INRIA

INSTITUT NATIONAL DE RECHERCHE EN INFORMATIQUE ET EN AUTOMATIQUE

Projet verso

Bases de Données

Futurs

THÈME 3A

Table des matières

1. Composition de l'équipe

Verso est une équipe commune avec le LRI (CNRS et Université Paris Sud), bientôt localisée à Orsay. Verso se termine en 2002. Le projet Gemo prend sa suite en 2003.

Responsables scientifiques
Serge Abiteboul [DR]
Marie-Christine Rousset [Professeur, Univ. Paris 11]

Assistante de projet
Geneviève Grisvard

Personnel INRIA
Ioana Manolescu [CR. sept-dec]
Luc Segoufin [CR]

Personnel des universités
Hélène Gagliardi [Maître de Conférence, Univ. Paris 11]
Nathalie Pernelle [Maître de Conférence, Univ. Paris 11]
Chantal Reynaud [Professeur, Univ. Paris X]
Brigitte Safar [Maître de Conférence, Univ. Paris 11]
Véronique Ventos [Maître de Conférence,Univ. Paris 11]

Conseillers scientifiques
Bernd Amann [Maître de Conférence, CNAM]
Christine Froidevaux [Professeur, Univ. Paris 11]
Michel Scholl [Professeur, CNAM]

Chercheurs invités
Tova Milo [Professeur, U. Tel Aviv]
Victor Vianu [Professeur, U.C. San Diego, 6 mois]

Chercheurs post-doctorants
Angela Bonifati [Politecnico Milano]
Francois Goasdoué [ATER, Paris 11]

Ingénieur expert
Jérôme Baumgarten [sep-dec]

Chercheurs doctorants
Omar Benjelloun [Boursier MENRT, Paris 11]
Grégory Cobéna [X-Télécom]
Irini Fundulaki [ATER, CNAM]
Gloria-Lucia Giraldo [Paris 11]
Amar-Djalil Mezaour [Boursier MENRT, Paris 11]
Benjamin Nguyen [Boursier MENRT, Paris 11]
Antonella Poggi [Roma University]
Alexandre Termier [Boursier MENRT, Paris 11]

3. Fondements scientifiques

Mots clés : *Bases de données*, *bases de connaissances*, *représentation de connaissances*, *intégration de données*, *intégration sémantique*, *langages de requêtes et optimisation*, *requêtes distribuées*, *données semistructurées*, *XML*, *Web*, *Services Web*, *contrôle des changements*, *logique et informatique*, *complexité*.

L'information disponible sur le Web est de plus en plus complexe, distribuées, hétérogéne et changeante. Les sevices Web, tels que SOAP doivent également être perçus comme délivrant de l'information qui doit être exploitée.

L'objectif de Gemo est double : D'une part, il s'agit d'étudier les problèmes fondamentaux soulevés par les systèmes de gestion de l'information et des connaissances. D'autre part, nous voulons proposer et évaluer des solutions nouvelles à ces problèmes.

La thématique principale de Gemo est l'intégration de données hétérogènes distribuées. L'information est perçue comme un concept générique qui couvre les données, les connaissances, et les services. Plus précisément, les problèmes adressés dans Gemo sont les suivants : la découverte de sources de données ou de services pertinents, ainsi que la compréhension de leur contenu ou objectifs, leur intégration, et enfin leur monitorage au cours du temps. Notre but est d'offrir un environnement à la fois puissant et flexible qui simplifie la mise en œuvre d'applications donnant un accès rapide aux données sur le Web.

La création d'entrepôts de données et de médiateurs, capables d'intégrer des sources de données multiples et hétérogènes, constitue un bon moyen d'atteindre ces objectifs.

La résolution de tous les problèmes que nous venons d'évoquer implique l'utilisation conjointe de techniques d'Intelligence Artificielle (comme la classification) et de techniques de Bases de Données (comme l'indexation)

Gemo est né de la fusion du Projet Verso de l'INRIA Rocquencourt, avec une partie du groupe IASI du Laboratoire de Recherches en Informatique (UMR 8623 CNRS-Université Paris-Sud). Pour l'instant, le projet est divisé géographiquement entre Rocquencourt et Orsay.

Nous proposons Gemo comme projet de recherche pour l'INRIA-Futurs, localisé à Saclay.

3.1.1. Principaux axes de recherche

Nous travaillons sur différents aspects de la gestion de données et de connaissances aussi bien d'un point de vue théorique que pratique. Nous décrivons d'abord les outils théoriques que nous utilisons, puis nous présentons les principaux thèmes de recherche auxquels se rattachent les projets applicatifs dans lesquels nous sommes impliqués. Les points de vue pratique et théorique sont étroitement liés dans tous nos travaux de recherche.

4. Domaines d'application

4.1. Introduction

Mots clés : *web*, *télécommunications*, *commerce électronique*, *ingénierie*, *portail d'entreprise*, *moteur de recherche*, *entrepôts de données*, *multimédia*.

