5

L’intégration de sources de données

Mohand-Saïd Hacid* et Chantal Reynaud†

*LIRIS, UFR Informatique
Université Claude Bernard Lyon 1
43, Blvd du 11 novembre 1918
69622 Villeurbanne
mshacid@bat710.univ-lyon.fr
†LRI, Bâtiment 490
Université Paris-Sud
91405 Orsay cedex
cr@lri.fr

Résumé

La diversité des sources d’information distribuées et leur hétérogénéité est une des principales difficultés rencontrées par les utilisateurs du Web aujourd’hui. L’infrastructure du Web sémantique doit permettre leur intégration donnant ainsi l’impression à l’utilisateur qu’il utilise un système homogène. Les solutions à l’intégration d’information proposées dans le cadre du Web sémantique tireront parti des recherches concernant les approches médiateurs et les entrepôts de données. Les premières réalisations sont en cours. Un des premiers verrous scientifiques à lever concerne le passage à l’échelle du Web. Parmi les travaux futurs, dont le développement doit être favorisé, figurent la mise en oeuvre de systèmes de médiation décentralisés, l’étude des problèmes liés à l’intégration de données multimédias, l’intégration temps réel et également la prise en compte de la complexité croissante des données à intégrer, signe d’une évolution vers une intégration de connaissances.

Mots-clés *: intégration d’information, approches médiateurs, entrepôts de données.*

Abstract

The diversity of the distributed information sources and their heterogeneity is one of the main difficulties met by Web users today. The infrastructure of the semantic Web should allow their integration giving the illusion that he uses an homogeneous system. Solutions to integration in the setting of the semantic Web will benefit from research in mediator systems and data warehouses. The first achievements are under development. One of the first scientific challenges to be addressed is Web scalability. Future research work must focalize on decentralized mediation systems, the study of problems arising when integrating multimedia data, real-time integration and must deal with the complexity of information which will lead to knowledge integration.

Keywords *: information integration, mediator approaches, data warehouses.*

1. PRÉSENTATION ET IMPORTANCE DE LA PROBLÉMATIQUE DU POINT DE VUE DES USAGES

La diversité des sources d'information distribuées et leur hétérogénéité sont une des principales difficultés rencontrées par les utilisateurs du Web aujourd'hui. Cette hétérogénéité peut provenir du format ou de la structure des sources (sources structurées : bases de données relationnelles, sources semi-structurées : documents XML, ou non structurées : textes), du mode d'accès et de requête ou de l'hétérogénéité sémantique : entre les schémas conceptuels ou ontologies implicites ou explicites sous-jacentes. Il est en effet illusoire de penser qu'une même ontologie " universelle " sera largement utilisée. Par ailleurs, les termes sont parfois exprimés dans des langues différentes.

La prise en compte de ces problèmes est une des clés de la mise en place d'applications Web sémantique. Elle s'avèrera encore plus fondamentale si l'on adhère à la vision, à plus long terme, d'agents logiciels capables de raisonner en accédant à des ressources variées. Dans ce contexte, le Web sémantique doit d'abord être une infrastructure dans laquelle l'intégration des informations d'une variété de sources peut être réalisée et facilitée. Le Web sémantique devrait donc tirer largement bénéfice des recherches déjà effectuées en intégration d'information,

concernant en particulier la réalisation de systèmes de médiation et la réalisation d'entrepôts de données et des résultats déjà obtenus.

L'aide apportée par les systèmes de médiation peut recouvrir différentes formes : découvrir les sources pertinentes étant donnée une requête posée, puis aider à accéder à ces sources pertinentes, évitant à l'utilisateur d'interroger lui-même chacune d'elles selon leurs propres modalités et leur propre vocabulaire, enfin combiner automatiquement les réponses partielles obtenues de plusieurs sources de façon à délivrer une réponse globale. De tels systèmes de médiation offrent à l'utilisateur une vue uniforme et centralisée des données distribuées, cette vue pouvant aussi correspondre à une vision plus abstraite, condensée, qualitative des données et donc, plus signifiante pour l'utilisateur. Ces systèmes de médiation sont, par ailleurs, très utiles, en présence de données hétérogènes, car ils donnent l'impression d'utiliser un système homogène. Parmi les différentes grandes catégories d'applications de ces systèmes de médiation, on peut citer les applications de recherche d'information, celles d'aide à la décision en ligne (avec entre autres l'utilisation d'entrepôts de données) et celles, de manière plus générale, de gestion de connaissances au sens large.

A titre d'illustration très simple du premier type d'applications, supposons qu'un utilisateur pose la requête suivante : quels sont les films de Woody Allen à l'affiche à Paris ce soir ? où ? leurs critiques ? Supposons l'existence de deux sources d'information. La première, Internet Movie Data Base, utilise un système de gestion de bases de données relationnel et contient une liste de films, précisant pour chacun le titre, les acteurs et le cinéaste. La seconde, Pariscope, qui peut utiliser des fichiers XML, contient, par film, les salles où le film peut être vu et, pour chaque salle, le nom de la salle et l'adresse. La réponse à la requête devra être construite en interrogeant chacune d'elles et en combinant les résultats de l'interrogation de façon à offrir à l'utilisateur une réponse globale.

Plus récemment, de nouvelles applications ont vu le jour dans les entreprises : eCRM, Business Intelligence, eERP , eKM, etc. Ces applications, que l'on désigne parfois sous le vocable de WebHouse [19] si elles sont menées dans le contexte du Web, s'appuyent sur la construction d'entrepôts de données sur le Web. Elles se trouvent également confrontées au problème de la médiation puisqu'elles mettent en œuvre un processus d'acquisition de données, souvent en temps réel, provenant de sources multiples, distribuées et hétérogènes. La conception d'outils de médiation intelligents entre les utilisateurs et les sources d'informations, accessibles via le Web ou stockées localement, est

nécessaire. Ils aident l'utilisateur à spécifier facilement les données qu'il recherche, celui-ci ayant l'impression d'utiliser un système unique et homogène.

