Métadonnées pour l'interrogation de sources de données hétérogènes et distribuées

Illustration sur un scénario couplant la santé et l'environnement

G. SALZANO, A. GUEMEIDA
E-mail :{salzano, guemeida}@univ-mlv.fr

Université de Marne-la-Vallée, 5, Boulevard Descartes, Champs-sur-Marne,
77454 Marne-la-Vallée Cedex 2.

Mots clés : Métadonnées, Interrogations, Qualité, Santé, Géographie, Risque

Keywords : Metadata, Querying, Quality, Health, Geography, Risk

Palabras clave : Metadata, Consultas, Calidad, Salud, Geografía, Riesgo

Résumé

Des évènements tragiques récents ont montré les difficultés liées à la gestion des émergences en santé publique, dans des situations de risque liées à l'environnement. L'analyse de ces catastrophes, montre que l'accès transparent à un grand nombre de sources d'information hétérogènes et reparties est un enjeu majeur des applications coopératives à large échelle.
Pour de telles applications, les métadonnées jouent un rôle essentiel. Ainsi, par exemple, la très récente norme ISO 11197, Information Technology - Metadata Registries, préconise d'associer explicitement des contextes applicatifs aux métadonnées, afin d'améliorer la pertinence et les performances des solutions d'interopérabilité.

Dans ce papier on souhaite analyser la pertinence des approches de médiation basées sur les métadonnées pour des applications à large échelle, couplant des informations relevant de la santé et de l'environnement. En particulier, on s'attaquera à la spécification des métadonnées d'application pour des services d'interrogation. Ces interrogations portent aussi bien sur les données contenues dans les sources d'information, que sur des critères contribuant à l'analyse de la qualité de ces sources. En analysant un premier scénario, on déterminera des éléments pour la médiation, sur la façon d'établir les correspondances entre le schéma global et les schémas locaux, ainsi que sur le type des métadonnées aux différents niveaux.

1 Introduction

Les recherches sur la gestion de l'information dans des situations de risque ont été beaucoup développées dans le milieu industriel, pour attaquer des problèmes comme les accidents liés au transport de matières dangereuses ou les incendies [1]. Elles sont moins avancées dans le domaine des risques liés à l'environnement et ayant un impact sur la santé, même si plusieurs organisations en mode réseau se développent. Par exemple, en France, dans le domaine de la santé publique, des réseaux-sentinelle ont été mis en place à différents niveaux, régional ou national, et des Observatoires Régionaux de Santé prennent en compte des spécificités locales pour l'analyse de la pollution atmosphérique.
Des évènements tragiques, comme la canicule qui a frappé la France en 2003, causant plus de 15000 décès, ou encore les inondations survenues à la Nouvelle Orléans en 2005, montrent les difficultés auxquelles sont confrontées les autorités responsables de la gestion des risques en santé publique.
La gestion du risque nécessite une vision transversale, pluridisciplinaire et à large échelle des systèmes d'information, basée sur le partage d'informations issues de plusieurs domaines. Nous considérons en particulier le couplage du domaine de la santé avec la géographie. Ce couplage est naturel, et donc nécessaire, pour évaluer les systèmes d'information d'organisations opérant dans un environnement global, régional, national ou international. Il facilite l'articulation de vues sur différents types de territoires, par exemple de santé, géographiques ou administratifs [2].
Néanmoins, la gestion du risque, dans toutes ses phases, doit faire face à des nombreux problèmes d'interopérabilité. Les deux domaines, santé et géographie, sont caractérisés par la multitude, l'hétérogénéité, la dispersion et l'évolutivité des sources d'information [3, 4].
Apporter progressivement des solutions à ces problématiques d'interopérabilité est une priorité pour les administrations, opérant à tous les niveaux, comme il est indiqué par exemple dans le Plan National Canicule, émis par le Ministère de la Santé et des Solidarités en France [5].

Objectifs et contenu de la contribution

Dans cette contribution, on souhaite analyser la pertinence des approches de médiation basées sur les métadonnées pour des applications à large échelle, couplant des informations relevant de la santé et de l'environnement.
En particulier, on s'attaquera à la spécification de métadonnées d'application pour des services d'interrogation. Dans ce papier, avec le terme "métadonnées d'application" on entend les métadonnées induites par les spécificités de l'application. Les interrogations visées portent aussi bien sur les données contenues dans les sources d'information, que sur des critères contribuant à l'analyse de la qualité de ces sources. En analysant un premier scénario, on déterminera des éléments pour la médiation, sur la façon d'établir les correspondances entre le schéma global et les schémas locaux, ainsi que sur le type des métadonnées aux différents niveaux.
Le plan de cet article est le suivant. On présentera un modèle conceptuel de données sur le partage de l'information pour des applications à large échelle (§2). Ce modèle met en évidence les problématiques d'hétérogénéité sémantique et d'indétermination. Une architecture de médiation, basée sur les métadonnées, est décrite dans le §3 et illustrée sur un scénario (§4), avant de conclure sur des perspectives de recherches ultérieures.