Les bases de données n'ont pas de champ d'application privilégiés. En effet, toute application mettant en jeu une quantité importante de données ou d'informations se doit d'utiliser des bases de données. Les technologies développées récemment dans le projet ont notamment de nombreuses applications dans le cadre des nouvelles applications du web (télécom et multimédia), des portails d'entreprises, des systèmes d'information pour la fabrication, etc. Verso a choisi de cibler principalement des applications dans le cadre des nouveaux services du web.

Nous mentionnerons en guise d'illustration une application, l'exploitation intelligente des données du web à l'aide d'entrepôts de données.

4.2. Un entrepôt sur le risque alimentaire

Mots clés : *internet*, *web*, *portail d'entreprise*, *moteur de recherche*, *entrepôts de données*, *risque alimentaire*.

De tels systèmes sont indispensables aux entreprises (par exemple) pour trouver l'information dont elles ont besoin sur le web et l'intégrer dans leurs systèmes d'information.

Notre but est développer des outils permettant de construire des entrepôts de données dans des domaines spécifiques en intégrant de manière automatique des informations découvertes sur le Web, avec des données privées et des données obtenues de fournisseurs de contenu. Ce travail devrait se situer dans le cadre du projet RNTL e.dot qui devrait démarrer début 2003. Il s'appuie sur le nouveau format du Web, XML, et de nouveaux

services comme ceux proposés par Xyleme, fondés sur des requêtes de haut niveau et sur le monitorage du Web. Les expérimentations auront pour cadre la création d'un entrepôt sur le risque alimentaire.

Ce travail s'appuie sur une coopération avec l'équipe BIA de l'INRA et la start-up Xyleme, issue du projet, qui propose des services sur le Web autour de XML. Le projet e.dot a comme application phare l'analyse du risque alimentaire, BIA ayant été choisi comme centre de compétences informatiques par le Ministère de l'Agriculture et le Ministère de la Recherche dans le cadre d'un programme national de recherche sur ce sujet.

5. Logiciels

- XyDiff : un outil de *diff* pour XML (logiciel libre)
- Thesu : un prototype d'interrogation de collections de documents utilisant la sémantique des liens, en collaboration avec l'Université d'Athènes.
- SPIN : un outil pour la surveillance des sites web.
- Active XML : un langage et un système basé sur des documents XML contenant des appels de services imbriqués [13]
- STYX : définition et mise en œuvre d'une plate-forme générique, permettant l'intégration et l'interrogation de ressources XML pertinentes pour une communauté web.
- OntoClass et OntoQuery : deux outils (brevetés par France Telecom R&D) pour la classification automatique de concepts, et pour la ré-écriture des requêtes conjonctives vers des plans de requêtes, développé dans le cadre du projet PICSEL.
- TreeFinder : un prototype pour découvrir des fragments d'arbres fréquents dans une collection de données XML.
- Zoom : un prototype pour construire et raffiner un treillis de classes sur des données semi-structurées, développé dans le cadre du projet GAEL.
- OntoMedia : un prototype pour la construction automatique de composantes d'ontologies à partir de DTDs, développé dans le cadre du projet PICSEL2.

6. Résultats nouveaux

6.1. Fondements théoriques

Participants : Serge Abiteboul, Tova Milo, Victor Vianu, Luc Segoufin.

Mots clés : *semi-structuré*, *langage de requêtes*, *automate*.

A cause du besoin de représenter et d'interroger des données et des connaissances, la logique joue un rôle central dans les travaux de recherche de Gemo. La logique du premier ordre (qui fonde les langages de requêtes relationnels) et les logiques de description (qui permettent de représenter et d'interroger des connaissances structurées complexes) sont des outils formels puissants pour appréhender le pouvoir d'expression et la complexité de l'évaluation de requêtes dans le contexte des bases de données relationnelles ou orientées-objets.

L'importance grandissante de formats d'échanges de données de type XML combinant du texte à une structure d'arbre nous conduit de plus en plus à utiliser des langages à base d'automates simples ou d'automates d'arbres. De ce fait, nous nous interessons aux logiques monadiques du second ordre qui caractérisent logiquement les automates.

Avec le Web, les bases de données gèrent maintenant un volume très important d'informations, et ne cessent d'échanger des données avec d'autres sources. La théorie de la complexité est utilisée pour obtenir une analyse fine des ressources nécessaires pour l'évaluation de requêtes. Par exemple, dans [30], nous initions une étude formelle du traitement d'un flot de données XML en utilisant des ressources mémoire limitées. Dans ce cadre, nous avons encore établi une forte connexion avec la théorie des automates.

6.2. Médiation entre données XML

Participants : Gloria Giraldo, Nathalie Pernelle, Chantal Reynaud, Marie-Christine Rousset, Michele Sebag, Alexandre Ternier, Veronique Ventos.

Mots clés : *Intégration sémantique*, *ontologies*, *regroupement automatique*.