L'approche médiateur a fait l'objet de nombreux travaux. Les résultats obtenus à ce jour sont intéressants mais ne peuvent être mis en oeuvre en l'état à l'échelle du Web. Dans le cadre du Web sémantique, l'intégration de sources d'information devra s'appuyer sur de *multiples* systèmes de médiation, ces systèmes participant de manière *distribuée et collective* au traitement des requêtes utilisateurs. Les connexions entre systèmes de médiation donneront au Web toute sa puissance, autorisant la recherche de données dans des sources non directement connectées aux sources du serveur interrogé.

2. MÉTHODES, TECHNIQUES ET OUTILS EXISTANTS SUR LESQUELS ON PEUT S'APPUYER

Les solutions à l'intégration d'informations proposées dans le cadre du Web sémantique tireront parti des recherches déjà effectuées dans le domaine. Nous présentons ci-dessous les deux approches d'intégration existantes : les approches médiateurs et les approches entrepôts de données.

2.1. L'approche médiateur

2.1.1. Présentation générale

L'approche médiateur [31] consiste à définir une interface entre l'agent (humain ou logiciel) qui pose une requête et l'ensemble des sources accessibles via le Web potentiellement pertinentes pour répondre. L'objectif est de donner l'impression d'interroger un système centralisé et homogène alors que les sources interrogées sont réparties, autonomes et hétérogènes.

Un médiateur (cf. FIG. 1) comprend un schéma global, ou ontologie, dont le rôle est central. C'est un modèle du domaine d'application du système. L'ontologie fournit un vocabulaire structuré servant de support à l'expression des requêtes. Par ailleurs, elle établit une connexion entre les différentes sources accessibles. En effet, dans cette approche, l'intégration d'information est fondée sur l'exploitation de vues abstraites

décrivant de façon homogène et uniforme le contenu des sources d'information dans les termes de l'ontologie. Les sources d'information pertinentes, pour répondre à une requête, sont calculées par réécriture de la requête en termes de ces vues. Le problème consiste à trouver une requête qui, selon le choix de conception du médiateur, est équivalente ou implique logiquement, la requête de l'utilisateur mais n'utilise que des vues. Les réponses à la requête posée sont ensuite obtenues en évaluant les réécritures de cette requête sur les extensions des vues.

L'approche médiateur présente l'intérêt de pouvoir construire un système d'interrogation de sources de données sans toucher aux données qui restent stockées dans leurs sources d'origine. Ainsi, le médiateur ne peut pas évaluer directement les requêtes qui lui sont posées car il ne contient pas de données, ces dernières étant stockées de façon distribuée dans des sources indépendantes. L'interrogation effective des sources se fait via des adaptateurs, appelés des wrappers en anglais, qui traduisent les requêtes réécrites en termes de vues dans le langage de requêtes spécifique accepté par chaque source.

2.1.2. Panorama des médiateurs existants

Les différents systèmes d'intégration d'informations à base de médiateurs se distinguent par : d'une part, la façon dont est établie la correspondance entre le schéma global et les schémas des sources de données à intégrer, d'autre part les langages utilisés pour modéliser le schéma global, les schémas des sources de données à intégrer et les requêtes des utilisateurs.

Concernant le premier point, on distingue l'approche Global As Views (GAV) de l'approche Local As Views (LAV). L'approche GAV, qui provient du monde des bases de données fédérées, consiste à définir le schéma global en fonction des schémas des sources de données à intégrer. Les systèmes suivant cette approche sont : HERMES [28], TSIMMIS [3,30], MOMIS [1]. L'approche LAV est l'approche duale. Elle est adoptée dans les systèmes suivants : Razor [8], Internet Softbot [6], Infomaster [9], Information Manifold [21, 23], OBSERVER [25], PICSEL [27]. Les avantages et inconvénients de ces deux approches sont inverses [26]. Selon l'approche LAV, il est très facile d'ajouter une source d'information, cela n'a aucun effet sur le schéma global. En revanche, la construction des réponses à des requêtes est complexe, contrairement à la construction de réponses dans un système adoptant une approche GAV qui consiste simplement à remplacer les prédicats du schéma global de la requête par leur définition.

