NOTICE INDIVIDUELLE ET CURRICULUM VITÆ

NOM : BENZAKEN
PRÉNOM : VERONIQUE
DATE ET LIEU DE NAISSANCE : 16 juin 1960 à Tarascon (13)
NATIONALITÉ : Française
SITUATION DE FAMILLE : Célibataire, une fille (1999)
PROFESSION : Professeur
ADRESSE : LRI, Université de Paris-Sud, bat 490, 91405 Orsay Cedex
TÉLÉPHONE : (bureau) 01 6915 6628, (domicile) 01 4284 3398
ADRESSE ÉLECTRONIQUE : Veronique.Benzaken@lri.fr

1 Encadrement

Le pourcentage de participation à l'encadrement est indiqué entre parenthèses carrées.

Thèses en cours

1. Kim N'Guyen "Algèbres de filtrage pour l'interrogation de documents XML", [50%]
2. Andrei Arion, "Optimisation à base d'index et mises à jour dans un système de gestion de données XML",[50%].
3. Cédric Miachon. *Requêtes dans* ℂDuce *: implantation, optimisation et mesures de performances*, (depuis septembre 2003).
4. Marwan Burelle. *Sécurité de transformations XML* (depuis septembre 2002) Co-encadrement avec G. Castagna, LIENS, ENS [50%].

Thèses soutenues

1. S. Praud à partir d'Octobre 1997. Co-encadrement avec S. Cerrito (LRI)
   Sujet : "Contraintes temporelles, analyse statique et déduction automatique". Thèse soutenue en Février 2002, Université de Paris XI. [50%], Actuellement en poste chez Thalès Recherche.

DEA

1. Elise Henaux, "Validation sémantique de données au format XML pour la norme CGNS", Année 2004-2005
2. Kim N'Guyen, "Algèbres de filtrage pour l'interrogation de documents XML", Année 2003-2004 [50%]
3. Andrei Arion, "Optimisation à base d'index et mises à jour dans un système de gestion de données XML", Année 2003-2004 [50%].
4. Cédric Miachon. "Requêtes dans ℂDuce : implantation, optimisation et mesures de performances", Année 2002-2003.
5. Marwan Burelle. "Sécurité de transformations de documents XML" DEA I_3, Année 2001-2002 [50%]
6. Alain Frisch. "Types récursifs, combinaisons booléennes et fonctions surchargées: application au typage de XML" DEA Programmation. Année 2000-2001. [50%]

2 Recherche : Langages de programmations et systèmes pour données semi-structurées

L'explosion récente d'Internet comme support de calculs globaux conduit à l'échange croissant d'informations. La nature hautement distribuée du réseau induit que les informations sont généralement collectées, traitées puis restructurées à partir de sources multiples et hétérogènes. Afin d'en permettre le traitement ces informations sont publiées sous un même format. XML devient le standard de-facto pour l'échange et la manipulation de documents sur le Web. De nombreux facteurs plaident en faveur de son acceptation : le fait que les documents XML soient lisibles par un humain ainsi que leur nature auto-descriptive ; le processus actif de standardisation du W3C et le fait que de nombreuses entreprises, parmi les plus représentatives dans le domaine des technologies de l'information, l'aient dores et déjà adopté. Dans un tel contexte nous nous intéressons aux points suivants :

1. Définition de langages de programmation pour XML typés et expressifs
2. Systèmes de stockage natifs et optimisation de requêtes
3. Sécurité orient´e langage.

2.1 Définition de langages de programmation pour XML typés et expressifs

Comme le reflète le processus actuel de standardisation, les documents XML sont intrinsèquement typés (notions de validité et de bonne formation relativement à une DTD, XML-Schema, Relax-NG). Or, les langages de transformation associés (Xpath/XSLT) sont essentiellement non-typés.

La définition de langages de transformations pour XML typés permettra d'obtenir une plus grande fiabilité du code et surtout d'offrir un contrôle (partiel) statique de correction de la transformation. Ainsi, dans un tel contexte, il sera possible de garantir que le résultat d'une transformation donnée est bien conforme au type (ou DTD, ou Schemas) attendu.

