donné.

4 Logiciels

4.1 Les modèles de Markov pour l'ECBD numérique

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

Dans le cadre de la fouille de données numériques, nous avons développé des méthodes d'apprentissage et d'inférence fondées sur une modélisation à l'aide des chaînes de Markov d'ordre 2 pour analyser des données temporelles et spatiales d'occupation du sol.

En collaboration avec des experts agronomes de l'INRA (station SAD de Mirecourt), nous avons écrit un logiciel permettant la construction de modèles de Markov d'ordre quelconque, et leurs apprentissages sur les données *Ter Uti* fournies par les Directions Régionales de l'Agriculture et de la Forêt (DRAF). 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.

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.

4.2 Un système de RCO pour l'ECBD symbolique

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

Le système d'ECBD ORPAILLEUR a été développé à partir d'une application médicale, menée conjointement avec le Registre Lorrain du Cancer de l'Enfant (RLCE, Hôpital d'Enfants de Nancy-Brabois). Ce système comprend principalement deux modules de fouille et deux modules de visualisation :

- Le module de classification par arbres de décision 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, pour traiter des données représentées dans un système de RCO.
- le module de classification par treillis repose sur un algorithme incrémental de construction de treillis de Galois, avec exploitation de connaissances du domaine. Des règles expliquant les données peuvent être extraites du treillis.
  Un module complémentaire qui s'appuie sur la théorie des « ensembles approximatifs » (*rough sets*) permet de prendre en compte le degré de confiance associé aux règles extraites.
- 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 un point de vue géographique sur les données. Ce module, conçu pour traiter le caractère essentiellement géographique de certaines données du RLCE, est utilisé pour mettre en

évidence la répartition géographique de facteurs d'étude liés aux données. La répartition obtenue est ensuite comparée à des cartes de répartition de référence pour faire apparaître d'éventuelles corrélations.

4.3 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 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 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.
- Lemmatiseur du français : le lemmatiseur du français produit le lemme d'une forme fléchie (développé en collaboration avec Fiammetta Namer, Université de Nancy 2).

4.4 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).

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

4.5 Le système KASIMIR

Participants : Benoît Bresson [correspondant], 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 PALETUVIER permet d'afficher la hiérarchie des classes manipulées par PAULETTE ou une sous-hiérarchie, avec des possibilités (encore réduite) d'édition. 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 a été développée. 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 est en cours de finition. Une extension plus générale à la classification élastique, selon une approche RÀPC, est à l'étude.

4.6 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 « IntoWeb », 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 pour manipuler des documents par leur contenu, dans le cadre d'un

système de connaissances. l'intégration de bases de données et la résolution de problèmes dans le domaine des données. 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.

4.7 Les systèmes RÉSYN et RÉSYN-ASSISTANT

Participants : Sandra Berasaluce [correspondante], Claude Laurenço [CCIPE 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 Actions régionales, nationales et internationales

5.1 Actions locales

5.1.1 La collaboration URI et Orpailleur

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

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.

5.1.2 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.

5.2 Actions nationales

5.2.1 L'ARC INRIA Escrire

Participants : Rim Al Hulou, Hacène Cherfi, Olivier Corby [ACACIA SOPHIA ANTIPOLIS], Rose Dieng [ACACIA, INRIA SOPHIA ANTIPOLIS], Jérôme Euzenat [EXMO, INRIA RHÔNE-ALPES], Amedeo Napoli, Emmanuel Nauer, Yannick Toussaint.

Cet ARC INRIA se fait en collaboration avec le projet ACACIA (Rose Dieng, INRIA SOPHIA-ANTIPOLIS) et l'avant-projet EXMO (Jérôme Euzenat, 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 des connaissances souvent difficilement accessibles 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 devient dès lors 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. Il existe différents formalismes de représentation des connaissances, mais aucune étude importante et poussée mettant en jeu la représentation et la manipulation de documents textuels n'a

encore eu lieu, pour permettre la comparaison de leurs qualités respectives selon un tel point de vue.

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), des avant-projets EXMO (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, en l'occurrence des documents textuels sur le génome, 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 est définie conjointement et introduite dans un format XML. Un ensemble de requêtes définies de manière coordonnée doit être évaluée dans chacun des contextes.