Dans le cadre de l'intégration sémantique de données XML, il est important de pouvoir construire aussi automatiquement que possible des ontologies (ou schémas médiateurs) pouvant servir d'interface de requêtes entre des utilisateurs et une collection hétérogène de documents XML. Des correspondances (ou mappings) doivent pouvoir être établies entre l'ontologie servant de schéma médiateur et les différentess DTDs des documents XML relevant de cette ontologie commune.

Les travaux effectués dans cet axe de recherche s'inscrivent dans le prolongement des travaux réalisés dans le cadre de Xyleme et du projet PICSEL.

Pour la construction automatique de schémas médiateurs au dessus d'une collection hétérogène de documents XML, nous travaillons sur des algorithmes de regroupement automatique de données semi-structurées. Le but est de regrouper dans des classes des documents XML présentant des similarités, et de calculer pour ces classes les descriptions les plus appropriées (les généralisations les plus précises de l'ensemble des instances de chaque classe). Nous avons implémenté dans le système $TreeFinder$ une méthode [32] de découverte de structures d'arbres dont la copie exacte ou perturbée apparait fréquemment dans une collection d'arbres étiquetés modélisant la structure de documents. Nous avons aussi dévelopé le système ZooM [9], basé formellement sur les treillis de Galois imbriqués, qui regroupe des données semistructurées selon deux niveaux d'abstraction différents. Une session du système ZooM commence par fournir une hierarchie grossière de classes. Ensuite, l'utilisateur sélectionne une classe-mère et une classe-fille dans la hiérarchie et ZooM construit un affinement de la hierarchie des classes comprises entre ces deux classes.

Dans le cadre de PICSEL2, nous étudions [26] comment construire semi-automatiquement une ontologie à base de classes à partir d'un ensemble de DTDs ou de XML schemas relatifs à un même domaine d'application (e.g., le tourisme). Une maquette, nommée $OntoMedia$, a été développée pour extraire les composants de l'ontologie (des noms de classes, de relations et de propriétés) à partir d'un ensemble de DTDs.

Le travail sur l'intégration de données génomiques qui avaient été initié dans l'équipe par Christine Froidevaux sera poursuivi dans l'équipe Bioinformatique qui vient juste d'être créée au LRI, et avec laquelle de forts liens scientifiques vont se mettre en place.

6.3. Mediation pour le Web sémantique

Participants : Serge Abiteboul, Bernd Amann, Christine Froidevaux, Irini Fundulaki, Chantal Reynaud, Marie-Christine Rousset, Brigitte Safar, Michel Scholl.

Mots clés : *Web sémantique*, *ontologies*, *intégration de données*.

L'objectif du web sémantique est de tendre vers un web dont la sémantique des données serait à la fois compréhensible par des utilisateurs humains et appréhendable par des entités informatiques (agents, moteurs de recherche, serveurs d'informations). Le marquage sémantique des données du web ouvre de nombreuses perspectives d'amélioration de la qualité des moteurs de recherche. Cependant, le passage à l'échelle du web est un véritable défi qui impose que les problèmes clés soient clairement identifiés et étudiés de manière approfondie en évitant les solutions ad-hoc ne passant pas à l'échelle.

Les trois problèmes clés qui nous intéressent dans Gemo sont : la médiation entre ontologies, la médiation entre sources de données, et la médiation entre le web et ses utilisateurs.

Dans le prolongement de nos travaux sur PICSEL et Xyleme, [2][3][7], qui correspondent à une vision centralisée de la médiation entre sources de données (basée sur un schéma médiateur unique), nous étudions [11] une approche "Peer-to-Peer" correspondant à une approche décentralisée et collaborative de la médiation entre sources.

Nous travaillons aussi sur l'utilisation d'ontologies pour permettre une interrogation plus interactive et coopérative entre le Web et ses utilisateurs. Dans [21], nous montrons comment une ontologie peut être utilisée

pour évaluer la *similarité* entre requêtes, et ainsi permettre de choisir la requête jugée la plus proche de la requête initiale, parmi un ensemble de requêtes prédéfinies construites à partir des sources disponibles.

Nous travaillons maintenant sur le problème dual de la spécialisation de requêtes obtenant trop de réponses.

Le prototype StyX [25] met en oeuvre un langage de mappings simple mais expressif permettant de décrire des ressources XML comme des vues locales sur un schema conceptuel global. Ces vues sont ensuite utilisées par un algorithme efficace de reformulation de requêtes en requêtes XPath ou en XQuery [19] qui sont ensuite évaluées sur les sources.

Enfin, nous projetons de renforcer notre travail [18][29] sur la modélisation et la représentation d'ontologies servant d'outils de médiation entre des données, des services et des utilisateurs.

6.4. Utilisation des Services Web

Participants : Serge Abiteboul, Jerome Baumgarten, Omar Benjelloun, Angela Bonifati, Gregory Cobena, Ioana Manolescu, Tova Milo, Benjamin Nguyen, Marie-Christine Rousset.

Mots clés : *Intégration de données*, *services Web*, *peer-to-peer*.