Enfin, dans le contexte de l'interrogation stricto-sensu de documents XML, il est indispensable d'offrir une interface plus déclarative qu'un langage turing-complet. Ce type d'interface est communément appelé langage de requête.

Notre contribution des quatre dernières années à cet axe est briévement développée.

Le Langage ℂDuce (`www.cduce.org`)

Les langages de transformation de documents XML actuels sont loin d'être satisfaisants. Dans le cas de DOM (Document Object Model) —un API pour XML très répandu— par exemple, il n'y a aucune manière de garantir statiquement que le document produit par un programme est du type attendu. Ceci est aussi vrai pour XSLT par exemple, à l'exception près que dans ce dernier cas des programmes non corrects n'engendrent pas du XML mal formé mais des documents vides. D'une certaine manière on peut considérer ceux-ci comme des langages non typés.

Proposer un langage typé pour manipuler des documents XML est assez naturel : XML n'est qu'une syntaxe, et différents sytèmes de types sont proposés : DTD, XML-Schema, ... ; un document ne prend de sens que si on le voit avec un certain type XML.

Dans le cadre du stage de DEA de A. Frisch que j'ai co-encadré avec G. Castagna nous avons étudié l'introduction du sous-typage sémantique ceci nous a permis d'établir le socle théorique sur lequel bâtir les développements futurs. Les résultats de ce travail sont présentés dans [7] et donnent lieu au développement du langage ℂDuce [6, 5] qui est disponible sur le site `www.cduce.org` et qui compte à ce jour une centaine d'utilisateurs.

Nous poursuivons ces recherches dans deux directions parallèles : (i) langages de requêtes, stockage et optimisation et (ii) sécurité des transformations de documents XML.

ℂQL un langage de requête pour XML:

Le système de types de XDuce sert de base à celui de XQuery, un langage de requêtes pour XML proposé par le W3C. L'adoption d'un tel système montre l'intérêt que porte le monde des bases de données à celui des langages typés adaptés à

XML. En effet, les types peuvent apporter plusieurs caractéristiques au monde des bases de données (comme ils l'ont toujours fait de part le passé).

Dans le cadre de la manipulation de documents XML, on peut considèrer le mécanisme de filtrage des langages fonctionnels de la famille d'OCaml ("pattern-matching") comme une primitive pour exprimer et exécuter une *requête*. Dans le contexte du stage de Cédric Miachon [8], puis de sa thèse en cours [3], nous avons suivi cette approche en intégrant dans ℂDuce un langage de requête, ℂQL, exploitant pleinement la puissance expressive des motifs ("patterns") ainsi que leur compilation efficace au moyen d'automates d'arbres non-uniformes. Il s'agit d'un langage déclaratif qui laisse toute latitude au compilateur pour exploiter au mieux des réecritures algébriques (en particulier de motifs) afin de générer des plans d'exécutions, en memoire centrale efficaces.

Les perspectives de ce travail sont diverses: (i) étendre l'opération de filtrage afin de permettre également des parcours en profondeur de l'arbre (à la XPath), (ii) définir un langages de requêtes graphique [10] utilisant comme briques élémentaires une bibliothèque de requêtes ℂQL et de fonctions ℂDuce (iii) offrir un mécanisme de persistence à ℂQL ainsi que des mécanismes d'optimisation physique de requête associés/e

2.2 Systèmes de stockage natifs et optimisation de requêtes

Les résultats existants en matière d'optimisation de requêtes sur des données (XML) persistantes peuvent être classés en fonction du modèle sous-jacent de stockage et de traitement de requêtes. Des travaux se sont appuyés sur des systèmes de gestion de bases de données relationnelles pour stocker des documents XML, et ont traité des requêtes XML en les traduisant en SQL. En conséquence, tous les aspects reliés au placement des données, au stockage, et à l'optimisation de requêtes ont été délégués au système relationnel. Cette approche a ses limites, puisque des langages de requêtes XML plus récents (tels que XQuery et CDuce sont bien plus complexes, et très différents de SQL.

Plus généralement, en collaboration avec Ioana Manolescu (INRIA - Futurs - Gemo), avec qui nous co-encadrons la thèse d'A. Arion,nous proposons un système de stockage natif ainsi qu'un optimiseur générique [1, 2, 9] pouvant être paramétré en fonction des besoins et caratéristiques des applications visées.

