Projet VERSO

Bases de Données

Localisation : *Rocquencourt*

Mots-clés : modèle de base de données, théorie des bases de données, optimisation de requête, SGBD objet, base multimédia, base de données cartographiques, document électronique, document virtuel, World Wide Web, génome.

1 Composition de l'équipe

Responsables scientifiques

Sophie Cluet [CR Inria]
Stéphane Grumbach [CR Inria]

Secrétaire

Danièle Moreau [AJA Inria, en commun avec le projet MEVAL]

Personnel Inria

Anne-Marie Vercoustre [DR Inria - Oct-Déc]

Conseillers scientifiques

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

Chercheurs extérieurs

Bernd Amann [Maître de Conférence, CNAM]

Chercheurs invités

Serge Abiteboul [Professeur Stanford, 1 mois]
Gabriel Kuper [Chercheur ECRC, 6 mois]
Tova Milo [Professeur, U. Tel Aviv, 2 semaines]
Guido Moerkotte [Professeur, U. Manheim, 1 mois]
Jianwen Su [Professeur UCSB, 1 mois]
Victor Vianu [Professeur, UC San Diego, 2 mois]

Chercheurs post-doctorants

Daniel Chan [Boursier TMR, Oct-Déc]
Cassio Souza dos Santos [Boursier post-doc industriel à O_2Technology, Janv-Sept, Collaborateur Extérieur, Oct-Déc]

Chercheurs doctorants

Sihem Amer-Yahia [Boursière Algérie, Univ. Paris 11]
Philippe Brèche [Univ. Paris 11, Ingénieur SSII]
Vassilis Christophides [ATER, CNAM Janv-Sept, puis Ing. Expert, Médiaculture, Oct-Déc]
Zoé Lacroix [ATER Paris 10, Janv-Août, puis U. Penn. Sept-Déc]
Luc Ségoufin [ENS, Service Militaire, UC San Diego, Janv-Avril, puis Collaborateur Extérieur, Mai-Sept, puis ENS, Oct-Déc]
Jérôme Siméon [Boursier Inria]
Fariza Tahi [Action Génome, ATER, CNAM]

Stagiaires

Patrick Laonet [Médiaculture, Juin-Oct]
Pini Mogilevski [U. Tel Aviv, 2 semaines, Déc]

2 Présentation du projet

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 appropriées. Suivant une tradition bien établie, nous étudions les problèmes sur le plan théorique et sur le plan pratique, le dernier s'appuyant sur les résultats du premier tout en s'adaptant aux réalités du monde industriel avec lequel nous avons des partenariats.

L'émergence du *World Wide Web* et des nouvelles applications multimédia a considérablement modifié le paysage de la recherche en bases de données. Les systèmes du futur permettront de trouver, d'intégrer et d'offrir des données variées et complexes, stockées suivant différents formats sur des sites distribués géographiquement. Verso se tourne naturellement vers ces nouveaux axes tout en continuant son travail sur l'amélioration des systèmes à objets et les fondements théoriques des bases de données. Notre action de recherche s'articule donc autour de quatre thèmes : (i) bases de données multimédia, (ii) intégration de données hétérogènes et distantes, (iii) amélioration des fonctionnalités des bases de données à objets et (iv) fondements des langages de bases de données.

Aucun système de bases de données existant ne permet de manipuler de façon satisfaisante les données multimédia. Le marché du multimédia est dominé par des spécialistes qui, dans la plupart des cas, utilisent des systèmes de fichiers ou qui adaptent difficilement des systèmes de bases de données à leurs besoins. Les expériences passées de Verso sur les données géographiques, hypertextuelles ou documentaires nous laissent penser qu'il est illusoire de vouloir créer un système universel. Nous pensons que l'avenir verra des systèmes spécialisés pouvant éventuellement communiquer entre eux. Ce premier axe se consacre donc à l'élaboration de tels systèmes. Plus particulièrement, nous nous intéressons maintenant aux bases de données spatio-temporelles et aux bases de données textuelles.

Nos travaux sur les données spatio-temporelles s'effectuent dans le cadre du nouveau projet européen Chorochronos, tandis qu'un prototype permettant le chargement et l'interrogation de données textuelles dans un système à objets a été développé dans le cadre du projet Aquarelle[1].

Complément naturel de ce premier axe, notre deuxième axe de recherche s'intéresse à la communication entre systèmes spécialisés. En particulier, nous pensons bien sûr à l'ouverture sur de nombreux sites accessibles via le *Web*. Les systèmes de bases de données peuvent jouer dans le contexte *Web* un rôle important de fédération. Des langages de requêtes évolués permettront une extraction intelligente de l'information distante dont seules les caractéristiques seront stockées dans les systèmes de bases de données. Ainsi, les systèmes de bases de données pourront être utilisés comme serveurs de *Web* et donner accès à des informations locales au système ou distantes de façon homogène. Ceci permettra

[1] Aquarelle est un projet Européen dont le responsable INRIA est A. Michard et dont le but est de développer une offre de services d'information et d'applications informatiques, permettant l'accès aux représentations numériques du patrimoine culturel européen.

d'améliorer considérablement la gestion des pages Web dont le contenu change souvent et doit être maintenu à jour. Parmi les nombreux problèmes que pose le Web, Verso travaille plus particulièrement sur les langages de requêtes et leur optimisation ainsi que sur les outils qui permettront l'intégration *virtuelle* des données, tant au niveau serveur qu'au niveau client (documents virtuels). Ce travail s'effectue dans le cadre des projets Esprit IV Opal et Wire.

Notre troisième axe de recherche s'attache à effacer certaines limitations importantes des systèmes de gestion de bases de données à objets. Cette année, tout en continuant des travaux sur l'évolution de schéma et les mécanismes de vues, nous nous sommes plus particulièment intéressés à la réutilisation de l'existant relationnel dans les systèmes à objet. La première partie de ce travail concerne la migration/réplication des données. La deuxième partie concerne l'utilisation du langage relationnel SQL dans les systèmes à objets. Les deux aspects sont traités dans le cadre d'une collaboration avec la société O_2Technology.