Nous étudions Active XML (AXML, en bref) [35], un cadre declaratif pour l'intégration de données, reposant sur les Services Web et qui est mis en oeuvre dans une architecture Peer-to-Peer. Il est basé sur la notion de *documents AXML*, qui sont des documents pouvant contenir des appels de services. Ce langage permet d'une part de spécifier de tels documents et d'autre part de définir de nouveaux services de manière déclarative, basés sur ces documents.

La notion de documents contenant des appels de fonctions n'est certes pas nouvelle. Cependant, AXML est le premier à faire des appels de Services Web inclus dans des documents un outil puissant d'integration de données à l'echelle du web. En particulier, le langage offre des fonctionnalités permettant de contrôler le déclenchement des appels de services, la durée de vie des données, ou encore le choix entre données extensionnelles et intensionnelles dans les échanges. Plusieurs scénarii sont capturés, tels que la médiation, l'entreposage de données et une forme restreinte de calcul distribué. En combinant des approches telles que l'entrepôsage et la médiation, nous avons été amenés à nous intéresser à l'intégration des Services Web eux-meêmes. Un objectif à long terme pourrait être la découverte de Services Web utiles à une certaine application et leur utilisation, de manière automatique ou semi-automatique.

Parmi les travaux en cours sur Active XML, on peut citer : (i) le développement d'un prototype, qui a été présenté en démonstration à la conférence VLDB'2002 [13], (ii) des travaux théoriques sur les fondements d'un modèle restreint, (iii) des expérimentations consistant à construire des entrepôts avec AXML (le système SPIN, mentionné ci-dessous) ; (iv) l'utilisation d'AXML dans un environnement Peer-to-Peer avec terminaux mobiles, dans le cadre du projet DBGLOBE (également présenté ci-dessous).

6.5. Entrepôts thématiques de données du Web

Participants : Serge Abiteboul, Omar Benjelloun, Jerome Baumgarten, Gregory Cobena, Tova Milo, Benjamin Nguyen, Marie-Christine Rousset.

Mots clés : *Entrepôts*, *données thématiques*, *spécification déclarative*.

Nous souhaitons développer une approche flexible et générique permettant de spécifier de façon déclarative les données utiles pour enrichir ou créer un entrepôt thématique, en simplifier l'acquisition à partir du Web ainsi que la surveillance[14], et organiser ces données en vue de faciliter leur interrogation ultérieure.

Nous avons commencé une première expérimentation basée sur Active XML. Pour cela, nous développons en Active XML une bibliothèque de services Web utiles pour la construction d'entrepôts de données thématiques.

Un autre travail a débuté sur ce thème en collaboration avec l'Université d'Athènes autour du projet Thesu. Dans Thesu, nous suivons une approche relationnelle pour spécifier le choix et l'enrichissement des ressources Web relatives à une certaine thématique. La caractéristique de l'approche de Thesu est de renforcer l'information donnée par les liens entre documents par une information sémantique extraite des

mots apparaissant dans le voisinage des liens. Cela nous a conduit à définir de nouvelles mesures de similarités entre documents, et de nouvelles techniques de regroupements de documents, qui permettent de réaliser une structuration de l'entrepôt. Un premier prototype a été implémenté. Pour la prise en compte du temps et la surveillance de l'entrepôt, nous envisageons de réutiliser les travaux réalisés sur la gestion des changements dans des documents XML [23] et les différences entre versions de documents XML [22].

Finalement, une thèse a débuté sur la recherche ciblée sur le Web (focused crawling). L'approche étudiée combine l'appels à des services Web existants (e.g., Google) avec des techniques d'apprentissage automatique dans le but d'obtenir une stratégie du type "le meilleur d'abord" pour explorer le Web, en fonction de la thématique visée.

Le projet e.dot labélisé en 2002 par le RNTL est directement rattaché à ce thème, ainsi que le travail concernant l'archivage du Web francais.

7. Contrats industriels

7.1.1. Xyleme

Xyleme SA a été créée en septembre 2000 à partir des travaux initiés dans le projet Verso, qui ont aussi impliqué des membres de l'équipe IASI (M.-C. Rousset). Xyleme s'est détaché de Verso en mars 2001 et est maintenant une entreprise d'environ 30 personnes. Sa mission est de *fournir une nouvelle génération de technologies de gestion du contenu de données, capable d'exploiter le potentiel de XML.*

La collaboration avec Xyleme continue, S. Abiteboul et G. Cobena sont des conseillers scientifiques pour Xyleme. Xyleme est l'un des partenaires du projet RNTL e.dot, et est aussi impliqué dans les travaux sur l'archivage du Web français.

7.1.2. Les projets PICSEL et PICSEL2

Ces deux projets sont le résultat d'une collaboration avec France Telecom R&D qui a commencé en décembre 1997.

Le but de PICSEL était la conception d'un environnement déclaratif pour construire des médiateurs basés sur des ontologies. Deux outils Java ont été développés et brevetés par France Telecom R&D : $OntoClass$, un outil pour la classification automatique de concepts définis utilisant une logique de description, et $OntoQuery$, un outil pour ré-écrire des requêtes conjonctives exprimées en termes de l'ontologie au niveau du médiateur, dans des plans de requêtes utilisant uniquement des vues décrivant le contenu des sources.

