Intégration de données hétérogènes basée sur la qualité

Dimitre Kostadinov – Verónika Peralta – Assia Soukane – Xiaohui Xue

Laboratoire PRiSM, Université de Versailles
45 avenue des Etats-Unis
78035 Versailles Cedex
France
{prenom.nom}@prism.uvsq.fr

RÉSUMÉ. *Les systèmes de médiation constituent une réponse architecturale pour un accès transparent à des sources de données distribuées. Cependant, leur mise en oeuvre pose un certain nombre de problèmes, tant en ce qui concerne la génération des liens sémantiques entre le schéma de médiation et les sources de données (requêtes de médiation) qu'en ce qui concerne l'adaptation de l'accès aux besoins des utilisateurs ou la mesure de la qualité des données obtenues. Ces problèmes sont d'autant plus cruciaux lorsque les sources sont nombreuses et hétérogènes. Nous proposons un atelier de conception qui permet de générer automatiquement les requêtes de médiation dans un contexte relationnel et XML et d'adapter ces requêtes aux besoins des utilisateurs en termes de qualité.*

ABSTRACT. *Mediation systems constitute an architectural solution for transparent access to distributed heterogeneous sources. However, their implementation poses several problems, especially the generation of semantic links (mediation queries) between the mediation schema and its related data sources, the adaptation of the computed data with respect to user's needs as well as the evaluation of the quality of this data. Furthermore, the problem is particularly important when there is a high number of distributed and heterogeneous data sources. We propose a set of CASE tools to support the automatic generation of mediation queries, both in a relational and in a XML context, and to adapt the queries to user preferences and quality needs.*

MOTS-CLÉS : *Systèmes de médiation, requêtes de médiation, adaptabilité, qualité des données, hétérogénéité des données.*

KEYWORDS: *Mediation systems, mediation queries, user adaptability, data quality, data heterogeneity*

CATEGORIE : *Jeune chercheur*

1. Introduction

De nos jours, les systèmes de médiation sont de plus en plus développés et connus. Leurs composants essentiels sont : le schéma global, les mappings du schéma global avec les sources, les fonctions de réécriture de requêtes et les fonctions de composition des résultats. Tous ces composants prennent en compte l'hétérogénéité qui est un des principaux problèmes pour lesquels les systèmes de médiation sont construits. D'autres problèmes de conception émergent lors de l'utilisation de ces médiateurs. Parmi ces problèmes, on distingue la définition du schéma global et la définition des mappings qui relient ce schéma global aux sources de données. En raison d'un grand nombre de sources de données, contenant éventuellement des informations redondantes et de qualité variée, il est également important d'adresser le problème d'adaptabilité du système de médiation aux besoins des utilisateurs, notamment en terme de qualité des données.

Les principales questions que l'on se pose sont : (1) Comment automatiser la génération des requêtes de médiation ? (2) Comment détecter et résoudre les problèmes liés à l'hétérogénéité ? (3) Comment évaluer la qualité du système de médiation ? (4) Comment donner la possibilité aux utilisateurs d'exprimer leurs préférences ? (5) Comment tenir compte des préférences de l'utilisateur dans la conception du système de médiation? Ces problèmes ne sont pas spécifiques aux médiateurs, on les retrouve également dans la conception des entrepôts de données et aussi dans les architectures peer-to-peer.

Dans ce contexte, nous proposons un système adaptatif d'aide à l'intégration de données hétérogènes, basé sur la qualité. Il a pour objectif d'une part de générer automatiquement les requêtes de médiation en tenant compte de l'hétérogénéité des données et d'autre part d'évaluer la qualité de ces requêtes pour la confronter aux besoins et aux préférences des utilisateurs. Cet article décrit un ensemble d'outils constituant un atelier d'aide à la conception d'applications sur des architectures de médiation. Plus précisément, nous décrivons la génération des requêtes de médiation dans un contexte relationnel et XML, l'évaluation de la qualité des données retournées aux utilisateurs et la gestion de profils utilisateur qui contiennent leurs préférences.

Le reste du document est organisé de la façon suivante : La section 2 décrit notre approche globale d'aide à l'intégration de données de multiples sources. Ensuite, la section 3 présente la génération de requêtes de médiation, et la section 4 présente l'adaptabilité des requêtes aux besoins des utilisateurs. Finalement, les sections 5 et 6 présentent les travaux connexes à notre approche et les conclusions de notre réalisation.

2. Approche globale

Un système de médiation est défini comme l'intégration de plusieurs sources de données distribuées et hétérogènes. Cette intégration s'exprime principalement à l'aide d'un schéma global, appelé schéma de médiation, et d'un ensemble de mappings associant les sources de données au schéma de médiation (appelées requêtes de médiation). La Figure 1 donne un aperçu d'une telle architecture.

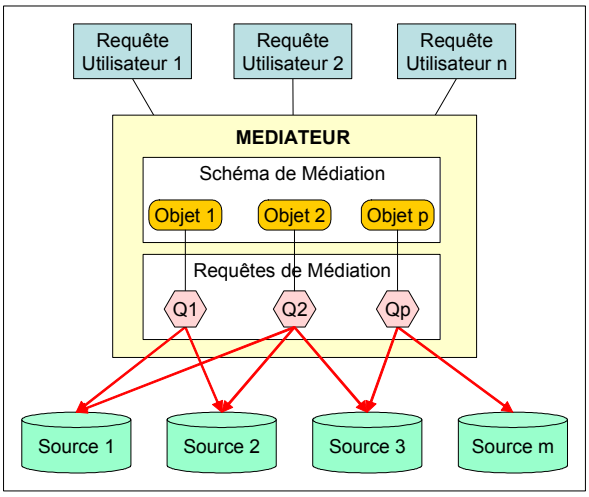

Figure 1. *Architecture d'un système de médiation*

Lors de l'utilisation d'un système de médiation, la génération des requêtes de médiation est l'une des tâches les plus fastidieuses à faire manuellement, surtout lorsque le nombre de sources devient important.