2 Présentation du modèle

Le risque est défini dans [6] comme un indicateur de l'état de danger qui est fonction de la probabilité d'occurrence (F) et de la gravité (G) des conséquences d'un évènement indésirable (EI). La criticité (C) d'un risque est liée à la pondération (P) des enjeux du système, C=P*G*F.
La gestion du risque est définie par l'ISO 9000, Version 2000, comme un "Processus régulier, continu, coordonné et intégré à l'ensemble d'une organisation, qui permet l'identification, l'analyse, le contrôle et l'évaluation des risques et des situations à risque, qui ont causé ou auraient pu causer des dommages à une personne ou à des biens". En effets, un événement potentiellement dangereux, un aléa, n'est un risque majeur que s'il s'applique à une zone où des enjeux humains, économiques ou environnementaux sont en présence [7]. Les risques majeurs sont regroupés en cinq grandes familles : risques naturels,

technologiques, de transports collectifs, de la vie quotidienne, liés aux conflits. Tous les risques peuvent affecter la santé des personnes.
Afin de mettre en évidence la difficulté à concevoir un cadre global de partage d'informations pour les applications d'e-gouvernement de gestion du risque, on présente un modèle conceptuel de partage d'informations, basé sur des règles de distribution [2].
Une règle de distribution est définie comme une règle logique pour déterminer la distribution des données. Chaque règle de distribution se décompose selon six dimensions, dont les mots clés sont : *Pourquoi*, *Quoi*, *Qui*, *Où*, *Quand*, *Comment.* Schématiquement :

- *Pourquoi* représente la motivation du partage de ressources d'information. La motivation est reliée à plusieurs *activités*, comme par exemple, l'évaluation de risques ou la diffusion d'information. A son tour, une *activité* peut être reliée à différents *domaines d'activité*. La protection civile, la santé et l'environnement sont concernés par toutes les activités. Selon la nature des risques, d'autres domaines peuvent s'ajouter : par exemple, l'industrie pour les risques technologiques, l'agriculture pour les risques alimentaires, le tourisme ou la pêche pour la pollution des littoraux.
- *Quoi* représente une source d'information, qui peut être reliée à un ou plusieurs *domaines d'activités*. Par exemple, les sources d'information décrivant les établissements de santé relèvent de plusieurs domaines, en priorité la santé mais aussi l'économie.
- *Où* représente un *territoire*. L'interprétation du territoire est liée au domaine d'activité. Ainsi les territoires ont plusieurs spécialisations : par exemple, *territoire de santé, administratif, géographique*.
- *Qui* représente les acteurs. On considère que les acteurs ont plusieurs spécialisations, par rapport aux domaines d'activité (environnement, santé, agriculture, …), aux liens institutionnels (services publics, collectivités, associations, …), aux activités de gestion de données (*producteurs, utilisateurs*, administrateurs, ...).
- *Quand* représente les aspects temporels du partage d'information. En gestion de risque, ces aspects concernent différentes phases, comme par exemple la prévention, la prévision, la gestion des crises, le retour d'expériences.
- *Comment* représente les aspects organisationnels du partage. Ces aspects incluent les moyens matériels et logiciels.

Le modèle de la figure 1 illustre ces éléments et met en évidence deux des principales difficultés au partage de ressources d'information: l'hétérogénéité sémantique et l'indétermination.

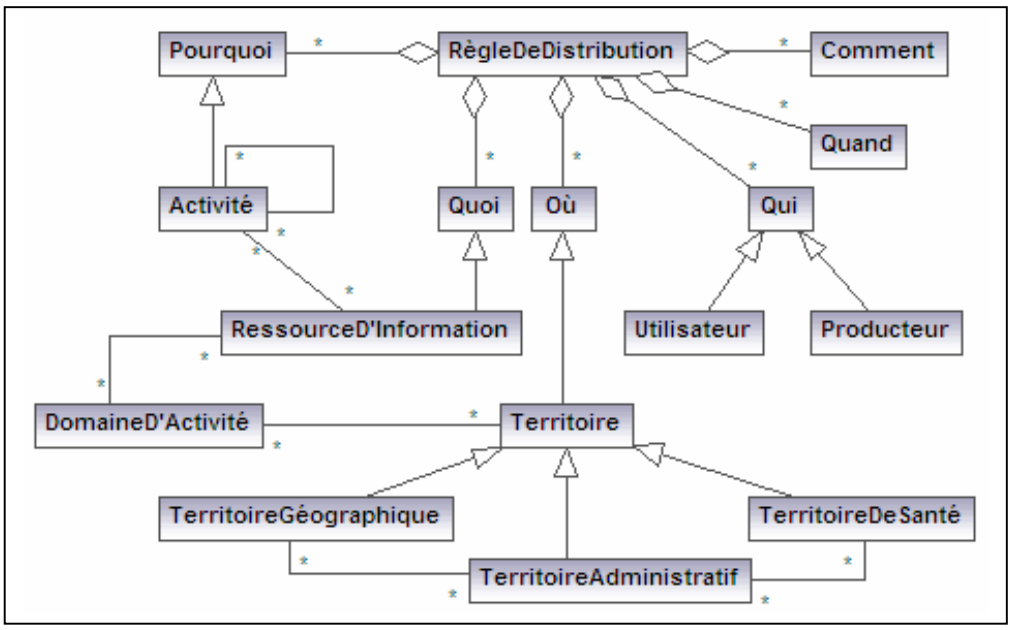

Figure 1 : Modèle conceptuel de partage des ressources d'information

L'hétérogénéité sémantique

L'hétérogénéité sémantique apparaît par exemple au niveau des territoires: les territoires de santé et géographiques, liés à des situations de risque, ont souvent des frontières floues, et en tout cas différentes de celles, figées, des territoires administratifs.

