INRIA

INSTITUT NATIONAL DE RECHERCHE EN INFORMATIQUE ET EN AUTOMATIQUE

Projet VERSO

Bases de Données

Rocquencourt

THÈME 3A

Table des matières

1 Composition de l'équipe

Responsable scientifique

Serge Abiteboul [DR]

Assistantes de projet

Danièle Moreau [TR, en commun avec le projet MEVAL, jan-avr]
Sandra Civino [Agent temporaire, en commun avec le projet MEVAL, avr-oct]
Karolin Usner [Agent temporaire, en commun avec le projet MEVAL, nov-déc]

Personnel INRIA

Stéphane Grumbach [DR, responsable des Relations Internationales]
Luc Segoufin [CR]

Chercheur associé

Tova Milo [Professeur, U. Tel Aviv]

Conseillers scientifiques

Claude Delobel [Professeur, Univ. Paris 11, jan-sept]
Michel Scholl [Professeur, CNAM]

Collaborateur extérieur

Bernd Amann [Maître de Conférence, CNAM]

Chercheurs invités

Victor Vianu [Professeur, U.C. San Diego, 6 mois]
Michalis Vazirgiannis [Assistant Professor, Athens University of Economics and Business, 8 mois]

Doctorants

Vincent Aguiléra [Ingénieur Ministère Equipement, Ecole des Ponts]
Grégory Cobéna [XTélécom]
Irini Fundulaki [ATER, CNAM]
Laurent Mignet [Boursier MENRT, CNAM, jan-nov]
Benjamin Nguyen [Boursier MENRT, Orsay]
Pierangelo Veltri [ATER, Université de Villetaneuse]
Omar Benjelloun [Boursier MENRT, Orsay, oct-déc]

Chercheur post-doctorant

Kjetil Norvag [Univ. Trondheim, jan-sep]

Stagiaires

Omar Benjelloun [DEA Orsay, avr-août]
Florin Cremenescu [Polytechnique, avr-juin]
Antonella Poggi [Polytechnique, avr-juillet]
Beiting Zhu [Ecole des Mines de Nancy, juin-août]

2 Présentation et objectifs généraux

Les données sont de plus en plus complexes, distribuées, hétérogènes, répliquées, multi-formes, changeantes. L'objectif du projet est l'étude des problèmes fondamentaux posés aux systèmes de gestion de bases de données existants et le développement de solutions novatrices

appropriées. Notre but est d'obtenir des systèmes plus ouverts à des données plus riches, orientés vers le réseau.

Les problèmes nouveaux suscités par l'explosion de l'accès à un grand nombre de sources de données, notamment via le Web, demandent souvent de combiner des techniques d'intelligence artificielle et de bases de données. Pour pouvoir les étudier plus efficacement, le projet Verso a formé depuis 2001 un projet commun avec l'équipe de représentation des connaissances du LRI, Orsay (M.-C. Rousset).

Axes de recherche

Verso s'éloigne des approches serveurs-centralisés utilisées dans Active Views et surtout Xyleme, pour aller vers des solutions plus Peer-to-Peer (comme dans Active XML).

- Entrepôts de données XML du web : Dans la continuité des travaux autour de Xyleme, ces travaux ont d'abord porté sur le diff XML, l'optimisation de requêtes, et le calcul de l'importance de pages du web. Nous avons aussi débuté de nouvelles études sur la création et la fouille de collections thématiques de données.
- Active XML : nous étudions l'intégration de données et de service du web. Notre approche est orientée autour de documents XML incluant des appels déclaratifs à des services web. Il s'agit de faciliter le développement d'applications ouvertes vers le réseau, permettant plus de coopérations entre des clients distribués.
- Interrogation de données hétérogènes et optimisation : Nous avons poursuivi les travaux sur l'indexation de XML et l'intégration de données hétérogènes. De nouveaux travaux ont démarré sur le filtrage de données XML avec des automates finis.
- Théorie de la modélisation de données : Nous considérons les aspects théoriques spécifiques d'une vision de l'informatique centrée sur les données en nous appuyant sur des outils classiques de logique et de complexité pour dégager les spécificités du calcul sur des collections (relations) ou sur des graphes irréguliers (Web).

Collaborations

- Démarrage d'un projet européen DbGlobe sur l'évaluation de requêtes sur le web.
- Démarrage d'un projet franco-canadien avec U. Toronto sur les nouveaux services web.
- En France : Équipe Vertigo, CNAM.
- Industriel : avec la société Matchvision (start-up disparue en 2001, projet GAEL) et la société Xyleme (start-up issue du projet).

3 Fondements scientifiques

Mots clés : base de données, XML, Web, requêtes, contrôle des changements, services web, intégration de données, sémantique, distribution.

Les données sont au centre des recherches du projet. Cela nous conduit à utiliser des techniques venant de nombreux domaines de l'informatique : système, langage, typage, programmation, etc. Notre approche centrée autour des données introduit un biais qui modifie les

problèmes. Par exemple, l'accès efficace aux données via des langages de requêtes, conduit à utiliser des techniques de compilation et d'optimisation. Ces techniques sont très différentes de celles utilisées pour les langages de programmation. L'étude des fondements du domaine situe nos travaux souvent aux frontières d'autres théories comme la théorie de la complexité (faibles classes de complexité), la logique mathématique (notamment logique des modèles finis) ou la théorie des types.

Nous nous intéressons depuis plusieurs années aux données du web, qui sont par nature distribuées, hétérogènes, complexes, changeantes. Malgré le formidable développement du web et son utilité pour de nombreuses communautés, la recherche et l'exploitation d'informations s'y révèlent encore très difficiles. Les outils actuels de recherche ne permettent que l'expression de requêtes vagues auxquelles sont données des réponses peu précises qu'il faut péniblement analyser pour obtenir l'information réellement recherchée. Les limitations actuelles sont principalement dues à l'utilisation du langage HTML. Ce langage est très pauvre : il n'apporte aucune information structurelle ou sémantique. Pour pallier cette déficience, le *World Wide Web* consortium propose de remplacer HTML par un modèle de données semi-structurées, XML.