Sachant qu'il peut exister plusieurs sources fournissant les mêmes types de données, plusieurs requêtes de médiation peuvent définir un même objet de médiation. Par exemple, la recherche de films pour enfants peut être faite dans les sources Disney, FNAC ou Virgin ou une combinaison de ces sources. Le problème est de savoir quelle source ou quelle combinaison de sources offre les résultats les plus satisfaisants à l'utilisateur, en fonction de ses préférences de thématique ou de qualité.

Notre approche permet de générer un ensemble de requêtes de médiation, d'évaluer leur qualité et de sélectionner les plus pertinentes en fonction des préférences de chaque utilisateur.

2.1. Exemple illustratif

Prenons comme exemple un système de médiation qui fournit des informations sur des livres scientifiques et leurs auteurs. Supposons que le schéma de médiation de ce système est composé d'une seule relation :

Publications (**titre**, **auteur**, affiliation, conference, annee, editeur)

Prenons aussi quatre sources des données AutorBase, ConfList, DBpubs et LP avec leurs schémas respectifs. AutorBase contient des informations sur des auteurs de publications scientifiques, ConfList représente un catalogue de conférences et DBpubs et LP permettent d'obtenir des listes d'articles publiés :

- Source AutorBase : Auteurs (**auteur**, adresse, email, affiliation, nationalite)
- Source ConfList : Conferences (**conference**, annee, ville, pays, editeur)
- Source DBpubs : Publications (**auteur, titre**, conference)
- Source LP : Pubs (**auteur, titre**, conference, annee, editeur)

Pour calculer la relation de médiation, il faut combiner les sources de données de sorte à obtenir l'ensemble des attributs de la relation de médiation. Sur notre exemple, la solution n'est pas unique. Les requêtes Q_1, Q_2, Q_3 (et toute combinaison entre elles : $Q_1 \cup Q_2$, $Q_1 \cap Q_3$, etc.), permettent chacune de définir la relation de médiation :

- Q_1 = LP.Publications $\bowtie_{auteur}$ AutorBase.Auteurs
- Q_2 = DBpubs.Pubs $\bowtie_{auteur}$ AutorBase.Auteurs $\bowtie_{conference}$ ConfList.Conferences
- Q_3 = DBpubs.Publications $\bowtie_{auteur}$ AutorBase.Auteurs $\bowtie_{auteur,\ titre}$ LP.Pubs

Bien que chacune de ces requêtes permet d'obtenir l'ensemble des attributs de la relation de médiation, elles peuvent produire des résultats dont la sémantique et la qualité différent en fonction des sources accédées. Par exemple, DBpubs peut contenir les articles publiés dans des conférences sur les Bases de Données et LP les publications en Intelligence Artificielle. Par conséquent la requête Q_1 va fournir des résultats plus pertinents que Q_2 pour un chercheur dans le domaine des Bases de Données. Un deuxième point de comparaison entre les requêtes de médiation est la qualité des résultats produits par chacune d'entre elles. Si la source DBpubs n'est pas mise à jour fréquemment, elle va fournir des données moins fraîches que LP et sera moins intéressante pour un chercheur expert qui ne s'intéresse qu'aux nouvelles publications.

Notre approche consiste à générer plusieurs requêtes de médiations, à évaluer la qualité des données produites par chacune d'elles et à confronter cette qualité aux préférences des utilisateurs.

2.2. Outils proposés

La figure 2 illustre l'ensemble des outils constituant l'atelier de conception que nous proposons. Trois types d'outils sont particulièrement décrits: (1) génération automatique de requêtes de médiation, (2) expression des préférences de l'utilisateur et (3) évaluation de la qualité des données.

Ces outils communiquent par le biais d'un serveur de méta-données qui gère l'ensemble des connaissances décrivant les sources de données, le médiateur, les mappings, la qualité et les profils utilisateur. Les sections suivantes donnent un aperçu des techniques sous-jacentes à ces outils.

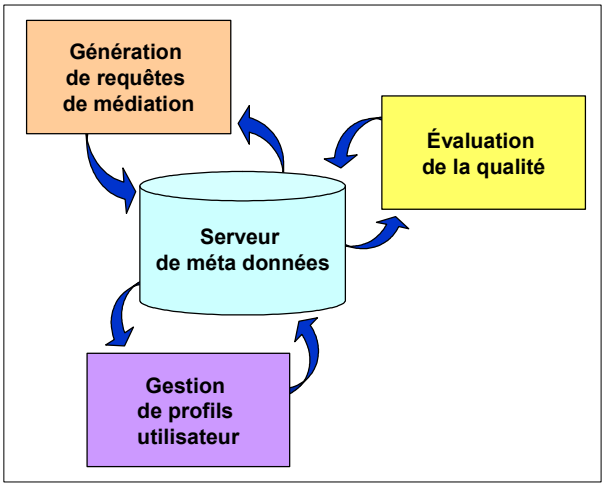

Figure 2. *Vision générale des outils*

3. Génération des requêtes de médiation

L'écriture manuelle des requêtes de médiation donne sans doute le résultat le plus pertinent au regard des besoins des utilisateurs. Cependant, il est difficile de l'envisager en raison du grand nombre de sources qui peuvent être impliquées (des dizaines ou des centaines) et du volume des méta-données les décrivant (description des schémas source et du schéma de médiation, correspondances sémantiques, etc.). L'approche de génération automatique tente de trouver toutes les solutions possibles au calcul des objets de médiation; ce qui rend le processus automatique aussi complexe. Mais, sous certaines conditions simplificatrices que nous justifions, il est possible de générer dans des temps raisonnables un ensemble de solutions.