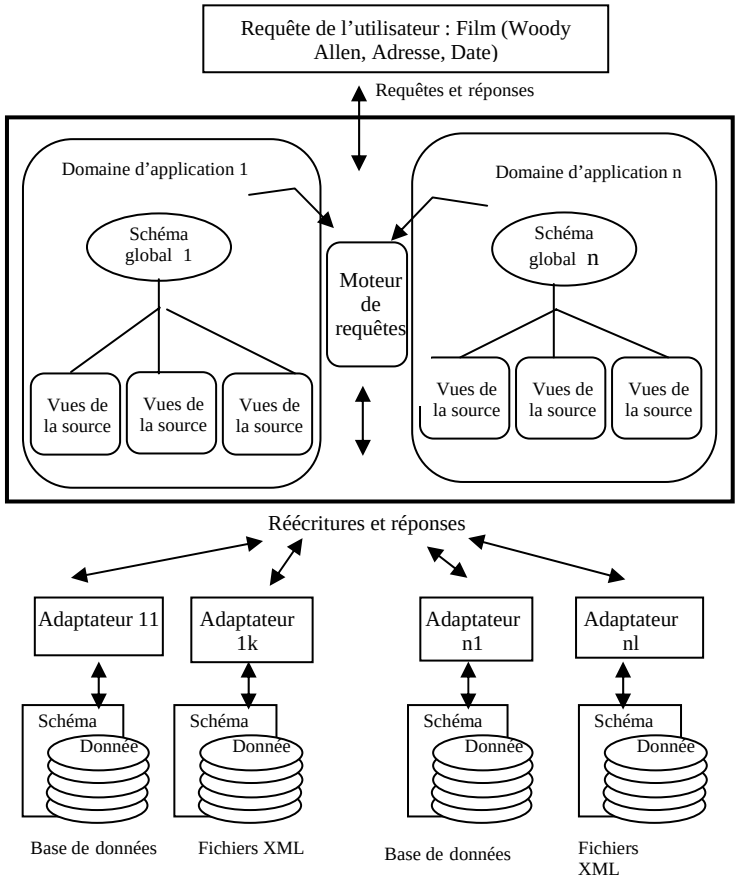

FIG. 1 - *Architecture d'un système médiateur*

Les systèmes existants se différencient également par le langage qu'ils utilisent pour exprimer le schéma global. On distingue les systèmes fondés sur un schéma global à base de règles (Razor, Internet Softbot, Infomaster, Information Manifold, HERMES), des systèmes fondés sur un schéma à base de classes (langage orienté objet (TSIMMIS)), logique de description (SIMS, OBSERVER, MOMIS), ou encore des systèmes combinant le pouvoir d'expression d'un formalisme à base de règles et d'un formalisme à base de classes (PICSEL). Enfin, plus récemment, sont

apparus des médiateurs au dessus de données semi-structurées ayant le format de documents XML (C-Web, Xyleme [23]). Ces systèmes sont fondés sur un schéma global à base d'arbres. Ils relèvent à la fois de l'approche GAV et LAV, la correspondance entre le vocabulaire du médiateur et celui des sources étant exprimée par de simples mappings de chemins.

2.1.3. Problèmes étudiés

Les travaux réalisés jusqu'alors dans le domaine des systèmes médiateurs se situent dans le contexte d'une médiation centralisée.

Dans ce cadre, des études ont porté sur les langages pour modéliser le schéma global, pour représenter les vues sur les sources à intégrer et pour exprimer les requêtes provenant des utilisateurs humains ou d'entités informatiques [11].

Des travaux ont porté sur la conception et la mise en œuvre d'algorithmes de réécriture de requêtes en termes de vues sur les sources de données pertinentes, celles-ci pouvant être connectées directement ou indirectement aux sources du serveur interrogé. Le problème à ce niveau peut consister à générer des expressions de calcul permettant de définir tous les objets du niveau global à partir des sources existantes. Le calcul de ces expressions nécessite la connaissance de l'ensemble des sources utiles à sa dérivation.

Enfin, plus récemment, certains travaux portent sur la conception d'interfaces intelligentes assistant l'utilisateur dans la formulation de requêtes, l'aidant à affiner une requête en cas d'absence de réponses ou de réponses beaucoup trop nombreuses [2].

L'idée de médiation entre sources de données utilisant des relations sémantiques locales n'est par ailleurs pas nouvelle. Ce problème a été également étudié dans le cadre des bases de données fédérées, consistant à étudier les mises en correspondance entre relations stockées. Dans le contexte du Web, toutefois, les techniques de bases de données fédérées ne sont pas réutilisables car le problème est étudié à plus grande échelle et les techniques proposées ne sont pas suffisamment flexibles. Il doit être bien plus facile de faire des ajouts ou des retraits de données et donc des mises en correspondance entre relations. Les systèmes accessibles via le Web sont par ailleurs particuliers dans la mesure où ils peuvent jouer des rôles multiples. Il peut s'agir de sources de données et/ou de systèmes intégrant des services.

2.2. L'approche entrepôt de données

Un Data Warehouse répond aux problèmes de données surabondantes et localisées sur de multiples systèmes hétérogènes, c'est une architecture capable de servir de fondation aux applications décisionnelles. Pour être exploitables, toutes les données provenant des systèmes distribués doivent être organisées, coordonnées, *intégrées* et enfin stockées pour donner à l'utilisateur une vue globale des informations.

2.2.1. Les étapes d'intégration

Nous distinguons deux niveaux dans la construction des entrepôts de données. Le premier niveau correspond à la construction des sources de données opérationnelles, et de l'entrepôt de données global. Le second niveau englobe tous les entrepôts de données locaux. La raison de cette distinction est, qu'à chaque niveau, sont associées différentes étapes de traitement et différentes difficultés techniques.

Au premier niveau, le processus de construction est décomposé en quatre étapes principales, qui sont : (1) l'extraction des données des sources de données opérationnelles, (2) la transformation des données aux niveaux structurel et sémantique, (3) l'intégration des données, et (4) le stockage des données intégrées dans le système cible. La figure 2 résume l'enchaînement de ces étapes de traitement.

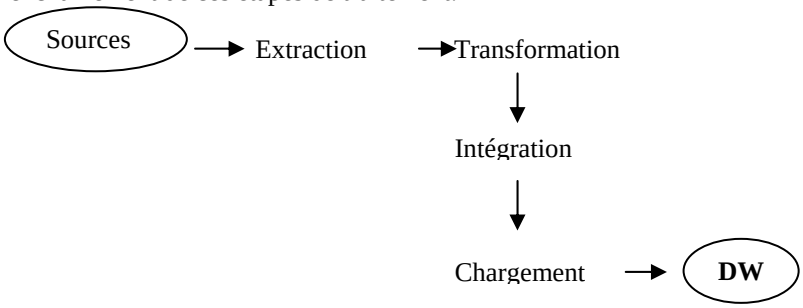

FIG. 2 - *Etapes de traitement du premier niveau de construction d'un entrepôt de données*

Notez cependant que cette décomposition est seulement logique. L'étape d'extraction et une partie de l'étape de transformation peuvent être groupées dans le même composant logiciel, tel qu'un « wrapper » ou un outil de migration de données. L'étape d'intégration est souvent couplée avec des possibilités de transformation de données riches dans un

même composant logiciel, qui, habituellement, réalise le chargement dans l'entrepôt de données. Toutes les étapes de traitement peuvent aussi être groupées dans un même logiciel, comme par exemple un système multibase. Quand les étapes d'extraction et d'intégration sont séparées, les données nécessitent d'être stockées entre les deux. Ceci peut être fait en utilisant un média par source ou un média pour toutes les sources. Une vue opérationnelle typique de ces composants est donnée par la figure 3. Les composants logiciels sont représentés par des rectangles. Les ellipses désignent des stockages intermédiaires des résultats de l'étape d'extraction/transformation. Toutes les données qui sont en entrée du composant intégration utilisent le même modèle de représentation de données. Finalement, un « wrapper » est associé à chaque source, fournissant ainsi une interface API à la source.