Enfin, nous continuons nos travaux sur les fondements théoriques des bases de données. Certains de ces travaux ont cette année rejoint des axes plus appliqués. Notamment, le modèle de bases de données contraintes qui est utilisé pour traiter les applications spatio-temporelles et de nouveaux travaux sur l'interrogation du *Web*. D'autres travaux concernent la complexité des langages de requêtes et les langages non-déterministes.

Notre travail complète naturellement les recherches réalisées dans le projet Rodin (Systèmes de Bases de Données) et notamment les travaux développés dans le cadre de l'action Disco. L'équipe Verso est aussi impliquée dans l'action Génome.

L'année 1996 a vu le démarrage de trois nouveaux projets européens dans l'équipe, autour desquels s'est concentrée une bonne partie de nos activités. Le projet Esprit IV OPAL (Integrated Information and Process Management In Manufacturing Engineering) a commencé en janvier, le projet Esprit IV Wire (Web Information Repository for the Enterprise) en juillet et le réseau TMR Chorochronos en août. Autre élément important de cette année, le stage post-doctoral de Cassio Souza dos Santos au sein de la société O_2 Technology. Ce stage a permis la réalisation de nouveaux outils commerciaux offrant une plus grande ouverture du SGBD O_2 à d'autres systèmes.

Notre collègue et ami, Paris C. Kanellakis, a disparu avec sa femme Maria-Teresa Otoya et leurs deux enfants, Alexandra et Stephanos, lors d'un accident d'avion, près de Cali en Colombie. Paris a très largement contribué aux activités du Projet Verso. Sa disparition nous a très profondemment touchés.

3 Actions de recherche

3.1 Bases de données multimédia

3.1.1 L'Espace et le Temps : Bases de données avec contraintes

Participants : Stéphane Grumbach, Gabriel Kuper, Zoé Lacroix, Luc Ségoufin, Michel Scholl

Les bases de données contenant des informations scientifiques, du type géométrique, temporel, géographique, etc. posent des problèmes particuliers que nous abordons dans le projet sous différents angles. Nous étudions le modèle de bases de données avec contraintes, qui connaît un intérêt grandissant dans la communauté, et que nous utilisons pour l'élaboration d'un prototype en cours de développement pour les bases de données géographiques.

Les nouvelles applications (e.g. géographiques) de bases de données conduisent à des bases de données infinies (e.g. sous-espace du plan réel) mais récursives et donc admettant une représentation effective finie. De telles bases de données peuvent être représentées à l'aide de contraintes sur des domaines numériques, par exemple les contraintes polynomiales sur les réels. Les bases de données avec contraintes généralisent les bases de données relationnelles et un certain nombre d'outils du modèle relationnel peuvent être utilisés dans ce contexte étendu. C'est le cas des langages de requêtes. Dans [392], nous

proposons des fondements théoriques au modèle de bases de données avec contraintes, et montrons certaines similarités avec la théorie des modèles finis. Nous étudions le pouvoir d'expression de la logique du premier ordre sur des bases de données contraintes [393], et introduisons de nouvelles techniques pour vérifier la non-définissabilité d'une propriété au premier ordre. L'expression de requêtes topologiques sur des données spatiales est étudiée dans [410]. L'extension de ces résultats en dimension supérieure à 2 n'est pas évidente. Des résultats d'indécidabilité ont été obtenus en dimension 4 et supérieure. En dimension 3, le cas général est complexe et peut être indécidable ; des cas particuliers peuvent être résolus.

Un prototype de bases de données géographiques basé sur le modèle avec contraintes est en cours d'implantation dans une collaboration avec le laboratoire Cedric du Cnam. Nous développons un langage de requêtes pour les données représentées à l'aide de contraintes linéaires. Nous montrons que les techniques d'optimisation de requêtes du modèle relationnel sont inefficaces en présence de contraintes et proposons des solutions préliminaires [386]. Par ailleurs, il est assez difficile de définir des fonctions agrégats pour des relations contraintes. Dans [404], nous montrons que sous certaines restrictions, ceci est possible. Dans [408], nous montrons que l'espace nécessaire au stockage des données géographiques sous forme de contraintes est du même ordre que l'espace utilisé par les techniques classiques des systèmes d'information géographique.

Nous étudions parallèlement les problèmes liés à l'implantation d'un modèle plus général permettant les contraintes polynomiales [409]. Nous montrons comment réduire la complexité de certains calculs par approximation finie.

Enfin, d'autres travaux portant plus spécifiquement sur les données temporelles ont été publiés[399].

3.1.2 Base de Documents

Participants : Serge Abiteboul, Vassilis Christophides, Sophie Cluet, Michel Scholl, Patrick Laonet

Nous nous sommes intéressés les années précédentes à la modélisation et l'interrogation des documents structurés (SGML ou autres) dans une base de données objet. Cette année, nous avons poursuivi ce travail et avons également abordé les problèmes de l'optimisation.

Nous avons retravaillé et implanté un langage de requêtes adapté aux données textuelles que nous avons appelé POQL (*Path Object Query Language*). Ce langage étend le langage OQL par l'introduction de nouveaux prédicats textuels et d'*expressions de chemin généralisées* (ECG), qui permettent d'exprimer des requêtes structurelles sans préciser toute la structure. Du point de vue de la modélisation, l'introduction dans le langage des prédicats textuels est relativement simple. Cependant, leur utilisation nécessite des techniques d'optimisation que nous avons étudiées dans le cadre plus général de l'interrogation des données virtuelles et qui seront abordées dans la section 3.2. Les expressions de chemin généralisées posent plus de problèmes. Grâce à l'utilisation de variables de chemins et d'attributs, une ECG permet l'interrogation simultanée des données et de leur structure, et ce, de façon simple et homogène. Les ECG posent des problèmes de typage et d'évaluation. Les premiers sont traités dans [385], les seconds dans [406, 385].