L'hétérogénéité sémantique se manifeste aussi, souvent, au niveau de la granularité des données. Les différentes granularités doivent être prises en compte pour prévenir la perte d'informations lors des opérations d'agrégation. Ainsi, par exemple, le déploiement du plan canicule en France nécessite des données acquises au niveau national, comme les données météorologiques ou les indices biomédicaux, et des données acquises au niveau régional, départemental ou local, concernant par exemple les maisons de retraite ou les services de transport médicalisés.

La standardisation facilite la resolution des problèmes d'hétérogénéité, en recommandant des modèles, des schémas, des formalismes. Les principaux acteurs du processus de standardisations sont :

- en santé, l'ISO TC 215, le CEN 251, HL7
- en géographie, l'ISO TC 211, le CEN 287, l'OpenGIS consortium
- de façon transversale, la Dublin Core Initiative et l'ISO, avec la norme 'Information Technology - Metadata Registries' (MDR, ISO/IEC 11179), qui vise à associer des contextes aux métadonnées [8]

L'interopérabilité est facilitée aussi par des actions publiques d'harmonisation des systèmes d'information et par la création de référentiels communs. Par exemple :

- en santé, la mission interministérielle MARINE (MARINE, Modernisation de l'Administration des Répertoires d'Identification Nationales et Études), vise à uniformiser les annuaires des professionnels de santé
- en géographie, le SANDRE, 'Secrétariat d'Administration Nationale des Données Relatives à l'Eau ', au sein du Réseau National des données sur l'eau (RNDE), développe un référentiel commun sur les données sur l'eau.

L'indétermination

L'indétermination est explicitée dans le modèle par la multiplicité des relations de type n-m, dont on a de nombreux exemples.

En géographie, par exemple, l'échelle de précision est un attribut essentiel pour les cartes et des critères de précision peuvent être spécifiés selon la problématique visée : l'information du public, l'aménagement du territoire, la mise en sécurité de zones particulières requièrent des cartes avec une précision croissante, avec des fourchettes de valeurs bien précises. De même, la zone de couverture, une commune, un département, une région, est liée aux objectifs de l'application. Une approche de recherche d'information guidée par la motivation permet de cibler le sous-ensemble de cartes qui mieux répondent aux objectifs [9].

De même, en santé, un très grand nombre de catalogues sont produits, au niveau des ministères, des associations scientifiques médicales, des associations des patients, des groupements pharmaceutiques, …

La pertinence de ces sources est liée fortement au contexte : évaluation des ressources sanitaires publiques, organisation de missions sanitaires s'appuyant sur des bénévoles, ….

L'indétermination peut donc être réduite en imposant des contraintes au niveau applicatif, afin de rendre plus cohérentes et performantes les recherches d'information, en fonction des usages et des contextes.

Dans la suite, on présentera une architecture de médiation, dans laquelle on envisage d'expliciter des contraintes pour améliorer globalement la qualité et la performance des interrogations.

3 Architecture de médiation

En suivant la définition générale donnée dans l'Action Spécifique 97 "Médiation via les métadonnées" du département STIC du CNRS, RTP 9, la médiation comprend tout service ou fonctionnalité qui tend à faciliter l'accès transparent des utilisateurs à des ressources reparties et hétérogènes [10]. Le catalogage, la localisation, la navigation et l'interrogation sont des exemples de services de médiation et les métadonnées sont au coeur des services [11]. Différents niveaux de métadonnées sont nécessaires à chaque service. En particulier, pour les services d'interrogation, les métadonnées sont associées aux trois niveaux : des sources, du médiateur et des applications.

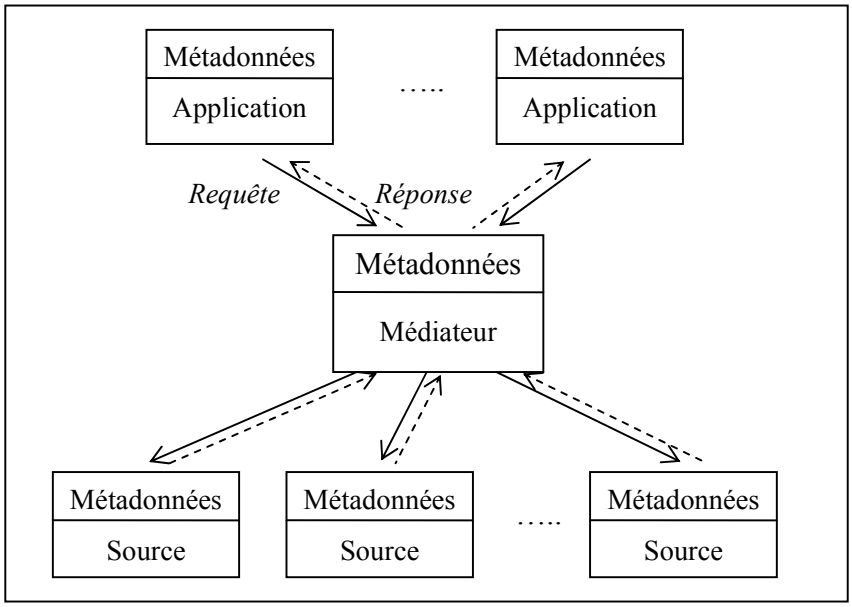

Figure 2 : Différents niveaux de métadonnées dans une architecture de médiation

Dans notre interprétation, les métadonnées incluent aussi bien les schémas, que d'autres informations pouvant augmenter la maîtrise de la qualité de l'intégration des sources, en termes de pertinence et performances des accès. Elles doivent permettre de représenter les besoins potentiels d'un large éventail d'applications, chacune avec son contexte. La figure 2, inspirée de [12], illustre le positionnement des métadonnées dans une architecture de médiation.