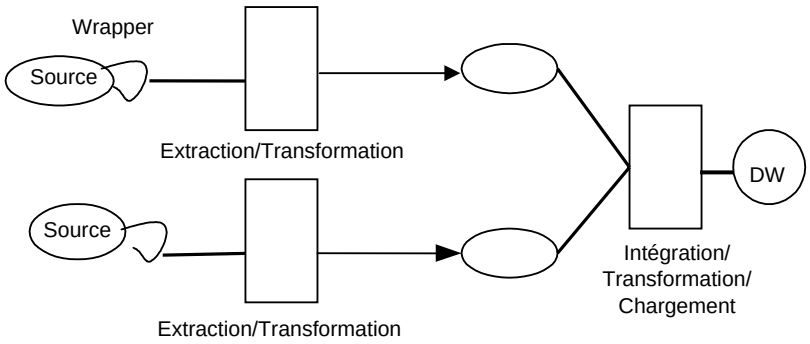

FIG. 3 - *Vue opérationnelle des composants utilisés pour la construction d'entrepôts de données*

Au second niveau, le processus de construction comporte trois étapes distinctes, qui sont : (1) l'extraction de données à partir d'une base de données (entrepôt de données local ou global), (2) le calcul des données dérivées pour l'entrepôt de données local cible, et (3) le stockage des résultats dans l'entrepôt de données local. L'étape d'extraction est un cas particulier de celle du premier niveau car les données de l'entrepôt sont stockées dans une base de données. A l'opposé, dans le premier niveau, l'extraction peut concerner des sources de données arbitraires, comme des fichiers par exemple. Le calcul des données dérivées est assez spécifique car il peut impliquer des requêtes complexes avec agrégats.

2.2.2. Les types d'intégration

Le type d'intégration réalisé dans la conception d'un entrepôt de données est celui que l'on réalise dans le domaine de l'intégration d'information, qui a été exploré dans différents domaines comme :

- les bases de données,
- les systèmes d'information coopératifs,
- les systèmes d'information globaux,
- la représentation des connaissances.

Une première classification des différentes approches repose sur le contexte d'intégration, et par conséquent, le type des entrées/sorties du processus d'intégration, et le but du processus lui-même. Nous distinguons l'intégration de schémas, l'intégration de données virtuelle, et l'intégration de données matérialisée.

- Intégration de schémas : Dans ce cas, l'entrée de l'intégration est un ensemble de schémas sources, et la sortie est un schéma de données correspondant à la représentation intensionnelle réconciliée de tous les schémas en entrée. L'entrée comporte également la spécification de la façon d'associer les schémas des données sources à des parties du schéma résultant (cible).
- Intégration de données virtuelle (médiateurs) : L'entrée est un ensemble de données sources, et la sortie est une spécification décrivant la façon de fournir un accès global et unifié aux sources dans le but de satisfaire certains besoins en information, sans interférer avec l'autonomie des sources.
- Intégration de données matérialisée : Comme dans le cas précédent, l'entrée est un ensemble de données sources, mais ici la sortie est un ensemble de données représentant une vue réconciliée des sources, à la fois au niveau intensionnel et au niveau extensionnel.

3. TRAVAUX ET RÉSULTATS EXISTANTS DU WEB SÉMANTIQUE

Face à l'ampleur du nombre de sources d'informations accessibles via le Web, le passage à l'échelle ne représente pas simplement un saut technologique. Il nécessite un véritable travail de recherche pour s'attaquer de façon fondamentale à certains verrous scientifiques qui sont des obstacles importants à la mise en œuvre d'une approche médiateur à l'échelle du Web. Des travaux relatifs à certains de ces verrous scientifiques ont déjà débuté.

Un des problèmes abordés concerne la construction d'ontologies comme support pour l'interrogation de données pré-existantes pouvant être nombreuses, sémantiquement hétérogènes et réparties dans des sources multiples. La construction d'ontologies est centrale dans le développement de systèmes médiateurs. La construction manuelle d'une ontologie, même assistée par des outils conviviaux, est un travail de modélisation long et difficile. Actuellement des travaux mettant en oeuvre différentes approches permettant d'automatiser la construction d'ontologies pour des systèmes médiateurs sont en cours de développement. Il s'agira ensuite d'exploiter les ontologies ainsi construites, qui factorisent et abstraient un ensemble important de sources d'informations, pour répondre aux requêtes des utilisateurs de façon efficace et coopérative. De telles approches ne sont possibles que si on se libère de l'hétérogénéité des formats des sources d'information considérées.

Dans le projet PICSEL 2 au Laboratoire de Recherche en Informatique de Paris-Sud, c'est le problème de l'intégration d'un grand nombre de sources de données ayant le format de documents XML qui est étudié. Un premier prototype, OntoMedia, a été développé pour extraire des composants d'une ontologie à base de classes à partir de DTD spécifiques à un domaine d'application donné [10]. Une expérimentation réalisée à partir de DTD élaborées par un organisme de standardisation de transactions commerciales laisse penser que l'approche peut permettre la construction de systèmes médiateurs plus « ouverts ». Ces derniers pourraient être capables de regrouper a priori tous les systèmes dont l'interface est conforme aux standards ayant permis la construction de l'ontologie mais qui, au moment de la construction de cette ontologie, ne

sont pas forcément identifiés. Une telle ouverture est intéressante dans une optique Web sémantique même s'il ne s'agit pas d'une approche complètement générale, s'appliquant à toutes les ressources identifiables via le Web, quelles qu'elles soient.