La complexité du processus de génération de requêtes est accrue lorsqu'on tient compte de l'hétérogénéité des données sources. Les opérations qui composent une

requête de médiation ne sont valides que si les conflits sémantiques liés aux instances sont détectés et résolus. Par exemple, une jointure de deux relations source sur l'attribut *prix* peut retourner un résultat incorrect quand les prix sont exprimés dans des monnaies différentes (ex. euro et dollars). La transformation préalable des données est une solution au problème; encore faut-il savoir détecter les conflits sémantiques et identifier les fonctions de transformation appropriées.

Nous proposons un système qui permet de générer automatiquement des requêtes SQL dans le contexte relationnel et des requêtes XQuery/XSLT dans le contexte XML. Il tient compte aussi de l'hétérogénéité des données. Pour cela, nous faisons l'hypothèse de l'existence d'une librairie de fonctions de transformation et celle de descriptions étendues des attributs, rendant explicite les unités de mesure, l'échelle des valeurs, la précision des données, etc.

Les principales étapes de notre approche dans le contexte relationnel (Kedad et al., 1999) sont : (1) Sélection d'un ensemble pertinent de sources pour le calcul du schéma de médiation ; (2) Recherche des opérations candidates entre les sources sélectionnées; (3) Recherche des transformations dans la librairie de fonctions pour résoudre les problèmes liés à l'hétérogénéité ; (4) Génération de requêtes de médiation, intégrant des fonctions de transformation, à partir de l'ensemble pertinent et des opérations candidates.

Dans le contexte XML, il émerge un problème supplémentaire qui vient de la structure hiérarchique de données. Pour résoudre ce problème, nous avons proposé une extension de l'approche dans le contexte relationnel pour générer des requêtes de médiation XQuery/XSLT. Cette approche (Kedad et al., 2005) compte trois étapes principales: (1) Décomposition du schéma de médiation en plusieurs sous-arbres, appelés *sous-arbres de médiation*. La racine de chaque sous-arbre est un nœud multivalué et tous les autres nœuds sont monovalués. (2) Recherche des différentes façons de définir chaque sous-arbre de médiation à partir des schémas de sources, appelés *mappings partiels*. La détermination des mappings partiels pour chaque sous-arbre de médiation est faite indépendamment des autres. Elle partage les mêmes étapes que la génération de requêtes pour une relation de médiation dans le contexte relationnel. (3) Combinaison de mappings partiels des différents sous-arbres de médiation pour générer les requêtes de médiation. Chaque combinaison doit satisfaire les liens sémantiques entre les différents sous-arbres.

Les figures 3 et 4 donnent une idée générale des interfaces d'utilisation de notre outil. La Figure 3 illustre l'interface de la génération de requêtes de médiation dans le contexte relationnel. L'arrière plan montre les méta-données nécessaires: le niveau supérieur et le niveau inférieur décrivent les méta-données respectivement au niveau de médiation et au niveau de source ; le niveau intermédiaire décrit les assertions entre les schémas des sources et le schéma de médiation (contraintes référentielles, correspondances linguistiques et fonctions de transformations). La fenêtre au premier plan illustre une des requêtes générée en SQL qui permet de calculer la relation de médiation (id : 1) du schéma de médiation (id : 1).

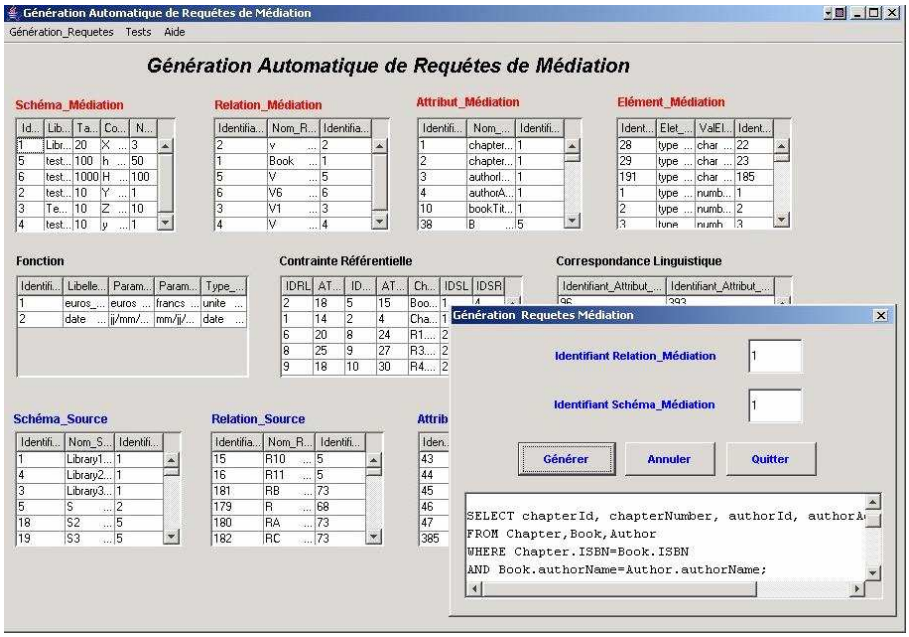

Figure 3. *Génération de requêtes de médiation dans le contexte relationnel*

La Figure 4 montre un exemple de résultat, à la fois graphique et textuel, obtenu par le générateur de requêtes XQuery. A gauche de la figure, il y a quatre requêtes de médiation générées. La requête sélectionnée (Query 1) est affichée sous forme d'une combinaison de 3 mappings partiels pour les 3 sous-arbres de médiation respectifs. Chaque mapping partiel est illustré sous forme d'un graphe : chaque nœud du graphe représente une partie source et chaque arc représente la jointure utilisée (libellé avec le prédicat de la jointure) pour combiner les deux parties correspondantes. La représentation de cette requête en XQuery est illustrée en bas à droite de la fenêtre.