Métadonnées pour la médiation

Avec ce terme "métadonnées opérationnelles" on désigne des métadonnées qui fournissent des informations sur la signification et l'usabilité des sources, avant d'atteindre les données qu'elles contiennent [2].

Nous visons à expliciter des connaissances sur l'application dans l'architecture de médiation pour améliorer la qualité du dispositif en général et des interrogations en particulier. Les critères de qualité le plus fréquemment mentionnés dans le domaine des bases de données sont : l'exactitude, la complétude, l'actualité, la fraîcheur et la cohérence [13]. Tandis que l'actualité mesure le taux de valeurs obsolètes par rapport à une date fixée, la fraîcheur compare la date de saisie avec la date courante.

Notre approche tend à prendre en compte les contraintes applicatives au niveau global pour induire des métadonnées au niveau local, sur les sources: l'analyse des valeurs des métadonnées sur les sources renseigne sur la pertinence et la qualité des sources par rapport aux objectifs visés par l'application.

Ainsi, grâce à cette approche de médiation basée sur les métadonnées, il est possible de poser des questions, portant aussi bien sur la qualité des sources par rapport à l'application, que sur le contenu même des sources. Les questions visées seront par exemple : « Combien de personnes âgées, vivant seules, vulnérables au risque X, sont dans la région Y ? », ou encore « Dispose-t-on d'informations assez récentes pour évaluer le risque X dans la région Y ?»

Infrastructure de médiation

Les approches de médiation sont proches de celles de fédération [14, 15]. Elles s'appuient sur la définition de vues, pour simuler un environnement centralisé et homogène, et sur l'interrogation des sources de données au travers de ces vues.

Ainsi, l'étude d'un premier scénario et la mise en perspective de ce scénario fourniront des éléments pour spécifier l'infrastructure supportant la médiation. Ces éléments concernent d'une part la façon d'établir les correspondances entre le schéma global et les schémas locaux à intégrer, et d'autre part les langages utilisés pour modéliser le schéma global, les schémas locaux ainsi que les requêtes.

Les deux approches majeures pour décrire le schéma global et son intégration avec les schémas locaux des sources sont appelées : GAV (Global As View) et LAV (Local As View) [12, 16] :

- L'approche GAV définit le schéma global comme une vue globale sur les schémas sources. Elle est particulièrement adaptée si l'on suppose vérifiée l'hypothèse du monde fermé, selon laquelle deux conditions sont vérifiées : toutes les sources sont connues au moment de la définition du schéma global et l'ensemble des données interrogées correspond à l'union des données dans les sources
- Dans l'approche LAV le schéma global décrit un domaine d'intérêt spécifique, dans notre cas, le domaine de la gestion des risques liés à l'environnement et ayant une incidence sur la santé, indépendamment des sources de données disponibles. Les sources sont intégrées progressivement et interprétées comme vues sur le schéma global.

Malgré sa plus grande complexité algorithmique, nous avons choisi l'approche LAV pour sa pertinence à notre contexte d'étude. Ce contexte, en effet, nécessite une approche souple, facilitant le passage à des échelles très larges, comme demandé par les applications d'e-governement. Pour ces applications, les sources ne sont pas toutes connues a priori, mais doivent être recherchées au fur et mesure que les besoins de représentation du schéma global sont spécifiés.
Nous utilisons l'approche développée dan le prototype STyX [17]. Dans cette approche on construit un graphe de concepts pour représenter une ontologie du domaine étudié. Cette ontologie est constituée d'un ensemble de concepts reliés avec des rôles. Les requêtes globales sont exprimées en fonction des termes du graphe de concepts à l'aide d'une variante simplifiée d'OQL.
Pour être évaluée sur les sources locales, la requête globale est décomposée en un ensemble de requêtes locales, à l'aide d'un algorithme de décomposition. Ce dernier calcule les liaisons entre la requête globale et les sources en utilisant des règles de transformation. Celles-ci associent aux chemins XPath de la source des chemins conceptuels dans le graphe. Chaque requête locale est traduite ensuite en XQuery. Les résultats partiels ainsi obtenus sont traités par un algorithme de jointure qui compose le résultat global.

4 Illustration sur un scénario

4.1 Contexte et objectifs

Pour illustrer la pertinence de cette approche de médiation utilisant les métadonnées on considère des sources reparties, relevant des deux domaines, la santé et l'environnement. Ces sources locales sont:

- Source 1 **:** LDeptRisqueX (NumD, NomD)
- Source 2 : LEtab (NumE, NomE, TypeE, NumD)
- Source 3.1 : LPaD1 (NumP, NomP)
- Source 3.2 : LPaD2 (NumP, NomP)
- Source 4 : LRisque (NumR, NomR)

Elles sont contenues dans des bases de données relationnelles et représentent respectivement :

1. Les départements soumis au risque X
2. Les établissements de santé, avec numéro, nom et type d'établissement, n° département
3. Les personnes âgées dépendantes, vulnérables par rapport à ce risque, sur deux départements, D1 et D2
4. Les risques, avec numéro et nom du risque

On suppose que les sources 1, 2 et 4 sont établies au niveau national, tandis que les sources 3.1 et 3.2 sont établies au niveau départemental. On a considéré le jeu de données suivant :

Source 1	
NumD	NomD
1	D1
2	D2
4	D4
6	D6

Source 2			
NumE	NomE	TypeE	NumD
1	E1	Public	1
2	E2	Public	2
3	E3	Public	2
4	E4	Public	3
5	E5	Public	4
6	E6	Privé	4