D'autres travaux portent sur la conception d'outils de « data mining » pour regrouper automatiquement un vaste ensemble de documents similaires [29]. L'objectif est ensuite de structurer les regroupements, de les caractériser par des descripteurs pertinents, puis de fusionner ces descripteurs, pouvant être vus comme des parties d'ontologies, de façon à obtenir une ontologie intégrée.

Un second verrou scientifique est la conception d'architectures de médiation décentralisées et facilement extensibles de façon à ce qu'un utilisateur puisse à tout moment ajouter de nouvelles données dans une source, modifier le schéma local d'une source ou les mises en correspondance entre schémas locaux, ceci quelque soit la source et son domaine d'application. Ce sont ces architectures qui permettront réellement le passage à l'échelle du Web. L'objectif est d'éviter la conception d'un schéma global unique, exigeant un gros travail de conception, difficilement extensible.

Des travaux réalisés dans le cadre des systèmes de gestion de données pair-à-pair (PDMS - Peer Data Management System) ont débuté dans cette direction à l'Université de Washington à Seattle aux Etats-Unis [13]. Ils s'appuient sur les travaux concernant les architectures distribuées pair-à-pair mais vont au-delà, ces architectures ne prenant absolument pas en compte la sémantique des données. Les premières réalisations portent sur la médiation de schémas dans les systèmes de données pair-à-pair, plus particulièrement sur l'étude et la conception d'un langage suffisamment flexible pour être utilisé dans le cadre d'une médiation décentralisée. Ce langage est une extension des formalismes d'intégration de données connus de façon à les rendre utilisables dans le cadre d'une architecture distribuée. L'objectif est d'avoir un langage qui reste très expressif mais beaucoup plus flexible. Le changement de contexte soulève deux problèmes majeurs. Le langage utilisable au sein d'une architecture distribuée doit, d'une part, permettre d'établir des mises en correspondance entre des schémas d'un système et celui de ses pairs, chaque système étant soit une source de données, soit un médiateur. Il ne s'agit plus d'établir des relations entre d'un côté, un médiateur, de l'autre côté, un ensemble de sources de données. D'autre part, le langage doit permettre de définir *localement* des relations sémantiques entre les schémas locaux de quelques systèmes et également de répondre *globalement* aux requêtes utilisateurs en exploitant le réseau des systèmes

reliés sémantiquement. Là encore, on se différencie de l'approche médiation centralisée fondée sur une architecture à deux niveaux pour laquelle les algorithmes et la complexité du problème de la reformulation des requêtes ont fait l'objet de nombreuses études.

Enfin, le troisième point que des travaux commencent à aborder concerne la mise en correspondance entre ontologies. Doan, Domingos et Halevy ont travaillé sur un problème similaire en intégration d'informations selon une approche médiateur. Le système GLUE [5] qu'ils proposent a été conçu à partir du système LSD [4] dont l'objectif était d'identifier, dans un contexte de médiation centralisée, des mises en correspondance entre un schéma global et le schéma (DTD) de sources d'information XML. Le système GLUE est appliqué au contexte du Web sémantique. Il permet d'assister le processus de mise en correspondance entre les taxinomies de deux ontologies en proposant d'utiliser plusieurs techniques d'apprentissage automatique, chacune exploitant des types d'information différents : les termes, leur format, leur fréquence, leur position, les caractéristiques des distributions de valeurs. Un méta-système combine l'ensemble des résultats obtenus. L'approche ne porte que sur des mises en correspondance de type 1-1. Le problème est posé en ces termes : étant donné un concept d'une taxinomie, quel est le concept le plus *similaire* dans la taxinomie d'une autre ontologie ?

4. RECHERCHES FUTURES POUR LE WEB SÉMANTIQUE

4.1. Vers des systèmes de médiation décentralisés

L'intégration de sources d'information hétérogènes dans le cadre du Web sémantique s'appuiera nécessairement sur de multiples systèmes de médiation. Certains systèmes pourront suivre une approche centralisée. D'autres suivront une approche décentralisée consistant à considérer une coalition de serveurs d'information, chaque serveur jouant indifféremment le rôle de serveurs de données ou de médiateurs avec ses pairs, et participant de manière distribuée et collective au traitement des requêtes des utilisateurs. Une telle architecture sera plus adaptée grâce à sa flexibilité. Dans ce contexte de médiation décentralisée apparaissent de nouveaux challenges.

Il est important de concevoir une nouvelle catégorie d'outils d'interrogation de données réparties au sein de systèmes multiples, ces

outils étant dotés de langages de requêtes riches. Il s'agit réellement d'outils d'un type nouveau dont l'utilisation doit être compatible avec la possibilité pour quiconque d'ajouter à tout moment de nouvelles données dans un des systèmes, d'établir des relations avec les concepts ou schémas déjà définis, de définir de nouveaux schémas *locaux* alors immédiatement utilisables pour poser des requêtes au niveau *global*, de définir des mises en correspondance entre schémas locaux. Une architecture pair-à-pair s'impose naturellement. Les recherches dans ce domaine sont ainsi fondamentales mais non suffisantes. A l'aspect décentralisé auquel les travaux sur les architectures pair-à-pair peuvent apporter des solutions, s'ajoute la dimension sémantique, indispensable pour connecter sémantiquement les systèmes mis en relation.