Ce prototype a été réalisé dans le cadre du projet MediaGRID[1] où un scénario concret constitué de sources de données génomiques a été utilisé (Bernot et al., 2004).

[1] Projet MediaGRID – ACI GRID : Infrastructure de médiation pour l'accès transparent aux données, http://www-lsr.imag.fr/mediagrid

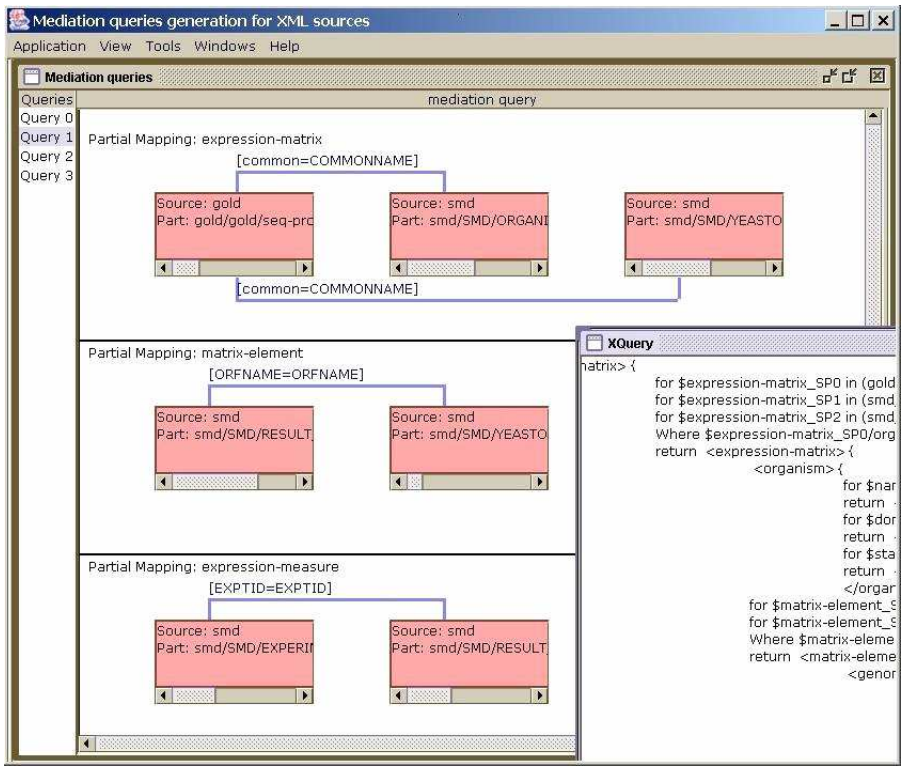

Figure 4. *Génération de requêtes de médiation dans le contexte XML*

4. Adaptabilité des requêtes aux besoins des utilisateurs

En raison du grand volume d'informations disponibles dans les sources et de la complexité croissante des systèmes d'information, l'adaptation de l'information délivrée aux utilisateurs est l'un des facteurs clés de leur succès ou de leur rejet. Dans ce contexte, la qualité joue un rôle fondamental dans la conception et l'exploitation des applications de médiation.

Notre approche est d'évaluer la qualité des résultats produits par des requêtes alternatives générés pour un objet de médiation et les confronter aux préférences des utilisateurs afin de délivrer des résultats adaptés à leurs préférences. Dans cette section nous décrivons d'abord comment exprimer les préférences des utilisateurs en utilisant des profils utilisateur et ensuite comment évaluer la qualité des données.

4.1. Gestion des profils de l'utilisateur

Les données décrivant les préférences des utilisateurs sont souvent sauvegardées sous forme de profils. Dans le cadre du projet APMD[2], nous avons défini un modèle de profil générique et extensible. Il est composé d'un ensemble de dimensions ouvertes, capables d'accueillir la plupart des informations caractérisant un utilisateur (Bouzeghoub et al., 2005). Chaque dimension est constituée d'un ensemble d'attributs dont les valeurs peuvent être simples (valeur numérique ou symbolique) ou complexes (expression logique, fonction d'utilité ou ordre de préférence par exemple). Certaines dimensions sont organisées en sous-dimensions selon la nature de leurs attributs. Nous avons identifié six dimensions du profil d'un utilisateur :

- données personnelles : des informations sur l'identité de l'utilisateur, des données démographiques, des contacts personnels et professionnels, etc.
- centre d'intérêt : définition du domaine d'expertise de l'utilisateur par des requêtes langagières et/ou des requêtes à mots clés
- qualité attendue : des facteurs de qualité des informations exprimant les exigences de l'utilisateur
- préférences de livraison : modalités de présentation des résultats (formats, taille, etc.), moment d'exécution de la requête ou type de notification de l'arrivée des résultats
- sécurité : le niveau de sécurité et de confidentialité que l'utilisateur souhaite obtenir
- historique des interactions de l'utilisateur : informations collectées implicitement ou explicitement sur le comportement de l'utilisateur (feedback)