À l'issue de ce travail, les différents formalismes doivent être comparés entre eux — mais aussi à la recherche plein-texte — selon un protocole prédéfini. Celui-ci doit s'appuyer sur des critères tant qualitatifs — expressivité des requêtes, accessibilité/lisibilité des informations, ... — que quantitatifs : temps de réponse à une requête, taux de précision/rappel des réponses, ... Cette évaluation va proposer une grille d'analyse des avantages et des inconvénients d'un langage de représentation formel vis-à-vis de la recherche d'informations sur le Web, et va tenter de déterminer les contextes favorables à l'exploitation de chacune de ces représentations.

5.2.2 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 2001 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 de 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.

5.2.3 Le projet KASIMIR

Participants : Benoît Bresson, Jean Lieber, Amedeo Napoli.

Le projet Kasimir (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 comme une activité de conception portant sur le savoir mis en œuvre, donc ici les référentiels (c'est-à-dire, les protocoles de traitement). Cela suppose que le savoir préexiste et qu'il est conservé dans une mémoire. Un des objectifs premiers du projet Kasimir 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 peuvent être considérées comme achevées, et une troisième est en cours de réalisation. 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 référentiel 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 référentiel, 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 référentiel.

La deuxième phase de ce travail est applicative : pour pouvoir raisonner avec le référentiel et le faire évoluer, il faut le connaître et le représenter informatiquement. C'est l'application KASIMIR/RÉFÉRENTIEL qui permet de le faire. Actuellement, KASIMIR/RÉFÉRENTIEL est sur le point d'entrer dans une phase opérationnelle et devrait pouvoir être utilisé par des médecins de la région Lorraine. Une évaluation de KASIMIR/RÉFÉRENTIEL par un ensemble de ces médecins est actuellement menée pour mettre en évidence l'impact de cette opérationnalisation.

La troisième phase est l'étude de l'adaptation du référentiel de traitement du cancer du sein aux situations dans lesquelles l'utilisation « littérale » de cette base de connaissances n'est pas satisfaisante. L'objectif est l'implantation d'un système de raisonnement à partir de cas — KASIMIR/HORS RÉFÉRENTIEL — permettant de réaliser ces adaptations. Un travail d'acquisition et de modélisation des connaissances d'adaptation nécessaires à un tel système a été réalisé et étudié dans [12]. Il s'appuie sur des discussions avec les experts sur des comptes-rendus du « comité de concertation pluri-disciplinaire » (CCP) 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 CCP. 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 $3,9\,\text{cm}$ 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.

Deux présentations de KASIMIR/RÉFÉRENTIEL (tel qu'il existe) et de KASIMIR/HORS RÉFÉRENTIEL (tel qu'il est envisagé) sont parues en 2000. Une étude de l'acquisition et de la modélisation des connaissances d'adaptation est présentée dans [12].

5.2.4 Le projet KVM

Participants : Benoît Bresson, François Charpillet [MAIA, LORIA], Claude Godard [ECOO, LORIA], Amedeo Napoli, Emmanuel Nauer.

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.

5.2.5 Une collaboration avec le Musée de La Villette

Participants : Jean-Charles Lamirel [CORTEX, LORIA], Yannick Toussaint.

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 de collaboration avec le Musée de La Villette repose sur deux constats. D'une part les responsables d'expositions aimeraient avoir accès à l'information des collections. D'autre part les collections sont en réalité des « mines » 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, et de représenter les connaissances du conser-

vateur, 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 — gestion des données et des connaissances — en favorisant le filtrage et la classification par points de vue des objets, et donner ainsi une meilleure appréhension de la base et de son contenu. De plus, l'utilisation d'un système d'ECBD peut permettre de découvrir et d'expliquer des éventuelles corrélations entre les objets, et de mieux comprendre l'évolution des objets au cours du temps, et faire ainsi émerger de nouveaux thèmes d'exposition.

5.2.6 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. Des modèles ont été proposés parallèlement dans l'équipe Orpailleur et au LISI à Lyon, par les chercheurs Béatrice Fuchs et Alain Mille. Une collaboration s'est engagée avec ces derniers, pour confronter ces modèles de l'adaptation et les enrichir par les expériences respectives de chacun des collaborateurs. 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 dans [9]. Il s'appuie sur les notions d'appariement entre problèmes et de dépendance entre un problème et la solution qui lui est associée. L'étude de telles généralisations dans une double optique théorique et applicative constitue une perspective de cette collaboration.