2.3 Sécurité de transformation de documents XML:

Bien qu'XML soit en passe de devenir le standard de-facto d'échanges et de manipulation d'informations au sein du *middleware*, peu d'attention a été portée sur la nature même de tels échanges. En particulier, les aspects de sécurité liés à la caractérisation et l'analyse des flots d'informations ne sont pas pris en compte dans ce cadre.

En matière de *sécurité (orientée langage)* par exemple, l'exploitation tant statique que dynamique d'un *système de types* sous-jacent constitue un premier outil pour analyser les flots d'information masqués par une transformation.

L'objectif de cet axe de recherche, mené dans le cadre de l'encadrement de la thèse de M. Burelle, est double : (i) offrir des mécanismes d'analyse de flots d'informations dans le contexte de langages de transformations de documents XML et (ii) permettre la définition de listes de contrôle d'accès (ACL) aux bases de documents XML. Une première étape consiste à exploiter le système de type de ℂDuce afin de définir de telles analyses basées sur la notion de propagation de labels [4].. Dans ce contexte, la difficulté et la nouveauté reposent sur le fait que les calculs *dépendent* des types : il n'est plus possible de recourir à des analyses classiques pour lesquelles une telle dépendance est absente. En outre, l'absence de flots d'information est généralement exprimée par la propriété de *non-interference*. A nouveau, dans le contexte de calculs dépendants des types, cette notion doit être adaptée afin de prendre en compte cette connaissance statique sinon les analyses obtenues conduisent à des résultats triviaux.

3 Animation Scientifique, Évaluation et administration de la recherche

Jurys et rapports de thèses

1. Membre du jury de la thèse de L. Pilard, soutenue le 15 décembre 2005 à l'Université de Paris 11.
2. **Rapporteur** de la thèse de D. Debarbieux, soutenue le 9 décembre 2005 à l'Université de Lille.

3. **Rapporteur** de la thèse de S. Medjdoub, soutenue le 8 décembre 2005 à l'Université de Versailles Saint Quentin.
4. Membre du jury de la thèse de N. Gesbert soutenue le 2 décembre 2005 à l'Université de Marne la Vallée.
5. **Rapporteur** de la thèse de D. Duarte, soutenue en juillet 2005 à l'Université de Francois Rabelais Blois.
6. Membre du Jury de la thèse d'O. Benjelloun, soutenue en septembre 3004 à l'Université de Paris 11.
7. Membre du Jury de la thèse de P. Letouzey, soutenue le 10 Juillet 2004 à l'Université de Paris 11.
8. Membre du Jury de la thèse de J. Forest, soutenue le 29 Septembre 2003 à l'Université de Paris 11.
9. Membre du Jury de la thèse de P-A. Veltri, soutenue le 1 Octobre 2002 à l'Université de Paris 11.
10. Membre du Jury de la thèse de J. Barbay, soutenue le 24 Septembre 2002 à l'Université de Paris 11.
11. **Rapporteur** de la thèse de L. Nemirovski, soutenue à l'Université de Paris 1 Janvier 2002.

Comités de rédaction

1. Membre du comité de rédaction de la revue I_3, depuis 1999.

Comités de programme

1. Membre du comité de programme de l' International Workhop on Programming Language technologies for XML (ACM-PLAN-X 2005) in conjunction with ACM POPL'2005.
2. Membre du comité de programme de l' International Symposium on Methodologies for Intelligent Systems 2005
3. **Présidente du comité de programme** de la conférence Bases de Données Avancées (BDA) 2005.
4. Membre du comité de programme du workshop ACM on XQuery Implementation and Experieces, Xime-P joint with ACM SIGMOD-PODS, Paris Juin 2004.
5. Membre du comité de programme de la conférence Bases de Données Avancées (BDA) 2003.
6. Membre du comité de programme de la conférence Bases de Données Avancées (BDA) 2002.

Autres

- Lecteur pour les revues Journal of Functional Programming, Information Systems, Theoretical Computer Science, Theory and Practice of Object-Oriented Systems et pour les conférences DBPL, PODS, VLDB, EDBT, ICDT, ICALP, ESOP, ECOOP.