L'émergence de XML comme format d'échange standard pour Internet ouvre des perspectives particulièrement intéressantes pour l'utilisation des données du web. En effet, à l'opposé de HTML, XML apporte de l'information sur la structure des documents et, par là même, sur leur sémantique. Il nous paraît primordial d'exploiter cette nouvelle propriété des données du web. Le domaine des bases de données a mis une trentaine d'années à s'installer avec des bases théoriques solides (calcul et algèbre relationnels), des logiciels efficaces et fiables (Oracle, DB2, etc.). L'arrivée de XML se fait un peu dans le désordre et dans l'urgence, sous la pression de l'industrie. On voit apparaître une nuée d'outils autour de XML. Pourtant, un langage de requêtes standardisé pour XML semble émerger avec plus de difficultés. Enfin, des bases théoriques solides manquent. Elles sont indispensables à l'épanouissement du domaine et à la maîtrise de la complexité des applications que l'on met en place.

Les techniques et théories utilisées étendent celles développées pour le relationnel. Il faut noter de nouvelles dimensions. Les systèmes exigent, de par la nature du web, beaucoup plus de distribution. Une approche Peer-to-Peer remplace souvent l'approche jacobine, centralisatrice des bases de données traditionnelles. La dimension syntaxique de XML (des mots dans des grammaires particulières) est aussi nouvelle et comme l'échange de données est au centre de l'approche, elle ne peut être ignorée. Surtout, les données XML sont des arbres et en cela, se rapprochent des modèles NF2 et objets. Elles génèrent un regain d'intérêt pour des outils comme les automates d'arbre. Enfin, comme le web se situe a-priori dans un cadre global, peu ou pas contrôlé, le recours à des techniques d'intelligence artificielle, pour comprendre le contenu des données ou leur structure devient essentiel.

Pour conclure ce rapide tour d'horizon du domaine, il faut souligner que les aspects dynamiques sont importants dans le cadre du web. Cela implique de considérer des fonctionnalités comme le contrôle des changements, les souscription de requêtes et le recours à des techniques comme les règles actives. Cela veut dire aussi que les données ne sont pas uniquement statiques (venant de fichiers ou bases de données) mais qu'elles sont souvent générées à la volée, par exemple par des services web (e.g., SOAP).

4 Domaines d'applications

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 champs 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 du commerce électronique, des nouvelles applications du web (télécom et multimédia), des portails d'entreprise, 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 nous contentons de mentionner plus en détail deux applications à titre d'illustration. La première porte sur l'exploitation intelligente des données du web. Dans le cadre du projet Xyleme, nous avons considéré la création et l'utilisation d'entrepôts massifs de données du web. 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. Nous avons aussi considéré une application très à la mode : le commerce électronique. Certains aspects de cette application s'appuient sur la gestion de données hétérogènes distribuées. Nos travaux ont porté sur les spécifications déclaratives de telles applications, notamment la description des flots d'information entre leurs acteurs.

4.1 Xyleme : exploitation des données du web

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

Depuis son adoption par le W3C comme format d'échange standard pour les données de l'Internet, XML a séduit tous les grands acteurs du marché informatique.

Contrairement aux données HTML, il est possible d'extraire la structure d'un document XML. Cette structure peut être utilisée de multiples façons. Notamment, elle permet d'interroger les documents de façon plus intelligente. Par exemple, il est concevable en XML d'évaluer une requête de type : « donner les documents contenant des produits dont le prix est supérieur à 10.000 Euros ». Sur un ensemble de documents HTML, la requête la plus proche de celle-ci serait : « donner les documents contenant les mots produits et prix ». Il est également possible d'envisager l'intégration automatique d'un ensemble de documents portant sur le même sujet mais dont les structures sont légèrement différentes. Ceci facilite l'interrogation mais également la manipulation des données par des applications (par exemple, une application sur des données génomiques collectées en différents endroits de la planète).

Verso parie sur XML. Nous pensons que ce format va gagner encore en popularité et que dans un futur relativement proche, la majorité des données publiées sur le web seront XML. Nous comptons être alors présents avec des outils permettant une bonne exploitation de ces données. Une base de données construite avec des données XML du web peut servir de support à un ensemble de services tels que : interrogation conviviale et pertinente, abonnement de requêtes (e.g. prévenez-moi si un nouveau projet apparaît sur le site de l'INRIA) ou requêtes

temporelles (e.g. quelle est l'évolution du nombre de projets à l'INRIA depuis 1980).

La société Xyleme a été créée et a développé un produit à partir d'un prototype construit dans Verso.

4.2 Commerce électronique

Mots clés : règle de production, base de données active, parallélisme, déduction, datalog, workflow, commerce électronique, calcul relationnel, non-déterminisme.

Un catalogue électronique met en jeu du partage de données entre des clients et des vendeurs. Le système doit aider les clients dans leur achat. Il doit s'adapter aux utilisateurs. Dans le cas des PME, il faut aussi pouvoir, sans avoir à faire appel à des bataillons d'informaticiens, mettre en place un catalogue électronique bien adapté à l'entreprise. Nous avons proposé pour ce faire, les « vues actives ». Des vues actives permettent à un client de voir les données qui le concernent et d'interagir avec le serveur et d'autres clients par un mécanisme de notification. Les aspects novateurs de nos travaux s'articulent autour de : (i) l'utilisation de données semi-structurées, (ii) l'étude de descriptions logiques des connaissances (celles du catalogue et des profils d'utilisateurs), et (iii) une spécification déclarative des notifications (Voir Projet RNRT GAEL.)