Source 3.1	
NumP	NomP
1	PA1
2	PA2
3	PA3

Source 3.2	
4	PA4
5	PA5

Source 4	
NumR	NomR
1	X
2	Y
3	Z

Nous classifions les requêtes en deux familles :

- Les requêtes s'intéressant aux données contenues dans les sources, comme par exemple pour lister et localiser les ressources de santé adaptées à prendre en charge les personnes vulnérables au risque considéré :
  1. Donner, pour les départements ayant le risque X, la liste des établissements publics, triée par numéro de département
  2. Donner par département à risque X, le nombre de PA, triée par numéro de département
- Les requêtes donnant des informations sur la fraîcheur ou l'exhaustivité des sources :
  3. Chercher les départements à risque X pour lesquels l'information sur les PA est antérieure à 2005
  4. Chercher les départements à risque X pour lesquels on n'a pas d'info sur les PA.

4.2 Infrastructure de médiation

On détaille ici les éléments de l'infrastructure de médiation présentée dans la figure 2.

4.2.1 Niveau application

Le modèle conceptuel au niveau global est illustré dans la figure 3.
Par rapport au schéma présenté dans la figure 2, ce modèle comprend deux sous-ensembles de classes :

- Les classes relatives au domaine d'application, comme Risque, Département, Etablissement, PA (Personne Adulte dépendante),
- Les classes relatives aux métadonnées au niveau application, notamment Source et Territoire.

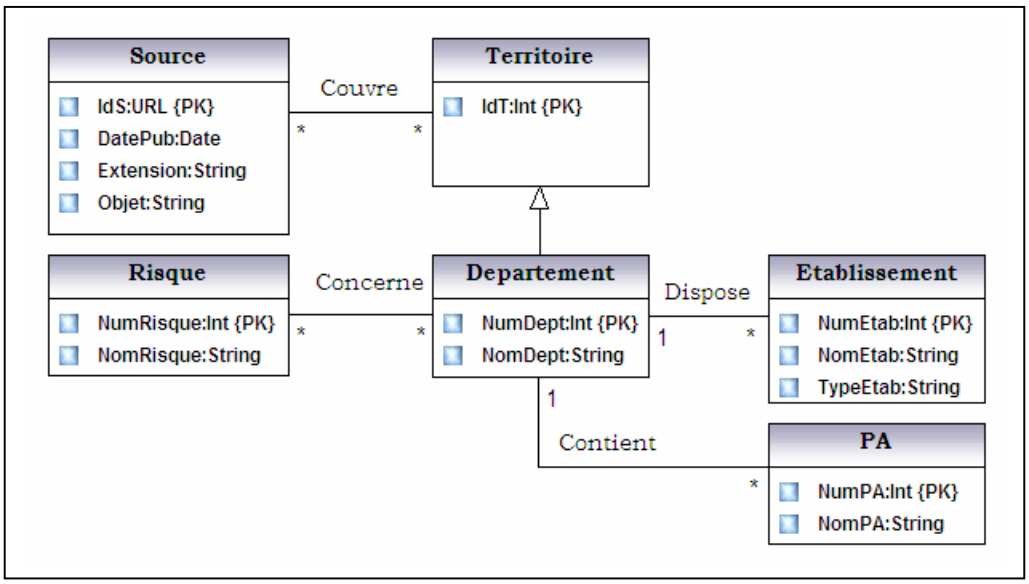

Figure 3 : Modèle conceptuel au niveau global

Le graphe de concepts associé à ce modèle est représenté dans la figure 4.