Animation et évaluation de la recherche

1. De Septembre 2002 à 2004, membre du comité de pilotage du RTP Bases de données et d'informations hétérogènes et distribuées (RTP 9).
2. Janvier 2003, rapporteur extérieur pour l'évaluation d'une proposition RNTL.
3. Membre titulaire de la **Commission de Spécialistes** Informatique de l'Université de Paris 11 depuis décembre 2003.
4. Membre du **Conseil du laboratoire** du LRI, UMR 8623 CNRS depuis 1999.
5. Membre du **Conseil Scientifique** de l'Université de Paris 11, de mars 2002 à mars 2005.
6. Mai 2003, membre du jury du concours CR2 pour l'Inria de Rennes.
7. Mai 2004, membre du jury du concours CR2 pour l'Inria de Rennes.
8. Mai 2004, membre du jury du concours CR2 pour l'Inria de Futurs.
9. Mai 2005, membre du jury du concours CR2 pour l'Inria de Rennes.

Contrats et collaborations

2003: Porteur d'un projet d'action spécifique "Langages, types, sécurité et intégrité pour données semi-structurées", dans le cadre du RTP 9 acceptée en mai 2003 (`http://www.lri.fr/~benzaken/ltsi.html`).

2003: Responsable pour l'équipe Bases de donnée d'un proposition soumise à l'appel d'offre RNTL : "GraphDuce", labellisée le 13 mars 2003.

2003: Responsable pour le partenaire LRI, du projet CASC, ACI-sécurité du CNRS-MENRT-INRIA.

2004: Responsable pour le partenaire LRI, du projet TRALALA, ACI-masse de données du CNRS-MENRT-INRIA.

4 Publications

Communications dans des conférences internationales avec comité de lecture

[1] Andréï Arion, Véronique Benzaken, and Ioana Manolescu. XML access modules: Towards physical data independence in XML databases. In *Proceedings f the Second International Workshop on XQuery Implementation, Experience and Perspective colocated with ACM SIGMOD-PODS 2005*, 2005.

[2] Andréï Arion, Véronique Benzaken, Ioana Manolescu, and Ravi Vijay. Uload: Choosing right storage for your XML application. In *Demonstration at the 31st International Conference on Very Large Databases*, 2005.

[3] Véronique Benzaken, Giuseppe Castagna, and Cédric Miachon. A full pattern-based paradigm for XML query processing. In *PADL, International Symposium on Practical Aspects of Declarative Languages*, pages 235–252, 2005.

[4] V. Benzaken, M. Burelle, and G. Castagna. Information flow security for XML transformations. In V. Saraswat, editor, *Eight Asian Computing Science Conference (ASIAN'03)*, Lecture Notes in Computer Science, Mumbai, India, 10-13 December 2003. Springer-Verlag.

[5] V. Benzaken, G. Castagna, and A. Frisch. ℂDuce: an XML-centric general-purpose language. In *ACM International Conference on Functional Programming (ICFP'03)*, 2003.

[6] V. Benzaken, G. Castagna, and A. Frisch. Cduce: a white paper. In B. C. Pierce and P. Wadler, editors, *International workshop on Programming Languages for XML (PLANX'02*, Pittsburgh, USA, 2002.

[7] A. Frisch, G. Castagna, and V. Benzaken. Semantic subtyping. In *Seventeenth Annual IEEE Symposium on Logic in Computer Science (LICS'2002)*, pages 137–146, Copenhagen, Denmark, 2002.

Communications dans des conférences nationales avec comité de lecture

[8] V. Benzaken, G. Castagna, and C. Miachon. CQL a pattern-based query language for XML. In *In Proceedings 20th Bases de Données Avancées (BDA'04)*, 2004.

Articles soumis ou en préparation

[9] A. Arion, V. Benzaken, I. Manolescu, and Y. Papakonstantinu. Algebra-based identification of tree patterns in XQuery. Technical report, LRI, 2006.

[10] V. Benzaken, D. Colazzo, and C. Miachon. Pattern by example: a graphical tool for querying XML. Technical report, LRI, 2006.

[11] V. Benzaken, D. Colazzo, and K. NGuyen. Type projectors. Technical report, LRI, 2006.