PICSEL2 est une continuation de PICSEL. Son but est de faire passer à l'échelle du Web l'approche médiateur mise en œuvre dans PICSEL. Le but est de faciliter la construction automatique d'un schéma de médiation sur plusieurs sources XML décrites par des DTDs et reliées à un domaine unique.

Un prototype (OntoMedia) a été développé, qui extrait des composantes d'ontologies automatiquement depuis un ensemble de DTDs. Dans PICSEL2, nous développons aussi des méthodes initiées sur PICSEL pour répondre de manière coopérative aux requêtes.

7.1.3. Le projet RNTL e.dot

Le but de e.dot est de développer un entrepôt XML de l'information concernant les risques alimentaires. Le projet a été labélisé RNTL en 2002 et devrait commencer bientôt. Il est composé de Gemo ainsi que du groupe BIA de l'Institut National de Recherche en Agronomie qui est spécialisé dans cette application, et de la compagnie Xyleme.

Dans ce projet, les aspects techniques comportent la surveillance des sites web, la classification et l'intégration d'informations hétérogènes, et le stockage de grands volumes de données dans le repository Xyleme. L'une des difficulté est d'accéder et d'intégrer des données pré-existentes dans ce domaine. Le projet constitue donc un banc d'essai excellent pour la technologie Gemo.

7.1.4. Archivage du web français

Le Web est de plus en plus vu comme une source importante d'information. Des organisations telles que Internet Archive et des agences gouvernementales essayent d'en archiver des portions. En France, la Bibliothèque Nationale de France relève ce défi pour le Web français. Nous travaillons avec eux sur ce sujet.

D'un point de vue technique, nous étudions des questions telles que la sélection des pages à traverser pour optimiser la consommation des ressources, et pour archiver plus souvent des sites importants. Ceci conduit aux recherches sur la délimitation du périmètres du Web français et des sites français, calculant l'importance des pages / des sites utilisant des critères multiples incluant la topologie du Web [5][15].

8. Actions régionales, nationales et internationales

8.1. Actions nationales

Des liens forts existent avec l'équipe de recherche en bases de données du LRI (N. Bidoit, P. Rigaux, E Waller), l'équipe en bio-informatique du LRI (C. Froidevaux, C. Rouveirol), l'équipe d'apprentissage automatique du LRI (M. Sebag), l'équipe Cedric du CNAM-Paris, le projet Caravel de l'INRIA-Rocquencourt (F. Llirbat), le groupe BIA de l'Institut National de Recherche en Agronomie (O. Haemmerlé, P. Buche, C. Dervin), LISI de l'Université de Lyon 1 (M. Hacid), LIRMM de l'Université de Montpellier (M. Chein, M-L. Mugnier) .

8.2. Actions financées par la commission européenne

Des liens forts existent avec l'Université de Mannheim (G. Moerkotte), l'Université de Marburg (T. Schwentick), l'Université d'Athènes (M. Vazirgianis) et l'ETH Zurich (H. Schek, R. Weber), l'Université de Madrid (A. Gomez-Perez), l'Université de Manchester (I. Horrocks), l'Université de Rome (M. Lenzerini), et Politecnico di Milano (S. Ceri).

8.2.1. Projet européen DBGlobe

DBGlobe, une approche basée sur les données pour le calcul distribué à l'échelle mondiale, est un projet IST composé de l'Université de Ioannina, l'Institut d'Informatique, l'Université d'Economie (Grèce), l'Université de Chypre (Chypre), l'Université de Californie à Riverside (USA), l'Université d'Aalborg (Danemarque) et l'INRIA.

Le but du projet DBGlobe est de développer des nouvelles techniques de gestion des données pour répondre aux problèmes soulevés par les calculs distribués à l'échelle mondiale. Au premier abord, le calcul distribué à l'échelle mondiale est un problème de bases de données : comment concevoir, construire, et analyser des systèmes capables de gérer des grandes quantités de données. Toutefois, l'approche traditionnelle en bases de données, qui consiste à stocker les données utiles dans des systèmes monolithiques de gestion de bases de données, devient inutile dans un tel environnement.

Dans la recherche actuelle en bases de données, les données sont relativement homogènes, présentent un petit degré de distribution (réduit à un petit nombre de sites), et sont passives (elles restent inchangées tant qu'elles ne sont pas explicitement mises a jour), et statiques (leur localisation ne change pas). Ces hypothèses ne sont plus valables dans un contexte de calcul distribué à l'échelle du Web. Ce contexte crée le besoin de nouvelles fondations théoriques dans tous les aspects de la gestion de données : la modélisation, le stockage, et l'interrogation.

DBGlobe adopte une approche centrée sur les données pour la conception et l'analyse des environnements dynamiques d'entités mobiles et autonomes en considérant : (i) des entités mobiles comme des unités de stockage de données primaires, et (ii) des entités mobiles comme des mini-servers (entités de calcul) qui protègent et encapsulent l'accès à leurs données.