Nous avons développé un gestionnaire de profils qui implémente le modèle générique. Son interface graphique (Figure 5) permet de créer et de manipuler manuellement les profils. La fenêtre principale de cette interface est divisée en deux parties : (1) la partie gauche qui affiche le modèle générique de profils et (2) la partie droite réservée aux instances de profils. Le modèle générique de profils regroupe l'ensemble des dimensions, des sous-dimensions et des attributs qu'un profil peut contenir. L'ensemble de ces données n'est pas forcément pertinent pour les besoins d'une application donnée. Dans notre approche, l'utilisateur a la possibilité de choisir les composants de son profil à partir du modèle générique ou de créer sa propre structure pour ensuite entrer les valeurs attendues des paramètres de personnalisation.

Pour garantir la flexibilité du modèle générique, l'interface graphique permet de définir de nouvelles dimensions, sous-dimensions et de nouveaux types d'attributs. Pour maintenir le profil générique le plus complet possible, chaque insertion d'une nouvelle dimension, sous-dimension ou attribut dans une instance de profil se fait

[2] Projet APMD - ACI Masses de Données : Accès Personnalisé à des Masses de Données, http://apmd.prism.uvsq.fr/

par le modèle générique. Par exemple si un utilisateur veut ajouter un attribut « *numéro de la carte bancaire* » à ces données personnelles, il est obligé de créer cet attribut dans le modèle générique de profil utilisateur pour ensuite le recopier dans son profil. De cette manière la typologie des données est préservée (par exemple le nom d'un utilisateur est une chaîne de caractères dans l'ensemble des instances de profils utilisateurs).

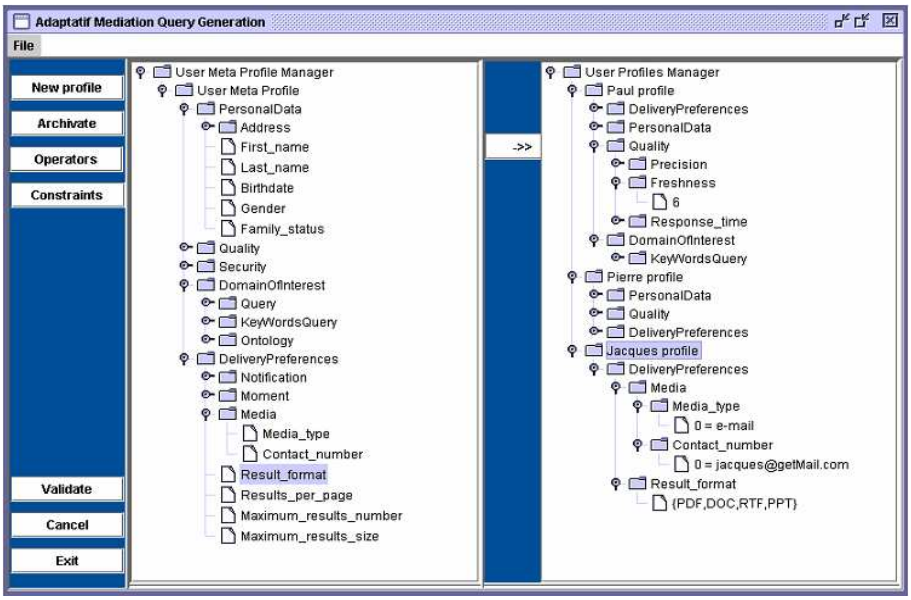

Figure 5. *Outil de gestion de profils*

Le gestionnaire de profils présenté dans cette section permet à l'utilisateur d'exprimer ses préférences qui seront confrontées aux valeurs de qualité des données que le système peut fournir. La prochaine section décrit l'évaluation de la qualité.

4.2. Evaluation de la qualité

Généralement, la qualité de l'information est exprimée ou caractérisée par un ensemble d'attributs ou de facteurs qui décrivent les données fournies aux utilisateurs (par exemple fraîcheur, précision, complétude) ou les processus qui produisent ces données (par exemple temps de réponse, fiabilité, sécurité). Par exemple, dans (Ballou et al., 1985) quatre dimensions de qualité ont été identifiés : précision, complétude, cohérence et fraîcheur, tandis que dans (Wang et al., 1996) les attributs de qualité sont analysés du point de vue de l'utilisateur. Pour un système donné, les facteurs de qualité à évaluer dépendent des utilisateurs et des applications.

Par exemple, certains utilisateurs peuvent s'intéresser au temps de réponse et d'autres à la fraîcheur des données.

L'évaluation de la qualité des données dans un système de médiation implique : (1) la sélection des facteurs de qualité à évaluer, (2) la sélection des métriques, (3) l'implémentation des algorithmes pour évaluer ces facteurs et (4) l'exécution des algorithmes pour mesurer la qualité des données produites pour le système.

Nous avons proposé un cadre générique d'évaluation de la qualité qui permet d'étudier différents facteurs de qualité, leurs métriques et des méta-données décrivant certaines propriétés du système (coûts, retards, politiques, stratégies, contraintes, etc.) (Peralta et al., 2004 A). Nous représentons les requêtes de médiation comme des graphes, isomorphes aux arbres algébriques des requêtes, adornés avec des valeurs des propriétés. La figure 6 montre un exemple du graphe adorné pour la requête Q_2 de l'exemple de motivation. Les nœuds en bas représentent des sources, le nœud en haut représente la requête de médiation et les nœuds intermédiaires représentent des opérations d'extraction et de jointure.