Nous proposons un traitement algébrique des ECG qui repose sur l'extension d'une algèbre objet. Trois nouveaux opérateurs sont introduits qui permettent la manipulation des chemins au niveau du schéma et de la base, et la transformation de chemins en données standard (c.a.d., des n-uplets). Cette approche offre une sémantique claire et opérationnelle de POQL et a été la base de l'implantation d'un interprète de requêtes que nous avons réalisé au-dessus du système O_2. L'algèbre ouvre également la voie à de nouvelles techniques d'optimisation que nous avons présentées dans [406] et qui s'appliquent aussi bien à des données non textuelles. Enfin, elle est utilisée dans nos travaux sur l'optimisation des requêtes sur les données virtuelles.

Le prototype développé est en cours de validation. Il est utilisé sur des applications réelles dans le cadre du projet européen Aquarelle. Nous avons également réalisé une première interface permettant, via le *Web*, l'interrogation de données textuelles stockées dans un système O_2.

3.2 Les données hétérogènes

3.2.1 Le Web et les Requêtes

Participants : Serge Abiteboul, Vassilis Christophides, Sophie Cluet, Jerôme Siméon, Victor Vianu

Dans le cadre du projet WIRE, nous travaillons sur l'interrogation des données dans le contexte du Web. Ce projet a une vision Internet. Le but est l'utilisation des systèmes de bases de données comme serveurs Web. Nous travaillons sur deux axes. D'une part, la conception et l'implantation d'un langage permettant une interrogation conviviale des données de la base depuis le Web. D'autre part, nous nous intéressons à la collecte des informations accessibles sur le Web et à leur interrogation depuis la base. Nous abordons donc deux aspects complémentaires.

Pour l'interrogation des données de la base depuis le Web, nous pensons étendre le langage POQL présenté dans la section 3.1.2. Avant d'étendre les fonctionnalités du langage, nous avons étudié de nouvelles techniques d'optimisation reposant sur l'utilisation des index plein-texte[405] sur la partie textuelle (éventuellement virtuelle) des données de la base. Ces techniques sont basées sur l'algèbre présentée dans [406].

Les données accessibles sur le Web le sont à partir d'interfaces très diverses : Web browsers, langages de bases de données, interfaces spécialisés, formats d'échanges de données (SGML, ASN.1, etc.). La structure de ces données est souvent implicite et irrégulière. Il est important de mieux comprendre les mécanismes de l'accès à ces données *semi-structurées* et de mieux les intégrer aux données stockées dans des SGBD. Serge Abiteboul, en sabbatique à Stanford Univ., travaille sur ces aspects dans le cadre du Projet Lore dont le but est un langage de requêtes pour données semi-structurées[414]. Il propose un état de l'art sur ce type de données dans [401]. Partant de l'évaluation de requêtes telle qu'elle est proposée dans le projet Lore, nous travaillons maintenant sur des techniques d'optimisation appropriées[415].

Enfin, Serge Abiteboul et V. Vianu (UCSD) se sont lancés dans une étude théorique des langages de requêtes pour le Web [400]. Ces travaux devraient se poursuivre lors de la visite de V. Vianu à l'INRIA en 1997.

3.2.2 Correspondances et Traduction

Participants : Serge Abiteboul, Sophie Cluet, Claude Delobel, Jerôme Siméon

Dans le cadre du projet Opal, nous nous intéressons à l'intégration de données hétérogènes et distribuées. OPAL a une vision Intranet. Son but est d'offrir un système unique dont l'interface serait le Web qui permettrait la gestion de processus de fabrication et qui manipulerait des données très hétérogènes stockées dans différents systèmes. Il s'agit donc plutôt d'une intégration virtuelle dans la mesure où les données restent dans leur système d'origine.

Dans [398], nous proposons un modèle et des outils permettant d'une part, d'établir les correspondances/liens existant entre des données hétérogènes et d'autre part, de traduire des données d'un format à l'autre. Le modèle est un modèle d'arbres très simple et qui devrait permettre de capturer n'importe quel type de format (SGML, ODMG, STEP, etc.). Les liens sont établis entre les noeuds des arbres d'une forêt et sont capturés par des règles à la *Datalog*. Une sémantique point fixe est utilisée. Nous montrons que, de la spécification de ces liens, nous pouvons dans la plupart des cas déduire des règles de traduction d'un format à l'autre. De plus, nous donnons des résultats de complexité très satisfaisants (Ptime dans la plupart des cas intéressants).

Cependant, il manque à ce modèle un niveau méta qui permettrait à l'utilisateur de spécifier les liens, non pas au niveau des données mais à un niveau plus abstrait permettant ainsi de réduire considérablement le volume d'informations manipulées par l'utilisateur. De plus, il est nécessaire d'introduire la notion de contraintes existant dans la plupart des formats que nous voulons capturer. Nous avons donc travaillé dans ces deux directions et un prototype écrit en Objective CAML est en cours de réalisation.

Ce prototype sera utilisé pour permettre l'intégration de données dans le système OPAL mais aussi pour la traduction des données OPAL en pages HTML.

Enfin, Serge Abiteboul travaille dans le projet Tsimmis dont le but est également l'intégration de données hétérogènes[411].

3.2.3 Génération de rapports et documents virtuels

Participants : Bernd Amann, Daniel Chan, Sophie Cluet, Anne-Marie Vercoustre

L'utilisation des données d'une base de données se fait traditionnellement au travers de formulaires ou d'applications spécialisées qui génèrent les données sous une forme appropriée (listes, tableaux, documents, etc). L'accès des bases de données depuis le Web requiert l'utilisation de HTML comme language de description de formulaire et de document.

Une approche classique est la génération automatique de pages HTML qui contiennent les réponses aux requêtes, avec d'éventuels liens vers d'autres éléments de la base sous forme de nouvelles requêtes. Ceci peut être réalisé de différentes façons que l'on peut classer en deux grandes familles : une approche centrée base de données et une approche centrée document.

La première approche, pour une base de données objets, consiste à attacher à une classe une *méthode* qui génère la page et les attributs à afficher pour les objects correspondants. Cette approche a l'avantage de permettre de construire potentiellement une infinité de pages en réponse à des requêtes quelconques, en utilisant un mécanisme générique et la puissance de l'héritage. L'inconvénient est qu'il est difficile de produire des documents complexes, en l'absence d'un modèle formel de document et d'éditeurs appropriés. De plus, cette approche ne permet pas à des applications différentes de définir leur propre modèle tout en partageant les mêmes données.