Un problème nouveau et important lié à cette dimension sémantique concerne la définition de correspondances sémantiques entre les ontologies manipulées par chacun des systèmes amenés à communiquer. Il faut pouvoir disposer d'une approche simple et naturelle de description de correspondances sémantiques entre ontologies. Le passage à l'échelle du Web n'est envisageable que si la conception de ces définitions peut être en partie automatisée. Il est donc nécessaire d'étudier comment cette automatisation est possible, sachant qu'elle devra pouvoir être établie entre des ontologies qui sont locales à des sources et qui sont hétérogènes. Les recherches pourront s'appuyer sur les travaux effectués sur la mise en correspondance de schémas proposant une automatisation partielle pour des domaines d'application particuliers. Dans le contexte du Web sémantique, néanmoins, il serait souhaitable que les solutions proposées au problème de mise en correspondance soient indépendantes de tout domaine d'application et prennent en compte toute la complexité des ontologies. En particulier, des travaux de recherche doivent s'intéresser à l'automatisation des mises en correspondance de type 1-n ou n-m, pas seulement de type 1-1. Ils doivent également chercher à exploiter les contraintes sur les attributs ou les relations définies au sein des ontologies.

Il faut ensuite pouvoir raisonner sur les correspondances entre ontologies. Il faut s'attendre à une explosion du nombre d'ontologies utilisées. Beaucoup décriront des domaines similaires mais n'utiliseront pas forcément les mêmes termes, d'autres décriront des domaines qui pourront se recouvrir. Il est nécessaire pour cela de développer des recherches portant sur la représentation explicite des mises en correspondance entre ontologies ainsi que sur la conception d'algorithmes de raisonnement efficaces et adaptés au traitement des mises en correspondance de différentes sortes : égalité, inclusion, recouvrement.

Enfin, ces systèmes distribués reposent sur l'exploitation d'ontologies elles aussi distribuées. Un champ de recherches à favoriser concerne alors la gestion à grande échelle de ce nombre très important d'ontologies pouvant couvrir des domaines identiques ou se recouvrant.

4.2. Intégration de données multimédias

La numérisation de l'information multimédia a permis l'apparition de nouveaux équipements et de nouvelles applications (enseignement à distance, télé-médecine, surveillance électronique, etc.). Cette production croissante de données multimédias numérisées amplifie les problèmes classiques de gestion de données multimédias et en crée de nouveaux tels que l'accès par le contenu, la personnalisation des contenus, l'accès à partir d'appareils mobiles, etc. Les problèmes majeurs concernent la modélisation, le stockage et l'indexation physique des données multimédias, l'intégration des données multimédias, et le traitement des requêtes sur ces données.

4.3. Intégration et analyse de données en temps réel

Les pressions résultant des demandes des clients et de la compétitivité liée à la nouvelle économie ont créé une demande insatiable pour une intégration et analyse, en temps réel, de l'information. Il n'est plus acceptable pour les décideurs de prendre des décisions en se basant sur des données datant de plus d'une semaine, voir même d'une journée. Les employés, les décideurs, les clients et tous les partenaires économiques ont besoin d'accéder à l'information quand elle est pertinente.

La possibilité d'accéder à temps et de façon simple à des données pertinentes au moyen d'outils d'interrogation et d'analyse est fondamentale pour les organisations qui souhaitent être compétitives. Cependant, avec la prolifération d'environnements hétérogènes qui doivent être intégrés à des systèmes d'aide à la décision, à des entrepôts de données, etc., les défis sont nombreux. Les données – données clients, données financières, données de navigations – constituent un avantage considérable sous réserve qu'elles soient intégrées et utilisées pour faciliter les échanges entre partenaires économiques. Une solution au problème de l'intégration de données en temps réel constituera une étape importante vers l'exploitation effective des possibilités de l'Internet dans le domaine de l'aide à la décision.

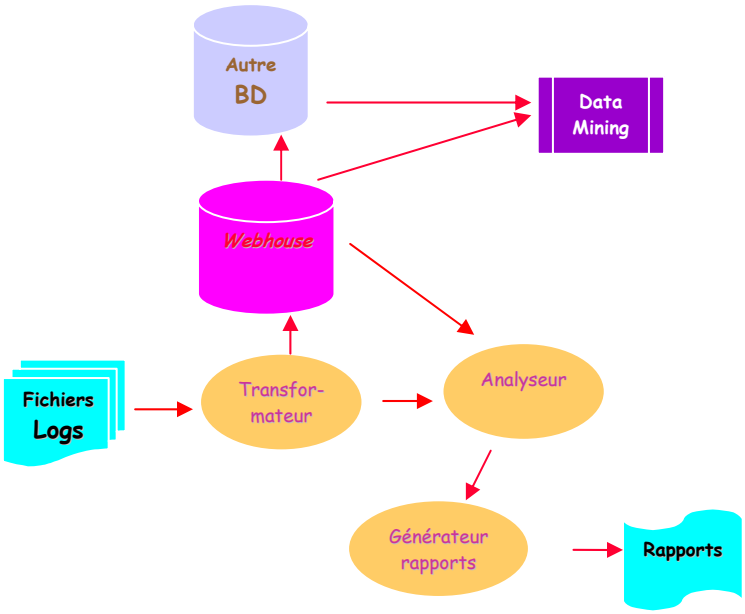

FIG. 4 - *Phases de transformation de données pour une analyse en temps réel de données collectées sur le Web*

Le traitement et l'intégration de gros volumes de données sur le Web posent des problèmes épineux comme le montrent les résultats de tests effectués sur un Pentium III, 700 MHz, 1 Go Ram et 100 Mbit Ethernet (cf. table 1).

Ainsi, dans le cas du WebHouse par exemple, le problème majeur reste celui de concevoir et de développer des agrégateurs incrémentaux efficaces. Des solutions à ce problème d'intégration de données pourraient conduire à terme à unifier proprement les différents services d'une entreprise géographiquement distribuée (cf. FIG. 5).