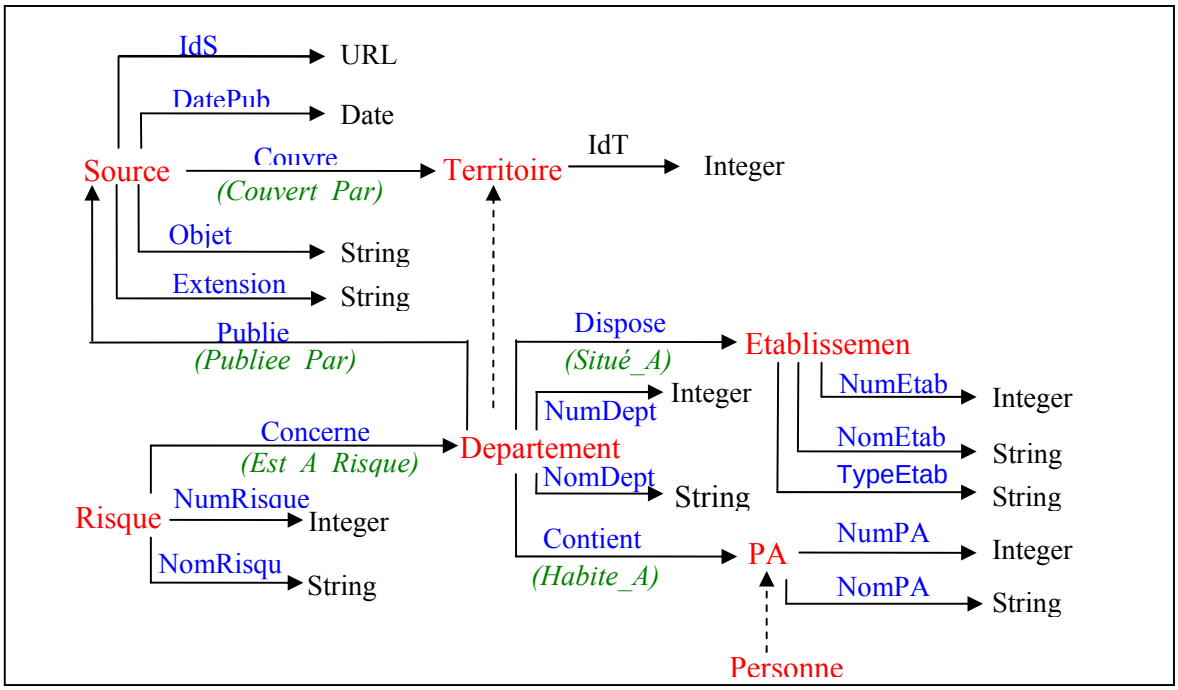

Figure 4 : Graphe de concepts

4.2.2 Niveau sources

L'extraction des données relationnelles au format XML pouvant se faire à l'aide de plusieurs outils, nous utiliserons XML comme format commun pour les sources. La structure des sources est décrite avec les XML Schémas illustrés dans la figure 5.

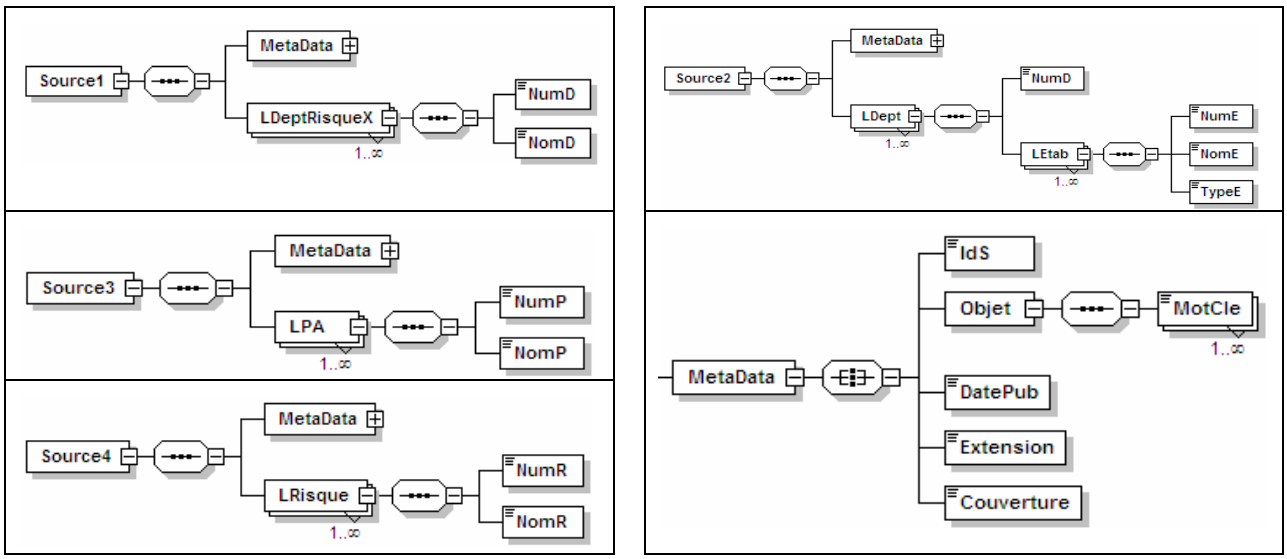

Figure 5 : Diagramme des XML Schéma

Le schéma de chaque source contient, en plus des éléments correspondants aux attributs des sources relationnelles, un élément MetaData induit par les besoins au niveau application. Dans ce scénario, l'élément MetaData est le même pour toutes les sources. Dans des cas concrets, il pourra dépendre des

sources et faire référence à plusieurs espaces de noms, pour référencer par exemple des éléments normatifs du Dublin Core [18] ou de la norme ISO 19115 [19].

4.2.3 Niveau médiateur

Les métadonnées de ce niveau décrivent les correspondances entre les schémas des sources locales et le schéma global.

Source 1 R1 :http://www.gestionrisque.fr/source1.xml//Source1 as u1 ← Risque R2 : u1/MetaData/Objet/MotCle[3] as u2 ← NomRisque R3 : u1/LDept as u3 ← Concerne R4 : u3/NumD as u4 ← NumDept R5 : u3/NomD as u5 ← NomDept	**Source 2** R1 : http://www.sante.fr/source2.xml//Source2 as u1 ← Departement R2 : u1/NumD as u2 ← NumDept R3 : u1/LEtab as u3 ← Dispose R4 : u3/NumE as u4 ← NumEtab R5 : u3/NomE as u5 ← NomEtab R6: u3/TypeE as u6 ← TypeEtab
Source 3.1 (analogue à Source3.2) R1 : http://www.pa01.fr/source3.1.xml//Source3 as u1 ← Departement R2 : u1/MetaData/Objet/MotCle as u2 ← NumDept R3 : u1/LPa as u3 ← Contient R4 : u3/NumP as u4 ← NumPA R5 : u3/NomP as u5 ← NomPA	**Source 4** R1 :http://www.environnement.fr/source4.xml//source4/LRisque as u1 ← Risque R2 : u1/NumR as u2 ← NumRisque R3 : u1/NomR as u3 ← NomRisque

Les règles de mise en correspondance pour l'élément MetaData, sur la Source 4 sont:

R4 : http://www.environnement.fr/source4.xml//Source4/MetaData as u4 ← Source
R5 : u4/IdS as u5 ← IdS
R6 : u4/Objet/MotCle as u6 ← Objet
R7 : u4/DatePub as u7 ← DatePub
R8 : u4/Extension as u8 ← Extension
R9 : u4/Couverture as u9 ← Publie_Par.NumDept

4.3 Formulation des requêtes, décomposition et résultats

On présente la formulation des requêtes 1 et 3, leur décomposition et les résultats obtenus. Nous avons utilisé l'outil Altova XMLSpy pour éditer et valider des schémas XML, des documents XML et les interroger en XQuery. La décomposition des requêtes sur le schéma global en requêtes sur les sources locales est faite manuellement.