5 Logiciels

5.1 Xyleme V0

Logiciel point de départ de la start-up du même nom. Entrepôt de données XML. Acquisition de données (crawl du web, calcul de l'importance des pages), stockage massif, indexation, requêtes, intégration sémantique de données, souscription de requêtes [terminé].

5.2 ActiveViews

Langage et système de vues actives et son application au commerce électronique (transfert, projet GAEL) [terminé].

5.3 XDIFF

Système de diff de documents XML (logiciel libre).

5.4 Thesu

Prototype de requêtes sur des collections de documents utilisant la sémantique des liens [en démarrage], collaboration avec l'U. d'Athènes.

5.5 SPIN

Outil de monitoring de site web utilisant des diff sur des collections de documents [en démarrage].

5.6 Active XML

Langage et système d'intégration de données XML et de services du web [en démarrage].

5.7 STYX

Définition et implantation d'une plate-forme générique et ouverte, permettant l'intégration et l'interrogation de ressources XML, utilisées par une communauté donnée.

6 Résultats nouveaux

6.1 Xyleme : Entrepôts de données XML du web

Participants : Serge Abiteboul, Sophie Cluet, Guy Ferran, Vincent Aguiléra, Grégory Cobéna, Laurent Mignet, Benjamin Nguyen, Pierangelo Veltri.

Mots clés : Xyleme.

En collaboration avec l'université de Mannheim et l'équipe IASI du LRI, Verso a développé en 1999-2000 le système Xyleme, un entrepôt dynamique de toutes les données XML du web. Xyleme soulève de nombreux et intéressants problèmes. Nous trouvons dans cette partie des travaux ayant été en grande partie réalisés l'an dernier même s'ils n'ont été publiés que cette année. Nous trouvons également des travaux qui prolongent cette recherche.

Acquisition et maintenance de données

La gestion de l'acquisition de nouvelles pages XML dans Xyleme ainsi que leur rafraîchissement sont présentés dans [44]. Il s'agit de guider la création d'un fond de pages XML trouvées sur le web et de le maintenir à jour dans un web changeant en permanence. On tient compte de facteurs tels que les importances respectives des pages ou les souhaits des utilisateurs du système. De nouveaux algorithmes pour calculer l'importance des pages ont été étudiés cette année.

Versions et souscriptions de requêtes

La représentation de versions en utilisant des « deltas » est présentée dans [33]. Il s'agit d'obtenir une représentation relativement compacte de l'histoire de certains documents facilitant des requêtes comme le calcul du changement depuis une version donnée. Un système de souscriptions à des requêtes a aussi été étudié [45]. Il s'agit de pouvoir être notifié, par mail ou sur une page web, lors d'événements survenus sur la base (e.g. nouveau document, changement du contenu d'un document, etc.) Enfin une analyse mathématique et expérimentale de l'algorithme de notification utilisé par Xyleme a été réalisée.

Dans cette continuation de Xyleme, se placent aussi un certain nombre de travaux sur l'évaluation de requêtes.

6.2 Active XML

Participants : Serge Abiteboul, Omar Benjelloun, Tova Milo.

Mots clés : XML, base de données active, parallélisme, distribution, datalog, service web, SOAP, WSDL, UDDI.

Avec Active XML, nous proposons une approche orientée-données de l'utilisation et de l'intégration de services web répartis. Nous introduisons un nouveau formalisme, basé sur l'inclusion d'appels de services à l'intérieur de documents semi-structurés, et s'appuyant sur des standards émergeants du web, tels que SOAP et WSDL. Active XML capture de nombreux scénarios classiques d'intégration de données, tels que la médiation et l'entreposage de données, et autorise la fusion de données par le biais d'identifiants d'objets. De plus, du fait qu'il autorise les paramètres d'appels de services ainsi que les valeurs qu'ils retournent à contenir à leur tour des appels vers d'autres services, ce formalisme permet une certaine forme de calcul distribué à l'échelle du web.

Ces travaux sont dans le prolongement de nos travaux sur les « vues actives ». Ils ont démarré cette année. Nous avons spécifié le langage ActiveXML. Un premier prototype a été implanté. Nous étudions la sémantique du langage en utilisant datalog et des logiques de point-fixe non-déterministes (asynchronisme de la distribution) et non-inflationistes (changements des données). Nos premiers résultats ont été présentés au workshop FMII sur l'intégration d'information [12]

6.3 Interrogation de données hétérogènes et optimisation

Participants : Serge Abiteboul, Bernd Amann, Gregory Cobena, Irini Fundulaki, Benjamin Nguyen, Michel Scholl, Michalis Vazirgianis.

Mots clés : Web, HTML, XML, requêtes, liens, contrôle des changements, hétérogénéité, intégration, ontologie, thésaurus, linéarisation de thésaurus.

Gestion de vues

Notre objectif est de permettre aux utilisateurs du web de formuler des requêtes précises (par exemple, le nom et l'adresse des responsables des ventes des sociétés informatiques de la région parisienne). Pour ce faire, nous nous proposons de découper le web en domaines (la culture, le tourisme, les affaires, etc.), chacun étant décrit par une structure simple (comparable à un formulaire avec imbrication de champs). Une interrogation consiste, alors, à annoter cette structure.