Taille fichiers Logs	Temps de Traitement des fichiers Logs sur le réseau	Temps de traitement des fichiers Logs sur une même machine
100 Mo	8 min	4 min
1 Go	44 min	23 min
2.5 Go	1h12 min	48 min
5 Go	2h08 min	1h32 min

Table 1 - *Temps de calcul nécessaire à l'intégration en temps réel de données sur le Web*

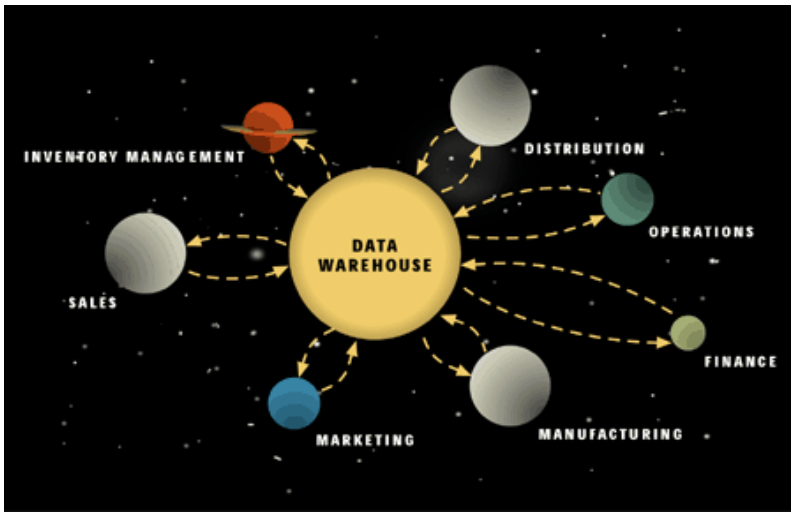

FIG. 5 – *Exemple de data Warehouse intégrant les différents services d'une entreprise géographiquement distribuée*

4.4. De l'intégration de données à l'intégration de connaissances

Bien que l'idée de construire un entrepôt de données intégré soit séduisante d'un point de vue conceptuel, elle est difficilement réalisable en pratique. Les observations indiquent que les architectures fédérées pour les entrepôts de données sont beaucoup plus pratiques des points de vues politique, opérationnel et technique [16, 7]. Les organisations réalisent de plus en plus leurs échanges via Internet et établissent des partenariats via des portails et des «extranets» avec leurs clients et leurs fournisseurs, les données pour une e-entreprise sont alors réparties entre plusieurs entités.

La notion d'entrepôt de données doit par ailleurs être étendue pour inclure non seulement les données orientées transactions, mais aussi des données créées par les employés au sein de l'entreprise. Nous devons, à l'avenir, pouvoir inclure des rapports techniques, des présentations vidéos, audio, etc.

Un autre facteur d'influence concerne le développement des services web [24], ceux-ci permettant la création de e-entreprises configurables dynamiquement. Les concepts et outils des entrepôts de données devront évoluer pour inclure des mécanismes d'accès à des bases de données de ces services web. Les informations obtenues devront pouvoir être intégrées et stockées dans des entrepôts de données fédérés. On peut ainsi imaginer des agents intelligents [17, 18] interagissant avec des fournisseurs des services web pour obtenir des informations pertinentes pour des entrepôts de données.

L'entrepôt de données deviendra alors petit à petit un entrepôt de *connaissances* comportant des données issues des entrepôts traditionnels mais aussi des connaissances du domaine, des ontologies, des méta données, etc.

5. RÉFÉRENCES

[1] Beneventano D. & Bergamaschi S. & Castano S. & Corni A. & Guidetti R. & Malzevezzi G. & Melchiori M. & Vincini M. (2000). Information integration: The MOMIS project demonstration. *In VLDB 2000 proceedings of 26th International Conference on Very large Data Bases*. September 10-14. Cairo – Egypte. p. 611-614.

[2] Bidault A. & Froidevaux CH. & Safar B. (2000). Repairing queries in a mediator approach. In 14th European Conference on Artificial Intelligence. p. 406-410. Berlin.

[3] Chawathe S. & Garcia-Molina H. & Hammer J. & Ireland K. & Papakonstantinou Y. & Ullman J. & Widom J. (1994). The TSIMMIS project: Integration of heterogeneous information sources. *In proceedings of IPSI conference*, Tokyo Japan.

[4] Doan A. & Domingos P. & Levy A. (2001). Reconciling Schemas of Disparate Data Sources: A Machine-Learning Approach. *Proceedings of the 2001 ACM SIGMOD International Conference on management of Data*. p. 509-520. Santa Barbara. CA: ACM Press.

[5] Doan A. & Madhavan J. & Domingos P. & Halevy A. (2002). Learning to map between Ontologies on the Semantic Web. *Proceedings of the international WWW Conference*.

[6] Etzioni O. & Weld D. (1994). A Softbot-Based Interface to the Internet. *Communications of the ACM*. Vol. 37(7). p. 72-76.

[7] Firestone J. M. (1999). DKMS Brief No. Nine: Enterprise integration, Data federation, and DKMS: A Commentary. *Executive Information Systems*.

[8] Friedman M. & Weld D. S. (1997). Efficiently executing information-gathering plans. *In 15th International Joint Conference on Artificial Intelligence*. p. 785-791, Nagoya. Japan.

[9] Genesereth M. R. & Keller A. M. & Duschka O. M. (1997). Infomaster: an information integration system. *In proceedings of SIGMOD 97*. p. 539-542. New-York.

[10] Giraldo G. & Reynaud Ch. (2002). Construction semi-automatique d'ontologies à partir de DTDs relatifs à un même domaine. *13èmes journées francophones d'Ingénierie des Connaissances*. Rouen.