La seconde approche propose de partir d'un modèle du document à générer et de définir le contenu du document comme résultat de requêtes à une ou plusieurs bases de données. Cette approche permet de définir des documents plus personnalisés, dépendants du contexte et plus adaptés à la réutilisation, sans duplication, d'informations distribuées.

Nous voulons travailler principalement sur cette deuxième approche, le premier objectif étant la définition d'un language de spécification permettant de définir de tels documents. Nous appelons *document virtuel* la spécification d'un document dans ce language de spécification. Son rôle est bien de réaliser l'intégration virtuelle de données sous la forme d'un document.

Nous utiliserons certainement SGML pour définir l'organisation du document avec, sans doute, une version simplifiée pour HTML. L'étape importante est la définition d'un language unifié de requêtes pour les différents modèles de données et l'intégration de données hétérogènes dans le document. Ce travail est donc parfaitement en ligne avec ceux décrits dans les sections 3.2.1 et 3.2.2.

Dans le cas particulier de documents virtuels permettant d'assembler des fragments de documents SGML existants dans une base de données, nous pensons utiliser DSSSL, le language standard pour définir des transformations et présentations pour les documents SGML.

3.3 Extensions des bases de données à objets

3.3.1 Vues et mise à jour de Schéma

Participants : Sihem Amer-Yahia, Philippe Brèche, Claude Delobel, Zoé Lacroix, Cassio Souza dos Santos

La mise à jour de schémas s'effectue dans la plupart des bases de données objet avec des primitives dites "élémentaires" qui effectuent des opérations comme *enlever*, *ajouter*, *modifier* une propriété d'une classe ou une classe dans un schéma. Pour faciliter la conception, la maintenance et l'évolution d'un schéma, nous avons travaillé à la conception d'outils permettant d'effectuer des mises à jour correspondant à des opérations plus abstraites (généralisation de classes, destruction de classes au milieu

de hiérarchies de classes en contexte multi héritage . . .) [403], mises à jour s'effectuant au niveau du schéma mais devant être suivies de mises à jour automatiques des données.

L'année dernière, nous avons présenté un prototype implantant un mécanisme de mise à jour de schémas et de données au-dessus du système O_2. Une première approche complète du mécanisme a été présentée dans [402] et dans les rapports du projet Goodstep. Le modèle est en cours d'adaptation et tente d'être unifié avec le modèle proposé pour les vues. Nous avons également continué à travailler sur le prototype en poursuivant le développement des primitives décrites dans [403].

3.3.2 Du Relationnel à l'Objet

Participants : Sihem Amer-Yahia, Sophie Cluet, Claude Delobel

Certains utilisateurs de systèmes relationnels considèrent les systèmes à objet pour écrire de nouvelles applications. Cependant, ils aimeraient écrire ces applications sur les données qu'ils possèdent déjà et le problème de la migration relationnel/objet se pose donc. Le problème ne réside pas dans la définition de la transformation des données relationnelles en données objet mais dans la migration des données à proprement parler. Cette migration peut prendre des heures voire des jours quand les volumes de données sont très importants. Il s'agit donc de permettre un chargement incrémental, ne bloquant ni les vieilles applications relationnelles ni les nouvelles applications objet.

Parmi tous les problèmes soulevés par le chargement incrémental, nous en avons sélectionné trois sur lesquels nous avons ou allons travailler : (i) définition de la transformation relationnel/objet de façon à pouvoir garantir la cohérence des deux bases, (ii) découpage logique du chargement de façon à ne pas bloquer les bases concernées, permettre la reprise sur panne et le maintien de la cohérence, et (iii) traduction des requêtes OQL en SQL, et vice-versa, pour reporter les mises à jour.

Ce travail ne fait que commencer et nous n'avons que peu de résultats. Nous avons défini un ensemble de primitives permettant de spécifier de façon simple les transformations relationnel/objet aux niveaux du schéma et des données. Nous travaillons maintenant à définir formellement les ensembles de transformations acceptables (c.a.d., permettant le maintien de la cohérence) tout en développant un prototype.

3.3.3 OQL et SQL

Participants : Sophie Cluet, Claude Delobel, Cassio Souza dos Santos

Toujours dans le cadre de l'héritage relationnel et dans le cadre d'accord avec la Société O_2Technology, nous avons travaillé d'une part, à l'harmonisation des deux langages OQL et SQL et, d'autre part, à la réalisation d'interfaces relationnels au-dessus du système O_2.

Les systèmes de gestion de bases de données orientés-objet ont un organisme de normalisation, l'ODMG, et un langage de requêtes standard, OQL. Le monde relationnel a également son standard : SQL. Cependant, il souffre de sa trop longue histoire et de ses trop nombreuses évolutions. En collaboration avec la société O_2Technology, membre de l'ODMG, nous faisons évoluer le langage OQL de façon à minimiser les différences qui existent entre ce langage et SQL. Cette évolution est prise en compte en temps (presque) réel dans le système O_2. Parallèlement, nous continuons à aider O_2Technology dans ses efforts de norme unique, efforts entrepris conjointement avec des membres du groupe ANSI X3H7 en charge du nouvel SQL.

Dans le cadre de son stage post-doctoral industriel, Cassio Souza dos Santos a appliqué la notion de vue à l'interopérabilité des systèmes objets et relationnels, ce sujet étant actuellement d'une importance stratégique pour la société O_2Technology. Parmi les sujets traités, l'un consistait à définir des vues relationnelles de schémas objets permettant l'accès à des bases objets via SQL. Les outils relationnels étant définis au-dessus de SQL, ce travail a permis leur réutilisation dans le contexte objet. Il a également requis l'application de certaines techniques d'optimisation de requêtes développées les années précédentes au sein de l'équipe.