Chaque site du web a une structure différente. Il serait illusoire d'imaginer que le passage à XML va changer cet état de fait. Pour associer à chaque domaine une structure unique, il est donc nécessaire de comprendre les correspondances qui existent entre leur structure réelle et celle proposée par Xyleme. Ces dernière années, nous avions travaillé sur un système de vues XML très puissant mais qu'il n'est pas concevable d'utiliser dans le contexte Xyleme, où le passage à l'échelle est primordial. En effet, la taille d'une vue (ensemble des correspondances)

est fonction du nombre de structures réelles. Ce nombre tend à croître de façon exponentielle et il n'y a, a priori, pas de limite à cette extension. Cependant, pour permettre une évaluation rapide des requêtes, il est nécessaire de stocker dans une seule mémoire, une synthèse de la vue permettant notamment de déterminer les machines qui vont pouvoir répondre à la requête. Également, il est important de comprendre comment traduire une requête portant sur une structure abstraite en des requêtes sur les documents réels de façon à minimiser la communication entre machines. Dans [23, 15], nous proposons une solution à ces problèmes, solution que nous avons implantée dans le système Xyleme.

Evaluation de requêtes

Xyleme doit pouvoir répondre en un temps raisonnable à des milliers de requêtes concurrentes portant sur des milliards de documents répartis sur de nombreuses machines. Pour ce faire, nous avons choisi de partitionner les documents suivants des critères sémantiques afin de limiter le nombre de machines impliquées dans une requête. Également, nous avons étendu la technique d'indexation plein-texte afin de prendre en compte la structure des documents. Plus particulièrement, nous associons à chaque occurrence d'un mot dans un document un code permettant, étant donné deux mots, de comprendre leur position relative dans le document. Cette technique permet une évaluation extrêmement rapide des requêtes de type formulaire imbriqué introduites précédemment.

Collections de documents

Nous étudions comment spécifier et construire des collections de documents dans un domaine spécifique, notamment en utilisant les outils du web. On veut ensuite pouvoir surveiller l'évolution de ces collections (contrôle des changements). Enfin, nous étudions de nouvelles techniques de requêtes basées sur les liens et de la sémantique que l'on peut attacher à ces liens.

Cweb

Nous avons commencé de développer le prototype *STYX* (Connecting the XML web to the World of Semantics). Cette recherche est dans la continuité des travaux du projet C-web (`http://cweb.inria.fr`), terminé en 2000. L'objectif de ce projet était la définition et l'implantation d'une plate-forme générique et ouverte, permettant l'intégration et l'interrogation de ressources XML, utilisées par une communauté donnée, qui souhaite partager les connaissances d'un domaine spécifique. Notre travail actuel concerne la *médiation de données* et illustre une nouvelle approche pour la publication et l'interrogation de ressources XML dans le web. Les résultats de ces travaux [18] [19] sont :

1. un *langage de règles d'association* entre les chemins qui existent dans une DTD XML qui décrit la structure d'une source XML et les chemins dans le schéma permettant à l'utilisateur de poser des requêtes ;
2. un algorithme de *re-écriture* qui traduit une requête arbre en un ensemble des requêtes XML et

3. un premier prototype qui démontre les fonctionnalités d'intégration et d'interrogation de documents XML originaires de ressources hétérogènes.

6.4 Modélisation de données

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

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

Données semi-structurées

Les données accessibles sur le réseau vont du très structuré, dans des bases relationnelles, au totalement non structuré, dans des fichiers de texte. De plus, l'intégration de données est souvent source d'irrégularité, des données similaires étant souvent représentées avec des structures différentes dans des sources indépendantes. On utilise le terme *semi-structuré* pour ces données qui ne sont pas vraiment structurées mais qui présentent une certaine structure même si celle-ci est peu régulière et implicite.

En pratique l'information présente sur le réseau est très incomplète. Cela peut être dû a des capacités de stockage limitées, la nature dynamique et changeante de la toile, etc. Dans [13], nous étudions la représentation et l'interrogation de données semi-structurées incomplètes. On montre la difficulté intrinsèque de l'approche tout en isolant un scénario restreint mais faisable.

Langages de requêtes

Le langage de requête est la partie émergée de l'iceberg des bases de données. En effet c'est l'outil principal mis à la disposition de l'utilisateur afin d'accéder aux données. Dans le cadre des bases de données relationnelles classiques, les données sont souvent considérées comme atomiques et le langage de requête principal est SQL.

L'évaluation de SQL est maintenant bien maîtrisée et se fait, après optimisation, de manière quasi optimale. La complexité correspondante n'est pourtant pas encore bien comprise. Si elle reste raisonnable par rapport à la taille des données : polynômiale dans le cas le pire, parfois linéaire voir logarithmique en présence d'index, elle est, dans le cas pire, exponentielle dans la taille de la requête. Pourtant, dans la pratique, l'évaluation s'effectue rapidement car nombreuses sont les requêtes ayant un algorithme d'évaluation polynômial.

Dans [26], on caractérise les familles de requêtes ayant une évaluation polynômiale, aussi bien dans la taille des données que celle de requêtes.

Lorsque les données ne peuvent plus être considérées comme atomiques, par exemple dans le cas de la bio-informatique où les données sont des chaînes de caractères correspondant à des séquences de gènes, les extensions de SQL proposées actuellement sont peut satisfaisantes. En effet soit le pouvoir d'expression du langage est trop limité soit il permet d'exprimer des requêtes non calculables. Dans [21] et dans [20] on propose divers langages de requêtes sur les chaînes de caractères. Pour chacun d'entre eux on étudie son pouvoir d'expression, la complexité nécessaire pour l'évaluer et on propose une algèbre physique correspondante.

Données spatiales

Verso a une longue expertise dans le domaine des données spatiales. Cette année a vu la publication dans des revues internationales de papiers reflétant notre apport : [9] [8], [7].

