APPROCHE DE RECHERCHE, SELECTION ET INTERROGATION DE SOURCES DE DONNEES HETEROGENES ET DISTRIBUEES

Illustration sur un exemple couplant la santé et l'environnement

Abdelbasset GUEMEIDA, Gabriella SALZANO
guemeida@univ-mlv.fr, salzano@univ-mlv.fr

Université de Paris-Est Marne la Vallée,
Laboratoire Sciences et Ingénierie de l'Information et de l'Intelligence Stratégique (S3IS)
5, Boulevard Descartes, Champs-sur-Marne, 77454 Marne-la-Vallée Cedex 2, France

Mots clefs :
Recherche et sélection de sources de données, Métadonnées, Catalogues, Qualité de données, Intégration de données.
Keywords:
Quality aware data search, Metadata, Catalogs, Data Integration.
Palabras clave :
Búsqueda y selección de datos, Meta datos, catálogos, datos calidad, datos integración.

Résumé

Dans ce papier nous présentons une approche de recherche, sélection et interrogation de sources de données distribuées et hétérogènes pour répondre à des besoins décisionnels. Cette approche est guidée par les besoins applicatifs et est basée sur les métadonnées. En correspondance d'une requête, l'approche déroule trois étapes s'appuyant d'abord sur les catalogues et leurs métadonnées et ensuite sur les sources de données contenues dans ces catalogues et leurs métadonnées. Les activités sont : (i) rechercher des catalogues référençant des sources de données pertinentes pour la requête, (ii) sélectionner dans les catalogues retenus des sources de données ayant une qualité suffisante par rapport aux besoins exprimés, (iii) interroger des sources sélectionnées par des techniques d'intégration. Nous détaillons une architecture globale supportant une vue unifiée de ces étapes, par l'association de systèmes de raisonnement et d'un système de médiation. Nous présentons aussi les choix techniques réalisés en termes de langages de spécification (OWL DL, langages de règles), et d'implémentation (Protégé, Pellet et Jess). L'approche est illustrée sur un exemple simplifié des besoins informationnels liés à l'environnement et à la santé.

1 Introduction

Face à l'explosion de la production d'information, les utilisateurs nécessitent de plus en plus d'aides automatiques pour rechercher et sélectionner des sources de données pertinentes pour leurs besoins décisionnels. Cette aide complexe consiste souvent en “robots” logiciels, en interfaces d’accès, à des catalogues ou métadonnées. La collecte et la sélection des données constituent une partie lourde et coûteuse du travail, suivie par la mise en œuvre du modèle de l’application, qui se termine par la prise de décision. Dans les grandes applications, comme en santé publique, nous pensons qu’il est indispensable de considérer la recherche et la sélection des données comme des activités à part entière.

L’objectif de ce papier est de proposer une approche de recherche, sélection et interrogation de sources de données basée sur les besoins préliminaires de qualité et s'appuyant largement sur les métadonnées. L'approche a pour buts d'optimiser les deux premières étapes (recherche et sélection), pour ensuite intégrer des données pertinentes.

En opposition à des approches appelées "agnostiques", les approches de recherche d'information intégrant des critères de qualité (quality-aware), constituent un sujet de recherche très critique pour tous les domaines liés à la veille stratégique, scientifique et technique : veille territoriale et recherche d'information dans tous les secteurs d'activités, de l'administration comme de l'industrie et des services. Ces approches contribuent aux processus décisionnels et peuvent être considérés en amont des approches d'aide à la décision de type OLAP [4].

Les comparaisons de contextes, des critères et des valeurs des indicateurs de qualité peuvent être utilisés ensemble pour améliorer les modes de recherche d'information, rendre celle-ci plus performante et plus précise [9]. En effets, la prise en compte des besoins préliminaires liés aux contextes applicatifs et à la qualité permettent de (i) réduire l'ensemble des sources candidates, adressées aux outils d'intégration, (ii) faciliter une vérification, par l'utilisateur, de l’usabilité, dans un contexte précis, des données qui lui sont proposées, (iii) garantir un certain niveau de qualité dans la prise de décision, avec une traçabilité des choix opérés. Ces remarques guident la conception d'e-services dans des systèmes à large échelle [5]

Le plan de cet article est le suivant. Dans le §2, on présente deux difficultés majeures pour l'élaboration de systèmes d’intégration de données : l'hétérogénéité et l'indétermination des sources. Dans le § 3, on présente une vue d’ensemble de l’approche d'intégration basée sur les métadonnées, pour donner un accès transparent à des sources d'information issues des deux domaines sectoriels, santé et géographie. A partir des exigences, des catalogues et des métadonnées, l'approche permet de (i) rechercher les catalogues référençant des sources de données pertinentes (ii) sélectionner des sources de données après leur évaluation sur des critères de qualité externe spécifiques au contexte et (iii) interroger des sources en décomposant les requêtes globales en requêtes partielles, extraire les contributions à partir des sources locales et les fusionner.