3.4 Fondements des Langages de Bases de Données

3.4.1 Complexité des langages de requêtes

Participants : Serge Abiteboul, Stéphane Grumbach, Zoé Lacroix, Victor Vianu

Différents langages pour les types structurés ont été étudiés dans le projet. Nous avons montré comment développer des langages de requêtes dont la complexité est raisonnable, et nous avons étudié en détail leur pouvoir d'expression. Nous proposons dans [391] une famille d'algèbres pour multi-ensembles et étudions leur complexité. En présence de certains opérateurs, nous obtenons des algèbres dont la complexité varie de l'espace logarithmique à l'hyperexponentiel. Une présentation générale des résultats des dernières années sur ce sujet est proposée dans [390].

Dans [395, 412], nous étudions la complexité et l'expressivité des modèles de SGBD actifs. Enfin, l'évaluation parallèle des programmes Datalog est étudiée dans [394].

3.4.2 Langages non-déterministes

Participants : Stéphane Grumbach, Zoé Lacroix, Victor Vianu

Les requêtes usuelles sont déterministes. Elles ont une unique réponse sur chaque instance. Il est possible de considérer des requêtes *non-déterministes* ayant plusieurs réponses possibles sur une instance. Les requêtes non-déterministes requièrent l'usage d'outils non-déterministes tant pour les exprimer que pour les calculer. Différents formalismes de requêtes non déterministes ont fait l'objet de recherches les années passées dans l'équipe. Le non-déterminisme dans les langages de règles est présenté en détail dans [397].

Nous étudions les correspondances qui existent entre les notions de non-déterminisme aux niveaux des modèles de calcul (machines de Turing) et des langages de requêtes (différentes extensions de la logique du premier ordre) [389]. En particulier, les requêtes déterministes des classes de requêtes non-déterministes étudiées sont caractérisées par des extensions du concept de définition implicite proposé et étudié par l'équipe l'année passeé [386].

4 Actions Industrielles

4.1 O_2Technology

Des liens étroits existent avec la société O_2Technology, le système O_2 nous servant de base principale d'expérimentation. Ces liens déjà anciens se sont resserrés encore cette année grâce à une participation commune à deux projets Esprit et à un stage post-doctoral réalisé par Cassio Souza dos Santos au sein de cette société.

Le travail post-doctoral a essentiellement porté sur l'utilisation de la notion de *vue* et son application dans les bases de données objet, sujet que Cassio Souza dos Santos avait étudié au sein de l'équipe Verso, dans le cadre de sa thèse de doctorat. Son travail a tourné autour de trois axes, à savoir : (i) la définition de vues relationnelles de schémas objets permettant le stockage d'objets Java et O_2, (ii) la réalisation d'une interface Java ODMG et (iii) la définition de vues relationnelles de schémas objets permettant l'accès à des bases objets par des outils relationnels.

Dans le cadre de la persistance pour le langage Java, les résultats d'une étude préliminaire réalisée pendant le stage post-doctoral[407] ont été présentés au Premier Colloque International sur Java et la Persistance qui a eu lieu au mois de Septembre 1996, en Écosse. Le transfert technologique des travaux sur les vues dans O_2 ont permis la mise sur le marché de nouveaux produits innovants dans un domaine où la concurrence est très âpre.

Le recrutement de Cassio Souza dos Santos par O_2 Technology à la fin de son stage est venu témoigner du succès de l'entreprise.

4.2 PSA

D'autres liens se créent avec les sociétés PSA, Bosch et Bull grâce au contrat Esprit IV Opal. La fin de ce contrat devrait voir la réalisation de deux applications de gestion de la fabrication de biens usinés par les sociétés PSA et Bosch à partir d'outils développés par les autres participants du projet. Verso participera à la réalisation de l'application pilote de PSA.

4.3 Euroclid

Dans le cadre du contrat Médiaculture entre l'Inria et le Ministère de la Culture, Michel Scholl était responsable d'une étude de faisabilité sur la représentation et l'accès à distance de documents multimédia du patrimoine produits par les Directions Régionales d'Action Culturelle pour la Direction du Patrimoine. La société Euroclid était sous-traitante pour la réalisation. Une maquette a été implantée par Euroclid avec le système O_2 qui permet d'interroger par le WEB des documents SGML multi-média stockés dans O2 en utilisant le traducteur SGML/O2 réalisé dans Verso.

5 Actions Nationales et Internationales

5.1 Actions nationales

S. Grumbach et F. Tahi participent, dans le cadre de l'action Génome, aux recherches entreprises dans le thème *Informatique et Génome*. F. Tahi travaille avec M. Régnier (du projet Algo) sur les problèmes de prédiction et d'énumération des structures secondaires d'ARN [413] (voir le rapport d'activité, *Action Transversale Génome et Calcul*). Cette activité devrait se terminer cette année avec la thèse de Fariza Tahi.

M. Scholl est membre du comité de pilotage du GIP Infobiogen.

A l'intérieur de l'Inria, des collaborations ont lieu avec les projets Rodin (séminaire commun), Algo (recherche dans les génomes avec M. Régnier). Des liens étroits existent aussi avec le LIPN (N. Bidoit, C. Tollu), le LRI (Claude Delobel, Emmanuel Waller), le Cedric (M. Scholl, B. Amann, P. Rigaux), et l'ENST (V. Vianu, P. Picouet). Dans le cadre de ses activités sur l'espace et le temps, l'équipe collabore activement avec le Cedric.

5.2 Actions internationales

5.2.1 Projet Esprit IV OPAL

Le projet OPAL a commencé en janvier 1996. L'objectif du projet est l'implantation d'une architecture ouverte pour l'intégration des informations nécessaires à la réalisation de processus de gestion de fabrication. L'intitulé exact du projet est : *Integrated Information and Process Management In Manufacturing Engineering*. Les participants au projet sont les sociétés AMT, Bosch, Bull, Eigner and Partner, Fatronick, O_2Technology, PSA, Tekniker et les centres de recherche FAW et l'Inria. Dans le cadre de ce projet, Verso devra essentiellement développer les outils de description, de manipulation et d'intégration de l'information hétérogène et distribuée.