Labélisation des noeuds de documents XML

Motivés par des applications comme Xyleme, nous étudions des schémas de labélisation des noeuds de documents XML. On doit pouvoir à partir des labels déterminer si un noeud est un ancêtre d'un autre noeud (par exemple, pour accélérer l'évaluation de requêtes). Un aspect important est la taille des labels qui détermine la taille des index. Nous avons étudié le problème dans le cas statique (sans mises-à-jour). Nous avons proposé de nouveaux schémas et les avons comparés à ceux utilisés par des moteurs de recherche et ce tant d'un point de vue théorique, que pratique (sur des documents trouvés sur le web). Nos schémas ont de meilleures performances [30]. Nous poursuivons cette étude par des travaux dans un cadre dynamique qui supporte des changements des documents. Nous avons un schéma qui est optimal à une constante près. Nous montrons aussi comment on peut utiliser des connaissances sur le document, comme des statistiques sur les futures insertions dans des labels particuliers.

7 Contrats industriels (nationaux, européens et internationaux)

7.1 Xyleme

La société Xyleme SA a été créée en septembre 2000 à partir de travaux du projet Verso. Xyleme a quitté Verso en mars 2001. Notre collaboration avec Xyleme continue.

7.2 MatchVision

Nous avons collaboré avec MatchVision dans le cadre du projet RNRT GAEL. MatchVision a déposé son bilan en 2001.

8 Actions régionales, nationales et internationales

8.1 Actions nationales

8.1.1 Projet RNRT GAEL

Ce projet, qui réunit une start-up MatchVision, l'équipe IASI du LRI (U. Paris Sud) et le projet Verso, s'est terminé cette année.

L'objectif du projet était de concevoir un générateur de catalogues électroniques et de services associés, permettant à des non informaticiens de développer leurs propres catalogues (boutiques, galeries marchandes) électroniques sans gros effort de programmation et en les personnalisant à leurs besoins. Pour ce faire, on combine des techniques de représentation des connaissances et d'agents intelligents. L'effort de recherche porte principalement sur une spécification déclarative des connaissances relatives au commerce électronique (les différents types de produits, de profils clients, d'actes commerciaux), ainsi que des informations présentées aux

clients en XML. Nous utilisons des agents intelligents pour doter les catalogues électroniques de comportements dynamiques permettant de réagir de façon adéquate aux achats en cours des clients en fonction du contenu du catalogue, de la stratégie commerciale spécifiée de façon déclarative par le vendeur, ainsi que du profil des clients.

Le responsable Verso pour GAEL est S. Abiteboul.

8.1.2 Livre

Saluons la publication d'un livre, *Spatial Databases*, P. Rigaux, M. Scholl, A. Voisard chez Morgan Kaufmann. M. Scholl est conseiller scientifique de Verso et A. Voisard a fait sa thèse dans Verso.

8.1.3 Autres Collaborations Nationales

Des liens étroits existent avec le Labri (B. Courcelle), le LRI (en plus du IASI, l'équipe BD, N. Bidoit, E Waller), le Cedric au CNAM (M. Scholl, B. Amann, P. Rigaux, D. Vodislav), et l'équipe Caravel de l'INRIA (F. Llirbat).

8.2 Actions financées par la commission européenne

8.2.1 Projet DbGlobe

Le projet DbGlobe (qui démarre en 2002) a pour but de développer de nouvelles techniques de gestion de données pour traiter du calcul global sur le web. Il s'agit de comprendre comment concevoir, construire, et analyser des systèmes permettant de gérer de gros volumes de données sur le web. Il faut tenir compte de l'autonomie des sources d'information, et éventuellement de leur mobilité.

Le projet est dirigé par l'université de Ioanina en Grèce. Le responsable Verso pour DbGlobe est S. Abiteboul.

8.2.2 Projet Mesmuses

Le Projet Européen Mesmuses (Metaphor for Science Museums) a démarré début janvier 2001. Il se situe dans la continuité du projet Cweb (voir rapport d'activité 2000). L'objectif est l'implantation de la plate-forme Cweb et l'installation de deux applications dans le contexte de la préparation d'expositions scientifiques. L'objectif principale est la réutilisation optimale de sources électroniques (documents, images, son, vidéo) créées pendant la préparation d'une exposition ou disponibles dans les médiathèques pour la création de sites web interactifs destinés au grand public.

Le responsable Verso pour Mesmuses est B. Amann.

8.2.3 GDR-PSIG Cassini

Cassini est un programme de recherche national pluridisciplinaire (informaticiens, géographes) financé par le CNRS et l'Institut Géographique National (IGN) sur l'information géographique. Dans ce programme, le projet Verso, en collaboration avec le CNAM, le LSR

de l'IMAG et le LAMA à Grenoble, étudie les problèmes de modélisation d'objets évoluant au cours du temps.

Le responsable Verso pour Cassini est M. Scholl.

8.3 Réseaux et groupes de travail internationaux

Verso est membre du réseau d'excellence Compulog (logic programming) et DELOS (European Digital Libraries) et participe au groupe « Bases de données » de l'Ercim.

8.4 Relations bilatérales internationales

8.4.1 Europe

Nous collaborons avec l'Université de Mannheim (G. Moerkotte), l'Université de Marburg (T. Schwentick), l'U. Athenes (M. Vazirgianis) et l'ETH Zurich (R. Weber).

8.4.2 Moyen-Orient

Un financement par l'Association Franco-Israélienne pour la Recherche Scientifique et Technique (AFIRST) sur le projet « Les Usines du Futur » s'est terminé en 2001. Nous poursuivons notre collaboration avec Tel Aviv (T. Milo) et Hebrew U. (C. Beeri).

8.4.3 Amérique du Nord

En Amérique du Nord, des liens existent avec notamment l'Université de Stanford (J. Widom), Pennsylvanie (P. Buneman), UC Santa Barba (J. Su), UC San Diego (V. Vianu), ATT (Sihem Amer-Yahia, Divesh Srivastava), Lucent-Bell (Jérôme Siméon). Un projet Franco-Canadien de collaboration avec l'université de Toronto (A. Mendelzon et Léonid Libkin) démarre en 2001. L. Mignet part en post doc à Toronto.