Les choix d'architecture sont présentés dans le §4. L'infrastructure technologique de cette approche couple un système d'inférence et un système de médiation (§5). Elle utilise des langages et outils de spécification et interrogation du web sémantique (OWL-DL, SWRL, Protégé, Pellet). L'application de l'approche à un jeu de données simplifié, représentatif d'un exemple couplant la santé et l’environnement, ainsi que les résultats obtenus, sont commentés (§6), avant de présenter des perspectives de recherche.

2 Hétérogénéité et indétermination des sources

Deux difficultés majeures pour les approches d'aide à la recherche, à la sélection et à l’intégration de données sont : l'hétérogénéité sémantique de données et l'indétermination des sources.

- L'hétérogénéité sémantique concerne les niveaux de la granularité (spatiale, temporelle et de spécialisation), des données qui doivent être prises en compte lors des opérations de filtrage ou d'agrégation. Par exemple, le déploiement du plan canicule en France nécessite de données acquises au

niveau national, comme les données météorologiques ou les indices biomédicaux, et de données acquises au niveau régional, départemental ou local, concernant par exemple des infrastructures de support médico-social (maisons de retraite, services de transport médicalisé). L'hétérogénéité sémantique apparaît particulièrement au niveau des territoires : les territoires de santé et géographiques, liés à des situations de risque, ont souvent des frontières floues, différentes de celles, figées, des territoires administratifs.

- L'indétermination est liée à la multiplicité des différentes dimensions des sources de données : thèmes, territoires, usages possibles. Par exemple, dans les sources géographiques, plusieurs échelles de précision et zones de couverture peuvent être associée à une même thématique, et la pertinence des sources est liée au contexte applicatif (prévention des risques, secours). Les difficultés liées à l'indétermination sont liées aux besoins de préserver et d'identifier une variété de vues utilisables selon le contexte.

3 Vue d'ensemble de l'approche

Le but de l'approche est de pouvoir s'appliquer à tout objectif particulier, comme l'étude de la propagation d'une maladie, ou la recherche des ressources médicales ou logistiques face à un risque environnementale.

Ainsi, l'approche doit guider les interrogations sur :

- **les catalogues** référençant des sources de données pertinentes pour l'étude;
- **la qualité**, de ces données, en adéquation (fitness) avec le but fixé;
- **les contenus**, qui doivent assurer un usage cohérent et opérationnel.

D'abord, nous accédons aux Catalogues, qui identifient les sources de données et fournissent, via les métadonnées, quelques informations sur la couverture, le format, parfois la date comme dans l'exemple MDWEB [8]. Ensuite chaque source cataloguée donne accès à ses métadonnées, qui fournissent des informations plus détaillées sur la Qualité, la description des domaines et du vocabulaire des sources données, le découpage en niveaux de spécialisation, etc. Enfin nous pouvons accéder aux Contenus.

3.1 Les trois étapes de l'approche

Etape1 : Recherche des catalogues

A partir de l'ensemble des catalogues et des métadonnées des sources référencées, on procède à une recherche selon les trois dimensions : Thème, Espace, Temps. Cette étape doit s'appuyer sur une stratégie de recherche et prendre en compte des relations de correspondances sémantiques, géométriques et topologiques, s'appuyant sur des thésaurus ou ontologies.

Etape 2 : Sélection des sources

La sélection des sources pertinentes se fait à partir de listes produites lors de l'étape précédente. Elle doit prendre en considération les critères de qualité imposés, par exemple par rapport à la complétude ou à la validité temporelle. Cette étape permet de réduire a priori la taille d'exploration en utilisant un ordre de préférence approprié entre les solutions.

Etape 3 : Interrogation sur les contenus

Lorsqu'une liste réduite de sources de données a été sélectionnée, on peut enfin confronter les données aux contraintes d'intégrité du schéma global. Cette tâche est coûteuse car les données sont volumineuses, et la probable détection de conflits peut rendre tout calcul exponentiel.
Finalement, le processus peut boucler jusqu'à une décision acceptée pour un niveau de risque d'erreur remis à jour.
Les deux premières étapes sont réalisables très tôt, dès l'accès aux catalogues et métadonnées. De plus, on peut anticiper certains paramètres et mémoriser certaines informations, pour préparer efficacement et donc améliorer l'étape ultérieure d'accès aux données. Cette démarche s'adapte à la directive européenne INSPIRE et aux normes associées.

3.2 Formalisation de l'approche

Une vue intégrée des trois étapes est représentée dans la figure 1.