5.2.2 Projet Esprit IV WIRE

Le projet Wire a commencé en juillet 1996. L'objectif du projet est le développement de répositoires pour l'entreprise et, plus précisément, l'implantation d'un ensemble d'outils permettant la construction de répositoires accessibles par les protocoles du World Wide Web. L'intitulé exact du projet est : *Web Information Repository for the Enterprise*. Les participants au projet sont les sociétés ARS, Grif SA,

O_2Technology, Zanussi Elettromeccanica Spa, FIZ Karlsruhe, et les centres de recherche Inria, Fraunhofer IGD et OSF RI. Dans le cadre de ce projet, Verso devra essentiellement développer un langage de requêtes pour l'interrogation des informations sur le Web et améliorer les techniques d'optimisation de telles requêtes.

5.2.3 Chorochronos

Le projet fait partie du réseau TMR Chorochronos démarré le 1er Aout 1996. L'objectif de ce réseau, constitué d'une dizaine de noeuds en Europe, est de coordonner et d'améliorer les efforts européens en matière de bases de données spatiales et temporelles. Le noeud de l'INRIA, en collaboration avec l'équipe Vertigo du Cedric (CNAM) s'intéresse plus particulièrement aux modèles de données, contraintes pour représenter et interroger l'information spatiale.

5.2.4 Autres Collaborations

En Amérique du Nord, des travaux en commun sont en cours avec l'Université de Toronto (A. Mendelzon), UC Santa Barbara (J. Su) [392, 409, 393], et UC San Diego (V. Vianu). Luc Ségoufin a fini son service national à UCSD et commence sa thèse dans le projet. Serge Abiteboul est en sabbatique au Département d'Informatique de l'Université de Stanford. Il travaille dans le groupe bases de données sur des problèmes d'intégration de données (projet Tsimmis entre autres). Il participe aussi à un programme de collaboration avec les Universités de Pennsylvania, Brown et l'Oregon Graduate School sur "Data Mapping and Matching : Language for Scientific Datasets" dans le cadre du CESDIS (*Center of Excellence in Space Data and Information Sciences*) de la NASA. V. Vianu collabore avec ATT. Bell Laboratories (Dan Suciu), U.C. Berkeley (Christos Papadimitriou), et la PUC. à Rio (Sergio Lifschitz). Zoé Lacroix a commencé un post-doc à l'Université de Pennsylvania.

Pour l'Asie, l'équipe est impliquée dans le *Programme de Recherches Avancées Franco-Chinois*. S. Grumbach est responsable d'une collaboration avec l'Université FUDAN à Shanghai sur le thème des bases de données à objets. Une collaboration plus étendue dans le cadre de l'Institut Franco-Chinois est envisagée.

Pour l'Europe, S. Grumbach collabore avec les Universités de Swansea (A. Dawar), et d'Athènes (F. Afrati). S. Cluet collabore étroitement avec l'Université de Manheim (G. Moerkotte) [406].

Verso participe à deux projets EC-NSF : l'un lié au projet Esprit BRA Fide2 (avec S. Cluet) et l'autre intitulé *Non-déterminisme et bases de données* (avec S. Grumbach) qui a pris fin cette année. Verso est membre du réseau d'excellence Compulog et participe au groupe bases de données de l'ERCIM. Par ailleurs, Verso participe également à un working group d'ERCIM sur la programmation avec contraintes en cours de création et participe à la création d'un nouveau groupe autour des bibliothèques numériques.

Enfin, l'équipe a une longue tradition de collaboration avec des équipes israéliennes. Un contrat de collaboration (Arc en ciel) entre le projet Verso, l'Université Hébraïque (C. Beeri) et l'Université de Tel Aviv (T. Milo) a été accepté.

6 Diffusion des résultats

6.1 Actions d'enseignement

C. Delobel est professeur à l'Université de Paris 11. M. Scholl est professeur au CNAM et est coresponsable pour le CNAM du DEA SI (Paris 6, CNAM et ENST). V. Vianu est professeur à UCSD et professeur invité à l'ENST.

B. Amann est maître de conférence au CNAM-Paris.

V. Christophides et F. Tahi sont ATER au CNAM. Z. Lacroix est ATER à l'université Paris X (1995-96).

Les cours suivants ont été assurés par divers membres de l'équipe.

SGBD relationnels, CNAM-Paris, B. Amann et V. Christophides ;
MIAGE et nouvelle formation d'ingénieurs, Paris 11, C. Delobel.

SGBD avancés, CNAM-Paris, M. Scholl et B. Amann ;
Standford University, S. Abiteboul.

SGBD à objets, DEA, Paris 6, B. Amann, V. Christophides et S. Cluet ;
Nouvelle Formation pour l'Ingénieur, IUT d'Orsay, Paris 11, S. Amer-Yahia et J. Siméon ;
ISERPA Angers et ENSEA à Cergy Pontoise, P. Brèche.

SIG, DEA BD, Paris 1,7 et 11, M. Scholl.

Théories des bases de données, DEA BD commun Paris 1, 7 et 11, S. Grumbach.

Systèmes Informatiques Cycle A, CNAM-Paris, V. Christophides.

6.2 Participation à des conférences et colloques

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

S. Abiteboul

- International Conference on Extending Data Base Technology (EDBT), Avignon, France (1996)
- International Conference on Very Large Databases (VLDB), Bombay, India (1996)
- International Conference on Management of Data (SIGMOD), Tucson, Arizona (1997).
- International Conference on Object-Oriented Information Systems, OOIS'97, Brisbane, Australia (1997)
- International Workshop on Research Issues in Data Engineering (RIDE'97), Birmingham, England (1997)
- International Workshop on Management of Semistructured Data, Tucson, Arizona (1997)

S. Cluet

- International Conference on Information and Knowledge Management (CIKM), Philadelphia, 1996.
- Conférence INFORSID, Bordeaux, France (1996)
- Seventh International Workshop on Persistent Object Systems Cape May, New Jersey, USA (1996)
- International Conference on Database Theory (ICDT), Athènes, Grèce (1997)
- International Conference on Management of Data (SIGMOD), Tucson, Arizona (1997).
- International Conference on Very Large Databases (VLDB), Athènes, Grèce (1997)
- International Conference on Data Engineering (ICD), Birmingham, Angleterre (1997)

- Sixth International Workshop on Database Programming Languages, Boulder, Colorado, USA (1997) **Président européen du comité de programme**
- Journées Bases de Données Avancées (BDA), Grenoble, 1997.