8.4.4 Asie et océan Pacifique

Après plusieurs années de collaboration avec le CSIRO à Melbourne, A.-M. Vercoustre est partie pour 2 ans dans cette équipe.

8.5 Accueil de chercheurs étrangers

Cette année, nous avons accueilli :
- Tova Milo, professeur à l'Université de Tel-Aviv (1 an)
- Michalis Vazirgiannis, chercheur à l'Université d'Athènes (8 mois)
- Victor Vianu, professeur, UC San Diego (6 mois)

Nous avons également accueilli pour de courtes visites d'autres chercheurs étrangers avec lesquels nous avons des nous avons des collaborations suivies : Vassilis Christophides (FORTH-Crète).

9 Diffusion de résultats

9.1 Actions d'enseignement

S. Abiteboul est professeur à temps partiel à l'Ecole Polytechnique.
B. Amann est maître de conférence au CNAM-Paris.
C. Delobel est professeur à l'Université de Paris 11. Il a pris sa retraite en septembre 2001.
M. Scholl est professeur au CNAM.
V. Vianu est professeur à UCSD.

Les cours suivants ont été assurés par plusieurs membres de l'équipe :

- *SGBD objets et avancés,* DEA Systèmes Informatiques, cohabilité Paris 6-Telecom-CNAM, B. Amann, M. Scholl.
- *SGBD,* E. Polytechnique, S. Abiteboul.
- *Théorie des modèles finis,* DEA de Paris VII, S. Grumbach, L. Segoufin et S. Abiteboul.
- *Bases de données semi-structurées,* DEA I3 de Paris XI, S. Abiteboul et V. Aguiléra.
- *Algorithmique et Structures de Données,* Travaux dirigés, Licence d'Informatique - Paris XI, B. Nguyen
- *Initiation à la programmation impérative,* Travaux dirigés, DEUG MIAS - Paris XI, B. Nguyen
- *Logique et matériel,* Travaux pratiques, DEUG MIAS - Paris XI, B. Nguyen
- *Base de données II,* Travaux dirigés, IUP troisième année, MIAGE - Paris XI, O. Benjelloun.
- *Intégration d'informations hétérogènes,* Travaux dirigés, CFA deuxième année, MIAGE - Paris XI, O. Benjelloun.
- *Systèmes d'information,* Module de deuxième année, Ecole Nationale des Ponts et Chaussées, V. Aguiléra
- *Bases de Données,* Troisième année, voie d'approfondissement Informatique, Ecole Nationale des Travaux Publics de l'Etat, V. Aguiléra
- *Base de données,* Travaux pratiques et Travaux dirigés, Première Année, Institut d'Informatique d'entreprise - CNAM, I. Fundulaki
- *Base de données,* Travaux dirigés, Deuxième Année, Institut d'Informatique d'entreprise - CNAM, I. Fundulaki
- *Systèmes d'Information Automatisés,* Travaux dirigés, Deuxième Année, Institut d'Informatique d'entreprise - CNAM, I. Fundulaki
- *Le Langage des Documents et des Données Semi-Structurées XML*, Séminaire, Deuxième Année, Institut d'Informatique d'entreprise - CNAM, I.Fundulaki
- *Informatique,* Travaux dirigés et Méthodologie, Première Année (MIAS), Institut d'informatique Galilée - Université de Paris XIII, P. Veltri
- *Bases de données,* Travaux dirigés, et Travaux Pratiques Première Année Formation d'ingénieurs, Institut d'informatique Galilée - Université de Paris XIII, P. Veltri

9.2 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
- ACM SIGMOD Conference on the Management of Data, 2002.
- PC Chair of International Conference on Very Large Databases, 2003.
- International Workshop on web Dynamics, London (2001)
- 4th Workshop on the web and Databases, Santa Barbara

B. Amann
- International Conference on Extending Database Technology (EDBT), 2002
- International Conference and Workshop on Database and Expert Systems Applications (DEXA, 2001 et 2002)
- Workshop Data Integration over the Web (CAiSE), 2001

S. Grumbach
- Thirteenth International Conference on Scientific and Statistical Database Management (SSDBM), 2001
- The Second International Conference on Web-Age Information Management (WAIM), 2001
- ACM Principles of Databases Systems (PODS), 2002
- 18th IEEE International Conference on Data Engineering (ICDE), 2002

L. Mignet
- Workshop Data Integration over the Web (CAiSE), 2001

T. Milo
- International Conference on Very Large Databases, 2001
- International Workshop on Formal Model for Information Integration, 2001.
- International Conference on Data Engineering, 2002.

M. Scholl
- ACM SIGMOD Conference on the Management of Data, 2001.
- Symposium on Spatial and Temporal Databases (SSTD01)
- Conférences sur les Bases de Données Avancées (BDA)
- Conference Cassini

L. Segoufin
- ACM Principles of Databases Systems (PODS), 2001.
- International Conference on Database Theory (ICDT), 2001.
- Bases de Données Avancés (BDA), 2001.

M. Vazirgianis
- International Symposium on Methodologies for Intelligent Systems, (ISMIS'02), 2002
- International Conference on Multimedia Modeling (MMM2001)
- Symposium on Document Engineering, 2001
- IEEE International Conference on Multimedia and Expo (ICME2001)

V. Vianu

- PC co-chair of International Conference on Database Theory, 2001.
- WWW-10 Conference, 2001.
- International Workshop on Database Programming Languages, 2001.
- Second International Conference web-Age Information Management, 2001.
- Association for Symbolic Logic Meeting, 2002.
- ACM SIGMOD Conference on the Management of Data, 2002.

9.2.1 Conférences invitées, tutoriels, cours, etc.

En 2001, S. Abiteboul a été conférencier invité à un séminaire Dagstuhl sur les langages pour données structurées et la International Workshop on Formal Model for Information Integration (Integration of data and web services). Il a participé à un VLDB panel sur les services web.