8.2.2. Le réseau européen OntoWeb

OntoWeb est un réseau regroupant différentes équipes académiques et industrielles européennes autour du but de supporter l'échange d'information et les transactions d'affaires par des ontologies. Le concept central est

que les ontologies sont appelées à jouer un rôle majeur dans le web sémantique, fournissant une sémantique de données qui peut être interprétée automatiquement.

8.3. Relations bilatérales internationales

8.3.1. Coopération avec les pays du Moyen-Orient

Des liens forts existent avec l'Université Hebrew (C. Beeri) et l'Université de Tel-Aviv (T. Milo).

8.3.2. Coopération avec les pays de l'Amérique du Nord

Un projet Franco-Canadien en coopération avec l'Université de Toronto (A. Mendelzon et L. Libkin) a commencé en 2001. Un projet Franco-Américain en coopération avec l'Université de Californie à San Diego (V. Vianu) a commencé en 2002.

Des liens forts existent avec l'Université de Stanford (J. Widom), l'Université de Pennsylvanie (S. Davidson), l'Université de Californie à San Diego (V. Vianu), AT&T (S. Amer-Yahia), Lucent-Bell Labs (J. Siméon), l'Université de Washington (A. Halevy), l'Université Rutgers (A. Borgida).

8.4. Accueil de chercheurs étrangers

Cette année, nous avons accueilli :

- Tova Milo, professeur à l'Université de Tel-Aviv (1 an)
- Victor Vianu, professeur, UC San Diego (2 mois)
- Catriel Beeri, professeur à l'Université de Jérusalem (1 mois)
- Michael Benedikt, chercheur aux Bell-labs de Lucent (1 mois)

9. Diffusion des résultats

9.1. Participation à des colloques

L'équipe a eu de nombreuses publications dans des conférences internationales et des colloques (voir la bibliographie). Enfin, certains membres du projet ont participé à des comités de programmes. La liste en est donnée ci-dessous.

S. Abiteboul

- PC Chair de la conférence international VLDB (Very Large DataBases), 2003.
- ACM-SIGMOD International Conference on the Management of Data (SIGMOD), Madison, Wisconsin (2002)
- International Conference on Database Theory (ICDT), Siena, Italy (2003)
- International Workshop on Web Dynamics (in WWW conference), Hawai (2002)
- International Workshop on E-Services and the Semantic Web Workshop, Toronto (2002)
- KRDB02, Internat. workshop on Knowledge Representation meets Databases, 2002

B. Amann

- International Conference on Extending Database Technology (EDBT), 2002
- International Conference and Workshop on Database and Expert Systems Applications (DEXA 2002)

I. Manolescu

- Web Technologies and Applications , special track at the 18th ACM Symposium on Applied Computing (SAC 2003), Melbourne, Florida, USA.

- Workshop on Web-Based Collabration (WBC'02), in conjunction with the DEXA 2002 conference.
- Workshop on Efficience and Effectiveness of XML Tools and Techniques (EEXTT), in conjunction with the VLDB 2002 conference, Hong Kong, China.
- Workshop Efficient Web-based Information Systems (EWIS), in conjunction with the OOIS 2002 conference, Montpellier, France.

T. Milo

- PC Chair de ACM International Symposium on Principles of Database Systems, 2003
- Workshop on Next Generation Information Technologies and Systems (NGITS), 2002
- International Conference on Very Large Databases, 2002

C. Reynaud

- 13th International Conference on Knowledge Engineering and Knowledge Management (EKAW-2002), Siguenza, Spain, 2002.
- EKAW-2002 Workshop on Knowledge Management through Corporate Semantic Web, Siguenza, Spain, 2002.
- ECAI Workshop on Machine Learning and Natural Language Processing for Ontology Engineering, Lyon, France, 2002
- ECAI Workshop on Knowledge Transformation for the Semantic Web, Lyon, France, 2002
- 13èmes Journées francophones d'Ingénierie des Connaissances (IC'2002), Rouen, France, 2002

M-C. Rousset

- PC Chair de 9th International Workshop on Knowledge Representation meets Databases (KRDB 2002)
- ACM International Symposium on Principles of Database Systems, 2003
- International Joint Conference on Artificial Intelligence (IJCAI), 2003
- European Conference on Artificial Intelligence (ECAI), 2002
- International Symposium on Methodologies for Intelligent Systems (ISMIS), 2002
- International Workshop on the Web and Databases (WebDB), 2002
- Journées Francophones d'Extraction et de Gestion des Connaissances (EGC), 2002
- Congrès Francophone de Reconnaissances des Formes et Intelligence Artificielle (RFIA), 2002

M. Scholl

- Conférences sur les Bases de Données Avancées 2002
- Workshop on distributed data and structures, 2002
- International Conference on Advanced Information Systems and Engineering, 2002
- ACM conference on Geographic Information Systems, Washington, 2002
- International Conference on Ontologies, Databases and Applications of Semantics, 2002

L. Segoufin

- Computer Science Logic (CSL), 2002.