C. Delobel

- Conférence INFORSID, Bordeaux, France (1996) **Président du comité de programme**

S. Grumbach

- International workshop on Logic In Databases : LID'96, San Miniato, Italie (1996)
- Workshop on Constraint Databases and their Applications, CDB'97, Delphes, Grèce (1997)
- 5th International Symposium on Large Spatial Databases : SSD97, Berlin, Décembre (1997)

G. Kuper

- ACM Symposium on Principles of Database Systems, PODS96, Montréal, Canada (1996)
- International Conference for Database Theory, ICDT97, Delphes, Grèce (1997)
- Workshop on Constraint Databases and their Applications, CDB'97, Delphes, Grèce (1997)

M. Scholl

- SAMOS workshop on GIS for Urban Planning, Grece, Avril (1996)
- Conférence INFORSID, Bordeaux, France, June (1996)
- Workshop on Constraint Databases and their Applications, CDB'97, Delphes, Grèce January (1997)
- 5th International Symposium on Advanced Spatial Databases,
  Berlin, Juillet (1997) **Président du comité de programme**
- ACM-GIS workshop on spatial databases, Decembre (1996)
- Conférence nationale sur les Bases de Données BDA, Grenoble, Septembre (1997)
- DEXA'97, Toulouse, Septembre 1997
- COSIT'97 Conference on Spatial Information Theory, ? USA (1997)

Victor Vianu

- International workshop on Logic In Databases: LID'96, San Miniato, Italy (1996)
- International Conference on Database Theory: ICDT'97, Delpi, Grece (1997)
- ACM SIGACT-SIGART-SIGMOD Conference on Principles of Database Systems: PODS'97

6.3 Conférences invitées, tutoriels, cours, etc.

Serge Abiteboul a participé au *DIMACS workshop on Finite Model Theory* (Princeton) où il a présenté les travaux de P. Kanellakis sur les bases de données objet. Il a également participé aux *ARPA workshop on data exchange* (Washington) et *CESDIS workshop on data integration* (Washington) où il a présenté un langage de requêtes pour données semi-structurées.

S. Cluet organise avec Nicole Bidoit une école d'été jeunes chercheurs en bases de données qui se tiendra en Mars 1997 à Port Camargue. Elle est co-responsable du *Sixth International Workshop on Database Programming Languages* qui se tiendra dans le Colorado en août 1997. Elle est responsable du programme de démonstrations de la prochaine conférence ICDE qui aura lieu a Birmingham en Avril 1997.

S. Cluet et S. Grumbach sont dans le comité de lecture d'un numéro spécial de TSI, Technique et Science Informatiques, sur les bases de données à paraître en 1998.

Stéphane Grumbach a été invité à faire une conférence intitulée "Towards practical constraint databases" pour la Journée du Fond National de la Recherche Scientifique belge, à Bruxelles en mai 1996, et un exposé sur la théorie des modèles finis et les bases de données dans le cadre du tutorial sur "Finite Model Theory, Problems, Methods and Applications" à University of Wales, Swansea, en juillet 1996.

V. Vianu a donné deux présentations invitées, au Logic Colloquium, Mathematics Department, U.C. Los Angeles, et au DIMACS Workshop on Finite Model Theory, Princeton, New Jersey (1996).

Les membres de l'équipe ont été invités dans de nombreux centres de recherche et universités. Sophie Cluet a été invité par l'université de Konstanz et de Tel Aviv. Stéphane Grumbach par les universités de Swansea, Bruxelles (ULB), et le laboratoire IASI du CNR à Rome.

6.4 Animations Scientifiques

Le prototype O_2 Views a été présenté au CEBIT 96 et la conférence WWW'96 à Paris.

7 Publications

Thèses

[385] V. CHRISTOPHIDES, *Documents Structurés et Bases de Données Objet*, thèse de doctorat, CNAM-Paris, Octobre 1996.

[386] Z. LACROIX, *Bases de Données: des Relations Implicites aux Relations Contraintes.*, thèse de doctorat, Université Paris-Sud, Juin 1996.

Articles et chapitres de livre

[387] S. ABITEBOUL, G. M. KUPER, H. G. MAIRSON, A. A. SHVARTSMAN, M. Y. VARDI, «In Memoriam: Paris C. Kanellakis», *ACM Computing Surveys*, 1996.

[388] S. ABITEBOUL, M. VARDI, V. VIANU, «Fixpoint Logics, Relational Machines, and Computational Complexity», *Journal of the Association for Computing Machinary*, à paraître.

[389] S. GRUMBACH, Z. LACROIX, «Non-deterministic Machines and Languages», *Annals of Mathematics and Artificial Intelligence*, papier invité, à paraître.

[390] S. GRUMBACH, L. LIBKIN, T. MILO, L. WONG, «Query Languages for Bags: Expressive Power and Complexity», *Sigact News 27*, 2, june 1996, p. 30–37.

[391] S. GRUMBACH, T. MILO, «Towards Tractable Algebras for Bags», *Journal of Computer and System Sciences 52*, 3, june 1996, p. 570–588.

[392] S. GRUMBACH, J. SU, «Finitely representable Databases», *Journal of Computer and System Sciences*, papier invité.

[393] S. GRUMBACH, J. SU, «Queries with Arithmetical Constraints», *Theoretical Computer Science 173*, 1997, Invited to a special issue.

[394] S. LIFSCHITZ, V. VIANU, «A Probabilistic Approach to Datalog Parallelization», *Theoretical Computer Science*, papier invité, à paraître.

[395] P. PICOUET, V. VIANU, «Semantics and Expressiveness Issues in Active Databases», *Journal of Computer and System Sciences*, papier invité, à paraître.

[396] V. VIANU, «Databases and Finite-Model Theory», *DIMACS Series of the American Mathematical Society*, papier invité, à paraître.

[397] V. VIANU, «Rule-Based Languages», *Annals of Mathematics and Artificial Intelligence*, papier invité, à paraître.