5.2.7 Le projet Supersélect

Participants : Bernard Nivelet [Bull Louveciennes], Amedeo Napoli, Rafik Taouil.

L'avant-projet Orpailleur est associé au projet Dyade (INRIA Rocquencourt) et Bull pour une étude qui concerne les machines parallèles NEC et l'exploitation par ces machines de bases de données volumineuses. Supersélect est un système de gestion de bases de données doté d'un moteur de requêtes très efficace qui s'appuie sur une représentation des données adaptée à l'utilisation de super-calculateurs vectoriels (NEC). Dans le cadre de la collaboration Dyade-Bull-Orpailleur, nous nous intéressons à l'extension des possibilités du moteur Supersélect vers des fonctionnalités de fouille de données, plus particulièrement vers l'extraction de motifs fréquents et de règles d'associations.

Nous approchons ces problèmes en nous appuyant sur le treillis des motifs fermés fréquents : formellement, étant donnée une base de données au format Supersélect, nous étudions des algorithmes efficaces pour la génération des motifs fermés fréquents et des règles d'associations, ainsi que des méthodes pour le stockage et l'interrogation des treillis. Ces algorithmes, une fois effectivement implantés dans l'environnement du moteur Supersélect, pourront bénéficier de

la puissance de calcul des super-calculateurs vectoriels. Outre l'aspect algorithmique, l'intérêt et l'originalité de notre approche réside dans l'utilisation des treillis pour la tâche de fouille de données. Ainsi, notre démarche montre une des facettes des nombreuses formes d'exploitations possibles des treillis (règles d'associations, classifications, regroupement, visualisation) dans des contextes et applications diverses (marketing, communication, médecine...).

5.2.8 Le GDR CNRS 1093 TICCO

Participants : Sandra Berasaluce, Jean Lieber, Amedeo Napoli.

Le GDR CNRS 1093 TICCO — *Traitement informatique de la connaissance en chimie organique* — en est à sa dernière année. Il réunit des chercheurs en chimie organique du CCIPE 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 est 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 nécessite 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 (CCIPE 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.

6 Diffusion de résultats

6.1 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.

6.2 Enseignement

- Enseignements et organisation scientifique de cours (en France et à l'étranger).
- Ouverture en septembre 2001 du DESS TEXTE pour « Traitements informatiques pour l'EXploitation de l'information dans les TExtes », sous la co-direction de Fiammetta Namer (Université Nancy 2) et de Yannick Toussaint.

- 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.

7 Bibliographie

Livres et monographies

[1] J.-F. MARI, R. SCHOTT, *Probabilistic and Statistical Methods in Computer Science*, Kluwer Academic Publishers, janvier 2001.

Thèses et habilitations à diriger des recherches

[2] E. NAUER, *Principes de conception de systèmes hypertextes pour la fouille de données bibliographiques multibases*, Thèse en Informatique, Université Henri Poincaré Nancy 1, Janvier 2001.

Articles et chapitres de livre

[3] F. LE BER, L. MANGELINCK, A. NAPOLI, « Étude de treillis de relations topologiques pour l'interprétation d'images satellitaires », *Géomatique 11*, 2001, p. 215–249.

Communications à des congrès, colloques, etc.

[4] R. AL HULOU, E. NAUER, A. NAPOLI, « XML et les systèmes de représentation de connaissances par objets pour la gestion de données semi-structurées », *in : Extraction et gestion des connaissances (EGC-2001), Nantes*, H. Briand, F. Guillet (éditeurs), *Extraction des connaissances et apprentissage Vol. 1(nos 1/2)*, Hermès, Paris, p. 363, 2001. Poster.

[5] M. CAPITAINE, S. LARDON, F. LE BER, J.-L. METZGER, « Chorèmes et graphes pour modéliser les interactions entre organisation spatiale et fonctionnement des exploitations agricoles », *in : Géomatique et espace rural. Journées CASSINI 2001, Montpellier, France*, T. Libourel (éditeur), p. 145–163, septembre 2001.