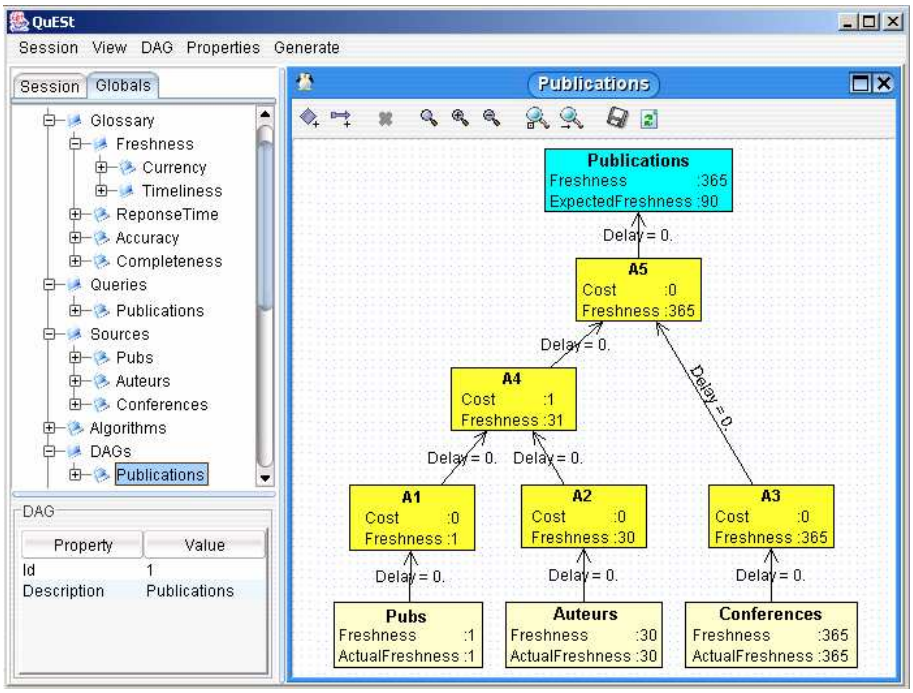

Figure 6. *Outil d'évaluation de la qualité*

L'évaluation de la qualité des résultats se fait en exécutant des algorithmes d'évaluation, chacun spécialisé dans le calcul d'un facteur de qualité. Les

algorithmes prennent en entrée les requêtes de médiation, les valeurs de qualité des données sources et les valeurs associées à un certain nombre de propriétés, combinent ces valeurs, et génèrent en sortie des valeurs qui expriment la qualité des résultats des requêtes.

Parmi les différents facteurs de qualité, nous avons choisi la fraîcheur des données pour faire une première étude de notre approche. Il existe différentes métriques pour mesurer la fraîcheur (Peralta et al., 2004 B), par exemple le temps passé depuis la création des données. La fraîcheur des données produites dépend de la fraîcheur réelle des données sources (actual freshness), des coûts d'exécution des opérations de la requête (cost) et les délais que peuvent exister entre l'exécution des opérations (delay) engendrés par exemple par la synchronisation entre ces opérations. L'algorithme de base pour mesurer la fraîcheur (Peralta et al., 2004 B) considère ces propriétés et estime la fraîcheur atteinte par chaque nœud du graphe, en utilisant les règles suivantes :

- Pour chaque noeud source A:

  Freshness (A) = SourceActualFreshness (A)

- Pour chaque noeud non source A et l'ensemble de ses prédécesseurs P:

  Freshness (A) = combine {Freshness(B) + Delay(B,A) /B $\in$ P} + Cost(A)

Pour les nœuds source, la fraîcheur de données est la fraîcheur réelle des données source. Pour les autres noeuds, la fraîcheur des données produites est calculée comme la fraîcheur des données d'entrée à laquelle on ajoute le délai et le coût. Quand un nœud a plusieurs prédécesseurs, la valeur de fraîcheur d'entrée est dérivée en utilisant une fonction de combinaison spécifique ; par exemple la valeur maximum parmi des valeurs d'entrée.

Les résultats de l'évaluation peuvent être utilisés d'une façon informative, pour avoir une mesure de la qualité des données produites par chaque requête de médiation et éventuellement comparer et trier les requêtes par rapport aux facteurs de qualité auxquels l'utilisateur s'intéresse. Les résultats de l'évaluation peuvent aussi être utilisés pour optimiser l'implémentation du médiateur (par exemple en utilisant la matérialisation), négocier avec des fournisseurs de données afin d'obtenir des données de meilleur qualité ou supprimer certaines contraintes sur les sources (comme la fenêtre de disponibilité par exemple) afin de satisfaire les besoins des utilisateurs, ou négocier avec des utilisateurs pour avoir de préférences plus précises ou plus ou moins restrictives.

Nous avons implémenté un outil d'évaluation de la qualité qui permet de choisir les propriétés les plus pertinentes pour une application donnée, associer des propriétés aux requêtes de médiation, incorporer dynamiquement des nouveaux algorithmes d'évaluation et les exécuter. La Figure 6 montre l'interface de l'outil. Dans la partie gauche, on peut gérer les différents composants du cadre de travail lequel inclut un catalogue de facteurs de qualité avec leurs métriques, des requêtes utilisateur, des sources de données, des algorithmes d'évaluation et des requêtes de

médiation. Dans la partie droite, on peut gérer les requêtes de médiation, acquérir et modifier des valeurs des propriétés, évaluer la qualité en exécutant des algorithmes et faire apparaître les requêtes pour lesquelles les besoins des utilisateurs ne peuvent pas être satisfaits. L'outil permet d'évaluer en parallèle la qualité de plusieurs requêtes de médiation. Il présente les résultats sous forme graphique, permettant ainsi de comparer les valeurs de qualité obtenues avec les préférences des utilisateurs exprimés dans leurs profils.