Communications à des congrès, colloques, etc.

[398] S. ABITEBOUL, S. CLUET, T. MILO, «Correspondence and Translation for Heterogeneous Data», *in : Proceedings of the International Conference on Database Theory*, Greece, 1997.

[399] S. ABITEBOUL, L. HERR, J. V. DEN BUSSCHE, «Temporal versus First-Order Logic to Query Temporal Databases», *in : Proceedings of the ACM SIGACT-SIGMOD-SIGART Symposium on Principles of Database Systems*, 1996.

[400] S. ABITEBOUL, V. VIANU, «Queries and Computation on the Web», *in : Int'l. Conf. on Database Theory*, Delpi, Greece, 1997. à paraître.

[401] S. ABITEBOUL, «Querying semi-structured data», *in : Proceedings of the International Conference on Database Theory*, Greece, 1997. (invited paper).

[402] S. AMER-YAHIA, P. BRÈCHE, C. S. DOS SANTOS, «Object Views and Updates», *in : Actes des 12 ièmes Journées Bases de Données Avancées*, Cassis - France, août 1996.

[403] P. BRÈCHE, «Advanced Primitives for Changing Schemas of Object Databases», *in : Proc. of the CAiSE'96 conference*, Springer Verlag, Heraklion, Crète, mai 1996.

[404] J. CHOMICKI, D. GOLDIN, G. KUPER, «Variable Independence and Aggregation Closure», *in : Proc. Fifteenth Symposium on Principles of Database Systems*, p. 40–48, Montréal, Canada, 1996.

[405] V. CHRISTOPHIDES, S. CLUET, G. MOERKOTTE, J. SIMÉON, «Un Cocktail de Bases de Données et d'Index Plein Texte», *in : XIV^{ème} Congrès INFORSID*, Bordeaux, 1996. Journée O_2 Technology.

[406] V. CHRISTOPHIDES, S. CLUET, G. MOERKOTTE, «Evaluating Queries with Generalized Path Expressions», *in : SIGMOD'96*, p. 413–422, Montréal, Québec, Canada, June 1996.

[407] C. S. DOS SANTOS, E. THEROUDE, «Persistent Java», *in : Actes du Premier Colloque International sur la Persistance et Java*, Drymen - Ecosse, septembre 1996.

[408] A. FRANK, P. HAUNOLD, W. KUHN, G. M. KUPER, «Representation of Geometric Objects as Set of Inequalities», *in : 12th European Workshop on Computational Geometry (CG'96)*, Mũnster, Germany, March 1996.

[409] S. GRUMBACH, J. SU, «Towards Practical Constraint Databases», *in : 15th ACM Symp. on Principles of Database Systems*, p. 28–39, Montréal, June 1996. Invited to a special issue of the Journal of Computer and System Sciences.

[410] C. PAPADIMITRIOU, D. SUCIU, V. VIANU, «Topological Queries in Spatial Databases», *in : 14th ACM Symp. on Principles of Database Systems*, Montréal, Canada, 1996. Full paper invited to special issue of JCSS.

[411] Y. PAPAKONSTANTINOU, S. ABITEBOUL, H. GARCIA-MOLINA, «Object Fusion in Mediator Systems», *in : Proceedings of Internat. Conf. on Very Large Data Bases*, Bombay, India, 1996.

[412] P. PICOUET, V. VIANU, «Expresiveness and Complexity of Active Databases», *in : Int'l. Conf. on Database Theory*, Delpi, Greece, 1997. à paraître.

[413] M. REGNIER, F. TAHI, «Enumeration and Asymptotics in Computational Biology», *in : Workshop "Mathematical Analysis of biological Sequences"*, Trondheim, Norvège, août 1996.

Divers

[414] S. ABITEBOUL, D. QUASS, J. MCHUGH, J. WIDOM, J. WIENER, «The Lorel Query Language for Semistructured Data», 1996, demo at ACM SIGMOD Intern. Conf. on Data Management.

[415] S. CLUET, G. MOERKOTTE, «Query Processing in the Schemaless and Semistructured Contex», 1996, en préparation.

[416] M. SCHOLL, A. VOISARD, J. PELOUX, L. RAYNAL, P. RIGAUX, «SGBD Géographiques, Spécificités», 1996, soumis.

8 Abstract

The objective of the Verso group is to develop novel solutions to problems of database management. We study both practical and theoretical aspects. This year, we mainly focused on the problems generated by the mass of heterogeneous multimedia information available on the *Web*. Among the many potential topics offered by this new context, we selected two. We work on developping softwares for (i) new multimedia applications and (ii) geographicaly distributed heterogeneous data applications. We also continue our work on (iii) the improvement of object-oriented database systems and (iv) the fundamental aspects of models and languages.

Our work on multimedia data concerns space-time and textual data. For the former kind, we are developping a system based on previous theoretical work on the constraint model. Concerning textual data, we extended the O_2 database system so as to be able to load/unload/query SGML documents.

In order to integrate heterogeneous data, one has to define the correspondance/mapping between data under different formats. We are developping a software allowing to do this in a declarative manner. The next step is the development of a graphical interface. Our other topics of research in this area concerns the interrogation of the data available on the *Web* and the interrogation of the database from the *Web*. For this we study new models, languages and optimization techniques.

Our work on object-oriented database systems concerns the development of new tools for sophisticated schema updates, the integration/migration of relational legacy data and, in collaboration with O_2Technology, the reuse of existing relational tools, the improvement and implementation of the OQL standard query language.

The theoretical part of the research done at Verso is concerned with understanding the fundamental aspects of database models and their links with the development of new tools. An important direction of research is the study of the complexity and expressiveness of database languages, taking into account aspects such as new data types and non-determinism.

Keywords: database model, database theory, query optimization, object-oriented database, multimedia base, geographic database, deductive database, electronic documents, virtual documents, World Wide Web, Genome.

Some scientific cooperations: participation in Esprit IV OPAL and Wire projects, in TMR network Chorochronos, in the Ercim group for databases; participation in two joint EC-NSF contracts; involvement in Franco-Chinese advanced research programs, collaboration with the Hebrew and Tel Aviv Universities.