[6] F. CHAKKOUR, A. NAPOLI, Y. TOUSSAINT, « Extraire des structures prédicatives à partir des textes, vers une indexation conceptuelle des textes », *in : Actes du* IX^{ème} *séminaire français de raisonnement à partir de cas*, B. Fuchs, A. Mille (éditeurs), 2001.

[7] F. CHAKKOUR, Y. TOUSSAINT, « Sentence Analysis by Case-Based Reasoning », *in : The Fourteenth International Conférence on Industrial and Engineering Applications of Artificial Intelligence and Expert SystemsIEA/AIE, Budapest, Hungary*, M. Ali (éditeur), *Lecture notes in artificial intelligence, 2070*, International society of applied intelligence, Springer Verlag, p. 546–551, juin 2001.

[8] H. CHERFI, Y. TOUSSAINT, « Extraction et Interprétation des Règles d'association pour la Fouille de Textes », *in : Actes de l'Atelier A3CTE-01 : Applications, Apprentissage, Acquisition des connaissances à partir de textes électroniques*, Plate-forme AFIA, p. 15–16, Grenoble, Juin 2001. Résumé (Version courte).

[9] B. FUCHS, J. LIEBER, A. MILLE, A. NAPOLI, « Un algorithme pour la phase d'adaptation du raisonnement à partir de cas », *in : Actes des journées nationales sur les modèles de raisonnement (JNMR'01), Arras*, A. Herzig (éditeur), p. 79–92, 2001.

[10] J.-C. Lamirel, Y. Toussaint, J. Ducloy, C. Czysz, C. Francois, « Reseaux neuronaux avances pour la cartographie de la science et de la technologie : Application a l'analyse des brevets », *in : VSST 2001*, octobre 2001.

[11] J.-C. Lamirel, Y. Toussaint, C. Francois, X. Polanco, « Using a MultiSOM approach for Mapping of Science and Technology », *in : ISSI 2001*, août 2001.

[12] J. Lieber, P. Bey, F. Boisson, B. Bresson, P. Falzon, A. Lesur, A. Napoli, M. Rios, C. Sauvagnac, « Acquisition et modélisation de connaissances d'adaptation, une étude pour le traitement du cancer du sein », *in : Actes des journées ingénierie des connaissances (IC-2001)*, J. Charlet (éditeur), Preses universitaires de Grenoble, p. 409–426, 2001.

[13] J. Lieber, « Des règles, des cas, des généralités, des spécificités, des applications, des adaptations, des chaînes, des combinaisons et des tartes », *in : Actes du* IX^{ème} *séminaire français de raisonnement à partir de cas*, B. Fuchs, A. Mille (éditeurs), 2001.

[14] E. Nauer, « Réhabilitons les doublons ! », *in : Veille Stratégique Scientifique et Technique, 1 (Full Paper)*, FPC/UPC - SFBA - IRIT, Barcelone, Espagne, Octobre 2001.

[15] P. Rapicault, A. Napoli, « Évolution d'une hiérarchie de classes par interclassement », *in : Langages et Modèles à Objets (LMO-01), Le Croisic, L'Objet 7(1/2)*, R. Godin, I. Borne (éditeurs), Hermès, Paris, p. 215–230, 2001.

[16] G. Stumme, R. Taouil, Y. Bastide, N. Pasquier, L. Lakhal, « Intelligent Structuring and Reducing of Association Rules with Formal Concept Analysis », *in : Proceedings of the Joint German/Austrian Conference on Artificial Intelligence, LNAI 2174*, 2001.

[17] R. Taouil, Y. Bastide, « Computing Proper Implications », *in : Proceedings of the 9th International Conference on Conceptual Structures (ICCS), Workshop on Concept Lattices-based KDD*, p. 49–61, 2001.

[18] Y. Toussaint, J.-C. Lamirel, M. D'Aquin, « Combining Symbolic and Numeric Techniques for Database Content Analysis », *in : AI/IEA01*, 2001.

Rapports de recherche et publications internes

[19] S. Hergalant, « Segmentation du génome de Streptomyces coelicolor par chaînes de Markov pour la recherche de réitérations », *Stage de dea*, septembre 2001.

[20] F. Le Ber, L. Mangelinck, A. Napoli, « Design and comparison of lattices of topological relations for spatial representation and reasoning », *rapport de recherche*, novembre 2001, `http://www.inria.fr/rrrt/rr-4321.html`.