5. Travaux connexes à notre approche

Très peu de travaux proposent la génération automatique de requêtes de médiation. Dans le contexte relationnel, le projet Clio (Miller et al., 2000) réalise la génération d'une requête de médiation à partir d'un seul schéma source en utilisant un ensemble de correspondances de valeurs prédéfinies. Chaque correspondance spécifie une façon de définir un attribut du schéma de médiation à partir du schéma source. A la différence de cette approche, nous générons un ensemble de requêtes en considérant que chacune entre elles peut être vu comme la meilleure solution pour un groupe d'utilisateurs particulier. Notre approche supporte aussi la transformation automatique de données, ce qui est fait manuellement dans Clio. En plus, Clio peut gérer un seul schéma source alors que nous pouvons traiter la génération à partir de plusieurs schémas source.

Dans le contexte XML, l'approche présenté dans (Popa et al., 2002) permet de générer des requêtes mais à partir d'un seul schéma source et dans un format particulier ad-hoc. Il propose également un algorithme de réécriture pour l'intégration de différentes sources en utilisant les requêtes ad-hoc (Yu et al., 2004). Les requêtes que nous générons sont dans un format XQuery/XSLT qui est supporté par un plus grand nombre de systèmes et la réécriture des requêtes utilisateurs est simplifiée du fait que l'intégration des sources est déjà faite par le processus de la génération des requêtes de médiation.

Il existe dans la littérature d'autres approches (Claypool et al., 2003) (Zamboulis et al., 2004) pour générer des requêtes de médiations à partir de plusieurs schémas source XML. Ces approches comptent principalement deux étapes : (1) restructuration de chaque schéma source pour l'homogénéiser avec la structure du schéma de médiation ; (2) génération de requêtes de médiation à partir des schémas restructurés. Notre approche n'impose pas la contrainte de restructuration ce qui rend possible l'existence de requête de médiation même dans le cas où le schéma n'est pas homogène.

L'intégration des données de multiples sources, et plus précisément lorsque ces données sont extraites du Web, est un problème bien connu et mature mais il manque des approches traitant de la qualité des données (Gertz et al., 2004). L'approche de (Naumann et al., 1999) traite l'intégration des données en tenant compte de la qualité. Plusieurs requêtes de médiation sont comparées entre elles selon un certain

nombre de facteurs de qualité. L'approche consiste à propager des valeurs de qualité des données source et à les combiner en utilisant des opérateurs arithmétiques (maximum, minimum, addition, produit, etc.). Notre approche diffère de celle décrite dans (Naumann et al., 1999) sur plusieurs points. D'abords, ils considèrent seulement des requêtes de médiation composées de jointures, sélections et projections, donc les opérateurs de propagation de la qualité sont très simples. Nous considérons aussi des transformation complexes de nettoyage de données et de la matérialisation ce qui amène à des algorithmes d'évaluation plus complexes qui balancent différent propriétés du système (Peralta et al., 2004, B). En plus, ils n'utilisent pas la notion de profil ou préférences utilisateur. Les calculs effectués servent uniquement à la comparaison des différentes requêtes de médiation.

Certains travaux se concentrent sur le sous-problème de sélectionner les sources pertinentes en tenant compte de la qualité (Zhu et al., 2002) (Mihaila et al., 2000). D'autres sont guidées par l'architecture et les besoins de qualité de systèmes spécifiques (ex. entrepôts de données) qui sont des cas particuliers de notre architecture de médiation (Jarke et al., 1999) (Hull et al., 1996).

Quelques travaux étudient la qualité des données du point de vue de l'utilisateur en capturant les attributs de qualité qui sont les plus importants pour les consommateurs d'information (Wang et al., 1996) (Ballou et al., 1985). Plusieurs approches utilisent des informations caractérisant l'utilisateur afin de lui délivrer des résultats pertinents. Cependant, dans le domaine de bases de données, (Koutrika et al., 2004) est une des rares approches qui utilisent la notion de profil utilisateur. Dans le domaine de la recherche d'information, le profil de l'utilisateur est généralement défini à l'aide d'un vecteur de mots clés avec éventuellement un poids associé à chaque mot (Ferreira et al., 2001), (Gauch et al., 1999). Certains travaux ont fait un effort d'élaboration des profils structurés (P3P projet) (Amato et al., 1999). Ces tentatives de structurations sont louables mais insuffisantes pour couvrir le champ de la personnalisation. Par ailleurs, elles se contentent de catégoriser les informations de profil sans expliciter les corrélations qui existent entre elles. En effet, les attributs de comportement peuvent être corrélés aux attributs personnels ou professionnels. De même, les données de sécurité peuvent caractériser aussi bien les données personnelles que les données collectées.

6. Conclusion

La génération manuelle de requêtes de médiation est difficilement envisageable en présence d'un grand nombre de sources de données. En réponse de ce problème, nous avons proposé une approche de génération automatique de requêtes de médiation en tenant compte de la qualité des données et des préférences des utilisateurs.

Dans ce contexte, nous avons présenté un atelier d'outils dont l'objectif est de générer automatiquement un ensemble de requêtes de médiation relationnelles et

XML, d'évaluer la qualité des données produites par chacune de ces requêtes et de donner à l'utilisateur la possibilité d'exprimer ses préférences. Ces outils représentent un premier pas dans la conception de systèmes de médiation à grande échelle, capables de fournir des résultats adaptés à chaque utilisateur.