[11] Goasdoue F. & Lattes V. & Rousset M.-CH. (2000). The use of the Carin language and algorithms for Integration Information: the PICSEL system. International Journal of Cooperative Information Systems. Vol. 9(3). p. 383-401.

[12] Gribble S. & Halevy A. & Ives Z. & Rodrig M. & Suciu D. (2001). What can databases do for Peer-to-Peer ? *WebDB01 - Workshop on databases on the Web*.

[13] Halevy A. Y. & Ives Z. G. & Suciu D. & Tatarinov I. (2003). Schema Mediation in Peer. *Data management Systems*. ICDE.

[14] Hammer J. & Garcia-Molina H. & Widom J. & Labio W. & Zughe Y. (1995). The Stanford Data Warehousing Project. In Data Engineering, *Special Issue on Materialised Views on Data Warehousing*. Vol. 18(2), p. 41-48.

[15] Hull R. & Zhou G. (1996). A framework for supporting data integration using the materialized and virtual approaches. *In proceedings of the ACM SIGMOD International Conference of the Management of Data*. p. 481-492.

[16] Kerschberg L. & Weishar D. (2000). Conceptual Models and Architectures for Advanced Information Systems. *Applied Intelligence*. Vol. 13. p. 149-164.

[17] Kerschberg L. (1997). Knowledge Rovers: Cooperative Intelligent Agent Support for Enterprise Information Architectures. *In Coopeartive Information Agents*. Vol. 1202, LNAI. P. Kandzia & M. Klusch Eds. p. 79-100.

[18] Kerschberg L. (1997). The Role of Intelligent Agents in Advanced Information Systems. *In Advances in Databases*. Vol. 1271, LNCS. C. Small & P. Douglas & R. Johnson & P. King & N. Martin Eds. p. 1-22.

[19] Kimball R. & Merz R. (2000). The data Webhouse Toolkit : Building the Web-Enabled Data Warehouse. John Wiley & Sons Inc.

[20] Kimball R.. (1996). The Data Warehouse Toolkit: Practical Techniques for Building Dimensional Data Warehouses. John Wiley & Sons, Inc.

[21] Kirk T. & Levy A. Y. & Sagiv Y. & Srivastava D. (1995). The Information Manifold. *In proceedings of the AAAI 1995 Spring Symposium on Information Gathering from Heterogeneous, Distributed Environments*, p. 85-91.

[22] Levy A. & Srivastava D. & Kirk T. (1995). Data Model and Query Evaluation in Global Information Systems. *Journal of Intelligent Information Systems*. Vol.5. p.121-143.

[23] Levy A. Y. & Rajaraman A. & Ordille J. (1996). Query answering algorithms for information agents. *In proceedings of the 13th National Conference on Artificial Intelligence (AAAI-96)*. p. 40-47.

[24] Mcllraith S. A. & Son T. C. & Zeng H. (2001). Semantic Web Services. *In IEEE Intelligent Systems*. p. 46-53.

[25] Mena E. & Kashyap V. & Sheth A. & Illarramendi A. (1996). OBSERVER: An approach for query processing in global information

systems based on interoperation across pre-existing ontologies. *In 4th Int. Conf. on Cooperative Information Systems*. p. 14-25. Bruessels. Belgium.

[26] Rahm E. & Bernstein P. A. (2001). A survey of approaches to automatic schema matching, *VLDB Journal*. Vol. 10. p.334-350.

[27] Rousset M.-Ch. & Bidault A. & Froidevaux Ch. & Gagliardi H. & Goasdoue F. & Reynaud Ch. & Safar B. (2002). Construction de médiateurs pour intégrer des sources d'information multiples et hétérogènes : le projet PICSEL. *Revue I3*. Vol.2. n°1. p.5-59.

[28] Subrahmanian V.S. & Adali S. & Brink A. & Emery R. & Lu J. J. & Rajput A. & Rogers T. J. & Ross R. & Ward C. (1995). HERMES: A heterogeneous reasoning and mediator system. *Technical Report. Univ. of Maryland*.

[29] Termier A. & Rousset M.-Ch. & SebagG M. (2002). Treefinder: a first step towards xml data mining. *In International Conference on data Mining ICDM02*.

[30] Ullman V. (1997). Information integration using logical views. *In proceedings of the 6th International Conference on Database Theory (ICDT'97)*. p. 19-40.

[31] Wiederhold G. (1992). Mediators in the architecture of future information systems, *Computer*, Vol. 25(3). p.38-49.

[32] Wiener J. L. & Gupta H. & Labio W. J. & Zhuge Y. & Garcia-Molina H. & Widom J. (1996). A System Prototype for Warehouse View Maintenance. *Proceedings of the ACM Workshop on Materialized Views: Techniques and Applications*. p. 26-33. Montreal, Canada.

[33] XYLEME L. (2001). A dynamic warehaouse for xml data of the web. *IEEE Data Engineering Bulletin*.

[34] Zhou V & Hull R. & King R. & Franchitti J.-C. (1995). Data integration and warehousing using HO2. *IEEE Bulletin of the Technical Committee on Data Engineering*. Vol. 18(2) p. 29-40.

[35] Zhou V & Hull R. & King R. & Franchitti J.-C. (1995). Using Object matching and materialization to integrate heterogeneous databases. *In proceedings of the 3rd International Conference on Cooperative Information Systems (CoopIS'95*). p. 4-18.

[36] Zhou V & Hull R. & King R. (1996). Generating Data Integration Mediators That Use Materialization. *In Journal of Intelligent Information Systems*. Vol. 6(2). p. 199-221.

[37] http://www.intelligententerprise.com/info_centers/data_int/

[38] http://www.pdit.com/

[39] http://www.datajunction.com/

[40] http://www.hummingbird.com/products/dirs/

[41] http://www.paladyne.com/