M. Vazirgianis a présenté un tutoriel sur la fouille de données à ADBIS 2001 Conference, Vilnius, Lithuania.

V. Vianu a été conférencier invité à PODS (A web Odyssey : From Codd to XML), au séminaire de l'INRIA-Rocquencourt et à un workshop NSF (The Unusual Effectiveness of Logic in Computer Science).

En 2001 L. Segoufin a été conférencier invité de l'École Jeunes Chercheurs du CNRS à l'ENS-Lyon.

9.2.2 Animations scientifiques

En 2001, S. Abiteboul est président du comité exécutif de l'ACM International Symposium on Principles of Database Systems, et membre des comités exécutifs de ACM SIGMOD on the Management of Data et de Logic in Computer Science.

V. Vianu est membre des comités exécutifs de PODS et ICDT, et du conseil scientifique de l'ENS Bucarest.

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

Participations à des revues scientifiques :

- S. Abiteboul est membre des comités d'édition de ACM Transactions on Database Systems, Information and Computation, et Journal of Digital Libraries.
- V. Vianu est membre des comités d'édition de JACM, ACM Transactions on Computational Logic, ACM Sigmod Digital Reviews, Discrete Mathematics and Theoretical CS ; éditeur de la Database Theory column of Sigact News et d'une issue spéciale de TCS sur ICDT2001.

10 Bibliographie

Ouvrages et articles de référence de l'équipe

[1] S. Abiteboul, P. Buneman, D. Suciu, *Data on the Web : From Relations to Semistructured Data and XML*, Morgan-Kaufman, New York, 1999.

[2] S. Abiteboul, R. Hull, V. Vianu, *Foundations of Databases*, Addison-Wesley, 1995.

[3] S. Abiteboul, V. Vianu, « Computing With First-Order Logic », *Journal of Computer and System Sciences(JCSS) 50(2)*, 1995, p. 309–335.

[4] S. Grumbach, P. Rigaux, L. Segoufin, « The DEDALE System for Complex Spatial Queries », *in : sigmod*, 1998.

[5] P. Rigaux, M. Scholl, A. Voisard, *Spatial Databases with application to GIS*, Morgan Kaufman, May 2001.

Livres et monographies

[6] P. Rigaux, M. Scholl, A. Voisard, *Spatial Databases with application to GIS*, Morgan Kaufman, May 2001.

Articles et chapitres de livre

[7] M. Benedikt, M. Grohe, L. Libkin, L. Segoufin, « Reachability and Connectivity Queries in Constraint Databases », *Journal of Computer and System Sciences (JCSS)*, To Appear.

[8] M. Grohe, L. Segoufin, « On First-order Topological Queries », *ACM Transactions on Computational Logic (TOCL)*, To Appear.

[9] S. Grumbach, P. Rigaux, L. Segoufin, « Spatio-Temporal Data Handling with Constraints », *GeoInformatica 5*, 1, March 2001, p. 95–115.

[10] M. Halkidi, Y. Batistakis, M. Vazirgiannis, « On Clustering Validation Techniques », *Intelligent Information Systems Journal*, 2001, To Appear.

[11] L. Xyleme, « A Dynamic Warehouse for XML Data of the Web », *IEEE - Data Engineering Bulletin 24*, 2, June 2001, p. 40–47.

Communications à des congrès, colloques, etc.

[12] S. Abiteboul, O. Benjelloun, T. Milo, « Towards a Flexible Model for Data and Web Services Integration », *in : Proc. of the International Workshop on Foundations of Models for Information Integration (FMII)*, Viterbo - Italy, September 2001.

[13] S. Abiteboul, L. Segoufin, V. Vianu, « Representing and Querying XML with Incomplete Information », *in : Proceedings of the ACM Conference on Principle of Database Systems (PODS)*, ACM, p. 150–161, Santa Barbara- California, May 2001.

[14] S. Abiteboul, « Semistructured Data : from Practice to Theory », *in : LICS 2001 : IEEE Symposium on Logic in Computer Science*, IEEE Computer Society, Boston - Massachusetts, June 2001.

[15] V. Aguilera, S. Cluet, P. Veltri, D. Vodislav, F. Wattez, « Querying a Web Scale XML Repository », *in : Sistemi Evoluti per Basi di Dati (SEBD)*, p. 105–118, Venice- Italy, June 2001.

[16] N. Alon, T. Milo, F. Neven, D. Suciu, V. Vianu, « Typechecking XML Views on Relational Databases », *in : LICS 2001 : IEEE Symposium on Logic in Computer Science*, IEEE Computer Society, Boston - Massachusetts, June 2001.

[17] N. Alon, T. Milo, F. Neven, D. Suciu, V. Vianu, « XML with Data Values : Typechecking Revisited », *in : Proceedings of the ACM Conference on Principle of Database Systems (PODS)*, ACM, Santa Barbara- California, May 2001.

[18] B.Amann, C. Beeri, I.Fundulaki, M. Scholl, A.-M. Vercoustre, « Mapping XML Fragments to Community Web Ontologies », *in : Informal Proceedings of the Fourth International Workshop on Web and Databases WebDB*, Santa Barbara, California, USA, May 2001.

[19] B.Amann, C. Beeri, I.Fundulaki, M. Scholl, A.-M. Vercoustre, « Rewriting and Evaluating Tree Queries with XPath », *in : Proceedings of the 17ièmes Journées Bases de Données Avancées (BDA)*, Agadir, Maroc, November 2001.

[20] M. Benedikt, L. Libkin, T. Schwentick, L. Segoufin, « A Model-Theoretic Approach to Regular String Relations », *in : LICS 2001 : IEEE Symposium on Logic in Computer Science*, IEEE Computer Society, Boston - Massachusetts, June 2001.