Requête 1 : Donner, pour les départements soumis au risque X, la liste des établissements publics, triée par département

La requête globale (Figure 6) doit parcourir les concepts : Risque, Département et Etablissement en utilisant les rôles Concerne et Dispose.

La liaison $\beta 1$= {a → R1, b → R3, c → R2, d → R4, f → R5} entre les variables de la requête et les règles de correspondance de la Source 1, est une liaison maximale qui n'est pas complète. Ceci implique que la Source 1 ne répond pas à l'intégralité de cette requête et une décomposition est nécessaire. On évalue sur la Source 1 la requête préfixe Q1 et sur la Source 2 la requête suffixe Q2 générée à l'aide de la variable de jointure b [20].

Q : Requête globale

```
Select    d, f, h, j
from      Risque a,
          a.Concerne b,
          a.NomRisque c,
          b.NumDept d,
          b.NomDept f,
          b.Dispose g
          g.NumEtab h,
          g.NomEtab j,
          g.TypeEtab k
where     c="X" and k="Public"
order by  d
```

Q_1 → Source 1

```
select    d, f
from      Risque a,
          a.Concerne b,
          a.NomRisque c,
          b.NumDept d,
          b.NomDept f,
where     c="X"
```

Q_2 → Source 2

```
Select    d, h,j
From      Departement b,
          b.NumDept d,
            b.Dispose g,
            g.NumEtab h,
            g.NomEtab j,
            g.TypeEtab k
Where     k="Public"
```

En reformulant les requêtes Q1 et Q2 à l'aide des règles de description de chaque source, on obtient:

```
select    d, f
from      doc(http://www.gestionrisque.fr/
          source1.xml)/Source1    a,
          a./LDeptRisqueX b,
          a./MetaData/Objet/MotCle[3] c,
          b./NumD d,
          b./NomD f
where     c="X"
```

```
select    d, h, j
from      doc(http://www.sante.fr/source2.xml)/Source2  b,
          b./NumD d,
          b./LEtab g,
          g./NumE h,
          g./NomE j,
          g./TypeE  k
where     k="Public"
```

La réécriture de ces deux requêtes en XQuery, sans le balisage XML, donne respectivement :

```
for       $a in doc ("Source1.xml")/Source1,
          $b in $a/LDeptRisqueX,
          $c in $a/MetaData/Objet/MotCle[3],
          $d in $b/NumD,
          $f in $b/NomD
where     $b="X"
return    $d, $f
```

```
for       $b in doc ("Source2.xml")/Source2/LDept,
          $d in $b/NumD
return    {$d,
          {for    $g in $b/LEtab,
                  $h in $g/NumE,
                  $j in $g/NomE,
                  $k in $g/TypeE
          where  $k="Public"
          return   $h, $j      }}
```

La requête XQuery ci-contre génère la jointure des résultats obtenus de Q1 et Q2, à l'aide de la variable « Departement» et de l'élément «NumDept» représentant la clé:

```
for     $a in doc("resultatQ1.xml")/Resultat/Departement,
        $b in doc("resultatQ2.xml")/Resultat/Departement
where                $a/NumDept=$b/NumDept
order by             $b/NumDept
return               $a/* union $b/* except $b/NumDept
```

Les résultats des requêtes Q1 et Q2 et de la requête de jointure, après une transformation XSLT, sont:

Résultat Q1

NumDept	NomDept
1	D1
2	D2
4	D4
6	D6

Résultat Q2

NumDept	Etablissement	
	NumEtab	NomEtab
1	1	E1
2	2	E2
	3	E3
3	4	E4
4	5	E5

Résultat Q (Global)

NumDept	NomDept	Etablissement	
		NumEtab	NomEtab
1	D1	1	E1
2	D2	2	E2
		3	E3
4	D4	5	E5

Requête 3 : Chercher les départements à risque X pour lesquels l'information sur les personnes adultes dépendantes (PA) est antérieure à 2005

Cette requête s'appuie sur l'élément MetaData de chaque source pour rechercher des informations sur les PA. La requête globale s'écrit ainsi :

```
select  d, f, g
from    Risque a,
        a.NomRisque b,
        a.Concerne c,
        c.NumDept d,
        c.Publie e,
        e.IdS f,
        e.DatePub g,
        e.Extension h,
        e.Objet i
where h= 'Departement'  and  i = 'PA'  and  g < '2005-01-01'  and  b="X"
```

La décomposition de cette requête génère la même requête préfixe Q1 associée à la Requête 1, et une requête suffixe, nommée Q3.2, à évaluer sur toutes les sources. Q3.2 et sa traduction en XQuery sont:

```
select  d, f, g
from    Departement c,
        c.NumDept d,
        c.Publie e,
        e.IdS f,
        e.DatePub g,
        e.Extension h,
        e.Objet/MotCle i
where   h= 'Departement'
and     i = 'PA'
and     g < '2005-01-01'
```

=> XQuery

```
for     $c in doc("Source3.1.xml")/Source3/MetaData,
        $d in $c/Couverture,
        $f in $c/IdS,
        $g in $c/DatePub,
        $h in $c/Extension,
        $i in $c/Objet
where   $h = "Departement"
and     $i/MotCle = 'PA'
and     $g < xs:date("2005-01-01")
return  $d, $f, $g
```

Le résultat de la requête Q3.2 est :

NumDept	Source	
	IdS	DatePub
1	http://www.pa01.fr/source3.1.xml	2004-12-20

Le résultat global de la requête 3 s'obtient en faisant l'union des résultats de la requête suffixe et ensuite la jointure avec les résultats de la requête préfixe (départements à risque X). Il est donné par.