Nos travaux actuels sont orientés, entre autres, vers l'étude de la manière de combiner différentes valeurs de qualités avec les préférences utilisateur afin de sélectionner la requête la plus appropriée par rapport à cet utilisateur. Nous allons aussi étendre notre approche pour traiter la maintenance automatique des requêtes de médiation générées en fonction des changements dans les schémas de médiation ou dans les schémas source.

7. Bibliographie

Amato G., Straccia U., «User Profile Modeling and Applications to Digital Libraries», *In: Proceedings of the Third European Conf. on Research and Advanced Technology for Digital Libraries*, Paris, France, 1999.

Ballou, D., Pazer, H.: «Modeling data and process quality in multi-input, multi-output information systems ». *Management Science, Vol. 31 (2) : 150–162*, February 1985.

Bernot G., Tahi F., Laurent D., «Scénario de démonstration MEDIAGRID, Application au domaine de la génomique»*, Rapport du projet MediaGRID, 2004.*

Bouzeghoub M., Kostadinov D., «Personnalisation de l'information: aperçu de l'état de l'art et définition d'un modèle flexible de profils », *Actes de la 2ème conférence en recherche d'informations et applications CORIA'2005*, Grenoble, France, 2005.

Claypool K. T., Rundensteiner E. A., «Sangam: A Transformation Modeling Framework», *Proc. of Eighth Int. Conf. on Database Systems for Advanced Applications DASFAA'2003*, Kyoto, Japon, 2003.

Ferreira J., Silva A., «MySDI: A Generic Architecture to Develop SDI Personalised Services», *Proc. of the 3rd Int. Conf. on Enterprise Information Systems*, Setubal, Portugal, July 7-10, 2001.

Gauch S., Pretschner A., «Ontology Based Personalized Search», *Proc. of the 11th IEEE Intl. On Tools with Artificial Intelligence*, pp. 391-398, Chicago, November 1999.

Gertz M., Tamer Ozsu M., Saake G., Sattler K., «Report on the Dagstuhl Seminar: Data Quality on the Web», *SIGMOD Record Vol. 33(1)*, March 2004.

Hull R.; Zhou G.: «A Framework for Supporting Data Integration Using the Materialized and Virtual Approaches». *Proc. of the Int. Conf. on ACM SIGMOD SIGMOD'1996*, Canada, 1996.

Jarke M.; Jeusfeld M.; Quix C.; Vassiliadis P.: «Architecture and Quality in Data Warehouses: An Extended Repository Approach», *Info Systems, Vol. 24(3): 229-253*, 1999.

Kedad Z., Bouzeghoub M., «Discovering View Expressions from a Multi-Source Information System», *Proc. of the 4th. Int. Conf. on Cooperative Information Systems CoopIS'1999*, Edinburgh, Scotland, 1999.

Kedad Z., Xue X., « Mappings generation for XML data sources: a general framework », To appear *in Proc. of Int. Workshop on Challenges in Web Information Retrieval and Integration WIRI'2005*, Tokyo, Japan, 2005.

Koutrika G., Ioannidis Y., «Personalization of Queries in Database Systems», *Proc. of the 20th Int. Conf. on Data Engineering*, Boston, Massachusetts, USA, April, 2004.

Mihaila G. A., Rashid L., Vidal M.-E., «Using Quality of Data Metadata for Source Selection and Ranking». *Proc. of the 3rd Int. Workshop on the Web and Databases, WebDB'2000*, Dallas, USA, 2000.

Miller R. J., Haas L. M., Hernández M. A., «Schema Mapping as Query Discovery», *Proc. of the 26th Int. Conf. on Very Large Data Bases VLDB'2000*, Cairo, Egypt, 2000.

Naumann F., Leser U., «Quality-driven Integration of Heterogeneous Information Systems», *Proc. of the 25th Int. Conf. on Very Large Databases VLDB'1999*, Edinburgh, Scotland, 1999.

P3P projet: Platform for Privacy Preferences project, http://www.w3.org/P3P/.

Peralta V., Ruggia R., Kedad Z., Bouzeghoub M., «A Framework for Data Quality Evaluation in a Data Integration System», *Proc. of the 19th Brazilian Symposium on Databases SBBD'2004*, Brasilia, Brazil, 2004. (A)

Peralta V., Ruggia R., Bouzeghoub M., «Analyzing and Evaluating Data Freshness in Data Information Systems», *Ingénierie des Systèmes d'Information, Vol 9 (5-6) :145-162*, 2004. (B)

Popa L., Velegrakis Y., Miller R. J., Hernandez, M. A., Fagin, R., «Translating web data», *Proc. of the 28th Int. Conf. on Very Large Data Bases VLDB'2002*, Hong Kong, China, 2002.

Wang R., Strong D., «*Beyond accuracy: What data quality means to data consumers*», *Journal on Management of Information Systems, Vol. 12 (4):5-34*, 1996.

Yu C., Popa L., «Constraint-based XML query rewriting for data integration», *Proc. of Int. Conf. ACM SIGMOD SIGMOD'2004*, Paris, France, 2004.

Zamboulis L., Poulovassilis A., «XML data integration by Graph Restructuring», *Proc. of the 21st Annual British National Conf. on Databases BNCOD21*, Edinburgh, Scotland, 2004.

Zhu Y., Buchmann A., «Evaluating and selecting web sources as external information resources of a data warehouse», *Proc. of the 3rd Int. Conf. on Web Information Systems Engineering WISE'2002*, Singapore, 2002.