[21] M. Benedikt, L. Libkin, T. Schwentick, L. Segoufin, « String Operations in Query Languages », *in : Proceedings of the ACM Conference on Principle of Database Systems (PODS)*, ACM, p. 183–194, Santa Barbara- California, May 2001.

[22] N. Bidoit, S. de Amo, L. Segoufin, « Propriétés temporelles indépendantes de l'ordre », *in : Bases de Données Avancées (BDA)*, Agadir - Maroc, October 2001.

[23] S. Cluet, P. Veltri, D. Vodislav, « View in a Large Scale XML Repository », *in : Proceedings of 27th International Conference on Very Large Data Bases*, Morgan Kaufmann, Roma - Italy, September 2001.

[24] G. Cobena, S. Abiteboul, A. Marian, « Detecting Changes in XML Documents », *in : Base de données avancées*, Agadir - Maroc, October 2001.

[25] G. Cobena, S. Abiteboul, A. Marian, « Detecting Changes in XML Documents », *in : International Conference in Data Engineering*, S. Jose - California U.S.A., Mars 2002.

[26] M. Grohe, T. Schwentick, L. Segoufin, « When is the Evaluation of Conjunctive Queries Tractable? », *in : 33rd ACM Symposium on Theory of Computing (STOC)*, Crete, Greece, July 2001.

[27] M. Halkidi, Y. Batistakis, M. Vazirgiannis, « Clustering algorithms and validity measures », *in : Proceedings of the Conference on Scientific and Statistical Database Management*, Virginia, USA, September 2001.

[28] M. Halkidi, M. Vazirgiannis, « A data set oriented approach for clustering algorithm selection », *in : Principles and Practice of Knowledge Discovery in Databases (PKDD)*, Freiburg, Germany, September 2001.

[29] M. Halkidi, M. Vazirgiannis, « Clustering Validity Assessment : Finding the optimal partitioning of a data set », *in : Proceedings of the IEEE International Conference on Data Mining (ICDM)*, California, November 2001.

[30] H. Kaplan, T. Milo, R. Shabo, « Comparison of Labeling Schemes for Ancestor Queries », *in : 13th ACM-SIAM Symposium on Discrete Algorithms, (SODA), To Appear*, 2002.

[31] H. Kaplan, T. Milo, « Short and Simple Labels for Small Distances and other Functions », *in : Proc. of the 7th International Workshop on Algorithms and Data Structures (WADS)*, Providence, August 2001.

[32] S. Lazaridis, M. Vazirgiannis, T. Sellis, « STEDEL : Spatiotemporal Definition Language, Modeling and rendering interactive 3D compositions », *in : Proceedings of International conference on Multimedia and Expo (ICME)*, Tokyo, August 2001.

[33] A. Marian, S. Abiteboul, G. Cobéna, L. Mignet, « Change-Centric Management of Versions in an XML Warehouse », *in : Proceedings of 27th International Conference on Very Large Data Bases*, Morgan Kaufmann, Roma - Italy, September 2001.

[34] F. Neven, T. Schwentick, V. Vianu, « Towards Regular Languages over Infinite Alphabets », *in : Symp. on Mathematical Foundations of Computer Science (MFCS)*, p. 560–572, 2001.

[35] B. Nguyen, S. Abiteboul, G. Cobena, M. Preda, « Monitoring XML Data on the Web », *in : Proceedings of the ACM SIGMOD Conference on Management of Data*, S. M. Timos Sellis (éditeur), ACM, p. 437–448, Santa Barbara- California, May 2001.

[36] G. P. R. Pitkaanen, M. Vazirgiannis, « The role of streaming in Interactive Multimedia Documents dissemination », *in : Proceedings of International conference on Multimedia and Expo (ICME)*, Tokyo, August 2001.

[37] S.Grumbach, L. Tininini, « Automatic Aggregationn Using Explicit Metadata », *in : Sistemi Evoluti per Basi di Dati (SEBD)*, Venice- Italy, June 2001.

[38] I. Varlamis, M. Vazirgiannis, P. Poulos, G. Akrivas, S. Ioannou, « X-Database. A middleware for collaborative video annotation, storage and retrieval », *in : Proceedings of the 8th Panhellenic Conference*, Cyprus, November 2001.

[39] I. Varlamis, M. Vazirgiannis, « Bridging XML-Schema and Relational Databases. A System for Generating and ManipulatingRelational Databases Using Valid XML Documents », *in : ACM Symposium on Document Engineering*, Atlanta- U.S.A., November 2001.

[40] I. Varlamis, M. Vazirgiannis, « Web document searching using enhanced hyperlink semantics based on XML », *in : IDEAS*, Grenoble, 2001.

[41] P. Veltri, « Constraint Database Query Evaluation with Approximation », *in : ITCC 2001, International Conference on Information Technology : Coding and Computing*, IEEE Computer Society, p. 634–638, Las Vegas-Nevada, April 2001.

[42] V. Vianu, « A Web Odyssey : From Codd to XML », *in : Proceedings of the ACM Conference on Principle of Database Systems (PODS)*, Santa Barbara- California, May 2001.

Rapports de recherche et publications internes

[43] L. Mignet, M. Preda, S. Abiteboul, A. Marian, B. Amann, « Acquisition and Maintenance of XML Data from the Web », *Rapport interne n°188*, INRIA / Verso Project, 2001.

Divers

[44] L. Mignet, V. Aguiléra, S. Ailleret, P. Veltri, « XyRo : The Xyleme Robot Architecture », First workshop Data Integration over the Web, June 2001.

[45] B. Nguyen, S. Abiteboul, G. Cobena, L. Mignet, « Querying Subscription in an XML Webhouse », First DELOS Workshop on Digital Libraries, December 2000.