NumDept	NomDept	Source	
		IdS	DatePub
1	D1	http://www.pa01.fr/source3.1.xml	2004-12-20

5 Conclusions

Dans ce papier, nous avons analysé une approche de médiation basée sur les métadonnées, en l'illustrant sur un scénario, pour aborder simultanément deux objectifs : interroger des sources hétérogènes et distribuées et évaluer la qualité de ces sources par rapport à des besoins exprimés au niveau global, de l'application.

Nous souhaitons poursuivre nos recherches dans ce deuxième objectif et approfondir la formalisation des spécifications des besoins applicatifs et de leur prise en compte dans un système d'interrogation pour des applications à large échelle, couplant des informations relevant de la santé et de l'environnement. En particulier, nous allons focaliser sur la formalisation des critères de qualité concernant la fraîcheur et l'exhaustivité des sources, qui sont très critiques dans les applications de type e-governement. D'autre part, nous souhaitons valider notre approche et la formalisation des besoins applicatifs sur des données réelles.

6 Références

[1] Guarnieri F., Garbolino E. : "Systèmes d'information et risques naturels", Presses de l'Ecole des Mines de Paris, 2004

[2] Salzano G.: "Modeling Metadata for Multidomain Health and Geography Information", International Symposium on Generalization of Information ISGI 2005, Horst Kremers, Berlin, Germany, 14-17 September 2005

[3] Beuscart R.,"Travail coopératif et Réseau", Informatique et Santé, (10), Paris Springer-Verlag, pp. : 3-10, 1998

[4] Lallement, R. : "Publier les métadonnées pour partager les données", Le Géo-événement 2005, France (2005)

[5] Ministère de la santé et des solidarités, Ministère délégué à la sécurité sociale, aux personnes âgées, aux personnes handicapées et à la famille, Plan National Canicule (PNC), France, 2005

[6] Poullain I., Lespy F.: "Gestion des risques. Guide pratique à l'usage des cadres de santé", Editions Lamarre, 2002

[7] Ministère de l'Ecologie et du Développement Durable, Prévention des Risques Majeurs, http://www.prim.net/

[8] ISO/IEC 11179, Information technology - Metadata Registries (MDR) http://metadatastandards.org/11179/

[9] Nedellec J.L.: "Optimisation des méthodes de cartographie des risques en fonction de la finalité", Ecole thématique UMLV-CNRS, Risques Mouvements de Terrain, 5 octobre 2005

[10] Action spécifique 97 du département STIC du CNRS, Médiation via les métadonnées, Animatrice T. Libourel, novembre 2003, http://www.lirmm.fr/~libourel/MM/MetaMedia.htm

[11] Sheth A.: "Changing Focus on Interoperability in Information Systems: From System, Syntax, Structure to Semantics", in Interoperating Geographic Information Systems M F Goodchild, M J - Egenhofer, R Fegeas and C A Kottman (eds), Kluwer Publishers, 1999

[12] Amann B.: "Du partage centralisé de ressources Web à l'échange de documents intensionnels", Document de synthèse présenté pour obtenir l'Habilitation à Diriger les Recherches, 18 novembre 2003, http://ftp.lip6.fr/lip6/reports/2003/lip6.2003.014.pdf

[13] Berti-Equille L.: "Un état de l'art sur la qualité des données", RSTI - ISI – Qualité des systèmes d'Information, pages 117 à 143, 9/2004

[14] Elmagarmid A., Rusinkiewicz M., and Sheth A. (eds): "Management of Heterogeneous and Autonomous Database Systemes", Morgan Kaufmann Publishers, Inc. ,1999

[15] Salzano G.: "Integration Methodology for Heterogeneous Databases" in Heterogeneous Information Exchange and Organizational Hubs, edited by H. Bestougeff, J.E. Dubois, B. Thurasingsham, Kluwer Academic Publishers, Netherlands, 2002, pages 1-16

[16] Levy A.: "Answering queries using views: a survey", VLDB Journal, 2001 http://www.cs.washington.edu/homes/alon/site/files/view-survey.ps

[17] Fundulaki I., Amann B., Beeri C., Scholl M.: "STYX: Connecting the XML World to the World of Semantics", 2002. Demonstration at EDBT'2002

[18] Dublin Core Metadata Initiative, DC**,** http://dublincore.org/

[19] UK Gemini: A Geo-spatial Metadata Interoperability Initiative - ISO 19115: Metadata Standard – Proposed Element Set (2003) http://www.gigateway.org.uk/metadata/pdf/ISO19115ProposedElements.pdf

[20] Amann B., Beeri C., Fundulaki I., Scholl M.: "Ontology-based integration of xml web resources", in International Semantic Web Conference (ISWC), Sardinia, Italy, 2002