9.1.1. Conférences invitées, tutoriels, cours, etc.

S. Abiteboul a été invité à présenter un exposé aux conférences internationales DEXA 2002 et WISE 2002 et, en France, à la Conférence sur les Bases de Données Avancées, à XML-France 2002 et aux journées RNTL 2002. Il a également animé un tutoriel sur les Services Web à l'école d'été Extending Database Technology (EDBT) International Summer School 2002 en Cargese (Corse).

I. Manolescu a donné un tutoriel sur « Adaptive and Self-Tuning Query Processing » à 2002 EDBT International Summer School à Cargese.

M-C. Rousset a présenté un exposé invité au symposium international sur les Methodologies pour systèmes intelligents (ISMIS 2002).

9.1.2. Animations scientifiques

Serge Abiteboul est membre (jusqu'en 2002) du comité exécutif de ACM International Symposium on Principles of Database Systems, et membre du comité exécutif de ACM SIGMOD on the Management of Data. Il est aussi président du comité de projets de l'unité INRIA Futurs.

Chantal Reynaud est co-chair de l'action spécifique « Web Sémantique » du département STIC du CNRS et co-organisatrice des Journées de l'action spécifique Web Sémantique du département STIC du CNRS, Paris (France), octobre 2002.

M-C. Rousset est membre du Comité de pilotage du RTP "Information et intelligence : raisonner et décider " du département STIC du CNRS, expert pour la Mission Scientifique Universitaire (DS1) et chargée de mission pour les Mathématiques et l'Informatique auprès du Délégué Régional à la Recherche et Technologie d'Ile de France pour le Contrat Plan Etat Région de l'Ile de France. Enfin elle est membre du comité de pilotage du PCRI (Pole Commun de Recherche en Informatique du plateau de Saclay.

M. Scholl est membre de la commission d'évaluation du RNTL et expert de la mission scientifique universitaire du MENRT.

9.1.2.1. Livres

Publication de *Comprendre XSLT* par B. Amann et P. Rigaux chez O'Reilly en 2002.

Publication d'une édition française de « Foundations of Databases » par S. Abiteboul, R. Hull et V. Vianu, et d'une édition portugaise de « Data on the web », par S. Abiteboul, P. Buneman et D. Suciu.

9.1.2.2. Édition

S. Abiteboul

- Information and Computation
- Journal of Digital Libraries
- Theory and Practice of Object Systems

B. Amann

- Revue Information - Interaction - Intelligence (I3)

C. Reynaud

- JEDAI (Journal Electronique d'IA de l'AFIA)
- Revue Information - Interaction - Intelligence (I3)

M-C. Rousset

- ACM Transactions on Internet Technology (TOIT)
- AI Communications (AICOM)
- Electronic Transactions on Artificial Intelligence (ETAI) (pour les domaines : Concept-based Knowledge Representation and Semantic Web).
- Revue Information - Interaction - Intelligence (I3)

M. Scholl

- Geoinformatica
- Annals of Telecommunications

10. Bibliographie

Livres et monographies

[1] B. AMANN, P. RIGAUX. *Comprendre XSLT.* O'Reilly, 2002.

[2] M.-C. ROUSSET, C. REYNAUD. *Picsel and Xyleme : two illustrative information integration agents..* Agentlink book, 2002.

Articles et chapitres de livre

[3] S. ABITEBOUL, S. CLUET, G. FERRAN, M.-C. ROUSSET. *The Xyleme Project.* in « Computer Networks 39 », 2002.

[4] S. ABITEBOUL, S. CLUET, T. MILO. *Correspondence and Translation for Heterogeneous Data.* in « Theoretical Comptre Science », numéro 1-2, volume 275, 2002, pages 179-213.

[5] S. ABITEBOUL, M. PREDA, G. COBENA. *Computing web page importance without storing the graph of the web.* in « IEEE-CS DataEngineering Bulletin », numéro 1, volume 25, 2002, pages 27-33.

[6] V. AGUILERA, S. CLUET, T. MILO, P. VELTRI, D. VODISLAV. *Views in a Large Scale XML Repository.* in « VLDB journal », 2002.

[7] C. DELOBEL, C. REYNAUD, M.-C. ROUSSET, J.-P. SIROT, D. VODISLAV. *Semantic Integration in Xyleme : a Uniform Tree-Based Approach.* in « Data and Knowledge Engineering Review », 2002.

[8] M. GROHE, L. SEGOUFIN. *On first-order topological queries.* in « TOCL », numéro 3, volume 3, 2002.

[9] N. PERNELLE, M.-C. ROUSSET, H. SOLDANO, V. VENTOS. *Zoom : a nested Galois lattices-based system for conceptual clustering.* in « journal JETAI (a paraitre) », 2002.

[10] M.-C. ROUSSET, A. BIDAULT, C. FROIDEVAUX, H. GAGLIARDI, F. GOASDOUÉ, C. REYNAUD, B. SAFAR. *Construction de médiateurs pour intégrer des sources d'information multiples et hétérogènes : PICSEL.* in « revue I3 », numéro 1, volume 2, 2002, pages 9-59.

[11] M.-C. ROUSSET. *The Semantic Web Needs Languages for Representing (Complex) Mappings Between (Simple) Ontologies.* in « IEEE Intelligent Systems 17 », 2002.

Communications à des congrès, colloques, etc.

[12] S. ABITEBOUL. *Issues in Monitoring Web Data.* in « International Conference on Database and Expert Systems Applications », 2002.

[13] S. ABITEBOUL, O. BENJELLOUN, I. MANOLESCU, T. MILO, R. WEBER. *Active XML : Peer-to-Peer Data and Web Services Integration (demo).* in « VLDB », 2002.

[14] S. ABITEBOUL, O. BENJELLOUN, T. MILO. *Web services and data integration.* in « International Conference on Web Information Systems Engineering », 2002.

[15] S. ABITEBOUL, G. COBENA, J. MASANES, G. SEDRATI. *A First Experience in Archiving the French Web.* in « Europan Conference on Digital Libraries », 2002.

[16] S. ABITEBOUL, G. COBENA, B. NGUYEN, A. POGGI. *Construction and Maintenance of a Set of Pages of Interest (SPIN).* in « Bases de Données Avancées (BDA) », (no proceedings), October, 2002.

[17] S. ABITEBOUL, M. PREDA, G. COBENA. *Computing web page importance without storing the graph of the web (extended abstract).* in « IEEE Data Engineering Bulletin », volume 25, March, 2002.

[18] B. AMANN, C. BEERI, I. FUNDULAKI, M. SCHOLL. *Ontology-Based Integration of XML Web Resources.* in « ISWC », 2002.

[19] B. AMANN, C. BEERI, I. FUNDULAKI, M. SCHOLL. *Querying XML Sources using an Ontology-based Mediator.* in « CoopIS (Cooperative Information Systems) also in BDA 2002 », 2002.

[20] A. BIDAULT, C. FROIDEVAUX, B. SAFAR. *Proximité entre Requêtes dans un contexte Médiateur.* in « RFIA », 2002.

[21] A. BIDAULT, C. FROIDEVAUX, B. SAFAR. *Similarity Between Queries in a Mediator.* in « ECAI », 2002.

[22] G. COBENA, T. ABDESSALEM, Y. HINNACH. *A comparative study for XML change detection.* in « Bases de Données Avancées (BDA) », (no proceedings), 2002.

[23] G. COBENA, S. ABITEBOUL, A. MARIAN. *Detecting Changes in XML Documents.* in « ICDE (Data Engineering) », 2002.

[24] E. COHEN, H. KAPLAN, T. MILO. *Labeling Dynamic XML Trees.* in « PODS », 2002.

[25] I. FUNDULAKI, B. AMANN, C. BEERI, M. SCHOLL, A.-M. VERCOUSTRE. *STYX : Connecting the XML World to the World of Semantics.* in « EDBT Demo », 2002.

[26] G. GIRALDO, C. REYNAUD. *Construction semi-automatique d'ontologies à partir de DTDs relatives à un même domaine.* in « IC », 2002.

[27] F. GOASDOUÉ, M.-C. ROUSSET. *Compilation and Approximation of Conjunctive Queries by Concept Descriptions.* in « ECAI 2002 », 2002.

[28] C. REYNAUD, B. SAFAR. *Aide à la formulation de requetes dans un mediateur.* in « RFIA », 2002.

[29] C. REYNAUD, B. SAFAR. *Representation of Ontologies for Information Integration.* in « EKAW », 2002.

[30] L. SEGOUFIN, V. VIANU. *Validating Streaming XML Documents.* in « PODS », 2002.

[31] A. TERMIER, M.-C. ROUSSET, M. SEBAG. *Mining XML Data with Frequent Trees.* in « DBFusion workshop », 2002.

[32] A. TERMIER, M.-C. ROUSSET, M. SEBAG. *TreeFinder : a First Step towards XML Data Mining.* in « International Conference on Data Mining ICDM02 », 2002.

Rapports de recherche et publications internes

[33] M. HALKIDI, B. NGUYEN, I. VARLAMIS, M. VAZIRGIANNIS. *THESUS : Organising Web Document Collections based on Semantics and Clustering.* rapport technique, Gemo, 2002.

[34] B. NGUYEN, S. ABITEBOUL. *A hash-tree based algorithm for subset detection : analysis and experiments.* rapport technique, Gemo, January, 2002.

Divers

[35] S. ABITEBOUL, O. BENJELLOUN, T. MILO, I. MANOLESCU, R. WEBER. *Active XML : A Data-Centric Perspective on Web Services.* Conférence sur les Bases de Données Avancées, 2002.

[36] V. AGUILERA, S. CLUET, T. MILO, P. VELTRI, D. VODISLAV. *Views in a Large Scale XML Repository.* VLDB Journal(to appear), 2002.