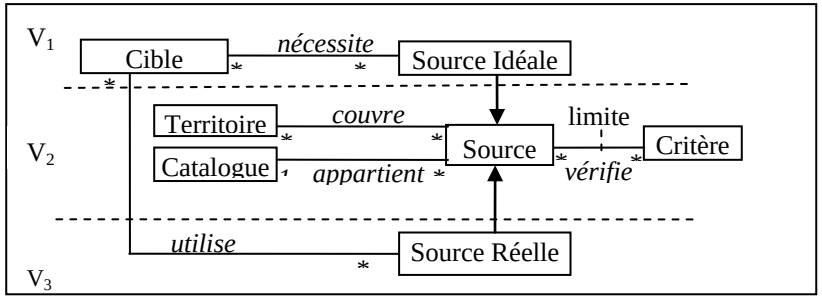

Figure 1. Vue d'ensemble des aspects d'existence, qualité et contenu des données

Les étapes 1 (recherche) et 2 (sélection) sont représentées par les niveaux V1 et V2, tandis que le niveau V3, est nécessaire seulement pour les requêtes de l'étape 3, portant sur les contenus des sources de données réelles.
Schématiquement, on considère:

En entrée :

(1) une application cible T, un schéma global Σ et une requête Q définie à partir de Σ. La requête Q est caractérisée par un triplet <O, G, D>, c.-à-d. Q porte sur un objet O, un territoire G, un temps (période, date) D (exemple : les personnes âgées dépendantes sur un territoire donné, une année donnée).
(2) un ensemble de critères Cr sur Q, avec leurs valeurs d'admissibilité V (exemple : fraîcheur sur les données)
(3) un ensemble X de catalogues C
- chaque catalogue est décrit par des métadonnées (exemple : objet, couverture, date de création, date de dernière mise à jour, …)
- chaque catalogue contient des sources de données S et chaque source de données est décrite par des métadonnées (exemple : objet, couverture, date création, date publication, …)

En sortie :

- des étapes 1 et 2:
  - un ensemble de sources de données S" appartenant aux catalogues C et concernant un (ou plusieurs) objet(s) O", un (ou plusieurs) territoire(s) G", une (ou plusieurs) période(s) D", avec O", G", D" en correspondance avec O, G, D.
  - une requête Q" <O", G", D">, évaluable sur Σ et vérifiant un ensemble de critères (Cr", V"), déduits des critères (Cr, V)
- de l'étape 3
  - résultat de la requête Q" <O", G", D">, sur les sources S" déterminées en sortie de l'étape 2.

Plus précisément :

Etape 1. Pour une cible donnée T (*target*), on note **rids(T)** = {S_1, S_2, ... S_m} l'ensemble des sources de données idéales nécessaires à T (*required ideal data sources*). L'étape 1 détermine l'ensemble des sources de données utilisables (*usable data sets*) **uds(T)** = {S'_1, S'_2, ... S'_k} où les sources *S'* vérifient les deux conditions :

- (1) il existe un catalogue C tel que S' ∈ C [1]
- (2) S' est en correspondance avec une source Si, i = 1, … m

Les correspondances sont définies selon la terminologie de [13].

Cette étape doit inclure un processus de spatialisation des objets et l'analyse de relations géométriques et topologiques. De plus, elle doit s'appuyer sur des services d'identification et de localisation des sources de données dans un ensemble de catalogues. Différentes stratégies de recherche des correspondances peuvent conduire à différents ensembles uds(T).

Etape 2. On note **Δ(T)** = **d(** rids(T), uds(T) **)** une fonction mesurant l'écart entre les sources de données requises et les sources de données utilisables par une cible T. L'étape 2 consiste à :

- évaluer la qualité des sources de uds(T), par rapport aux critères et limites dérivés à partir de ceux définis sur les sources de données idéales, rids(T),
- choisir un ensemble optimal uds(T), noté **ouds(T)**, qui minimise Δ(T). Des aspects organisationnels, conceptuels, syntaxiques, techniques, peuvent guider ce choix.

Il est très rare de trouver des correspondances exactes et totales entre données requises et données disponibles. La situation optimale est rids(T) ≡ ouds(T), i.e.: la formule [1] est vérifiée avec des correspondances qui sont des identités, tandis que la situation la plus défavorable a lieu si ouds(T) =∅. En pratique, l'étape 2 consiste à trouver le meilleur compromis entre les deux.

Etape 3. On effectue les requêtes sur les sources de ouds(T), si ouds(T)≠Ø.

4 Choix d'architecture

4.1 Couplage d'un système de raisonnement et d'un médiateur

Pour supporter les trois étapes de l'approche, on propose une architecture qui couple un système de raisonnement avec un système de médiation (figure 2).

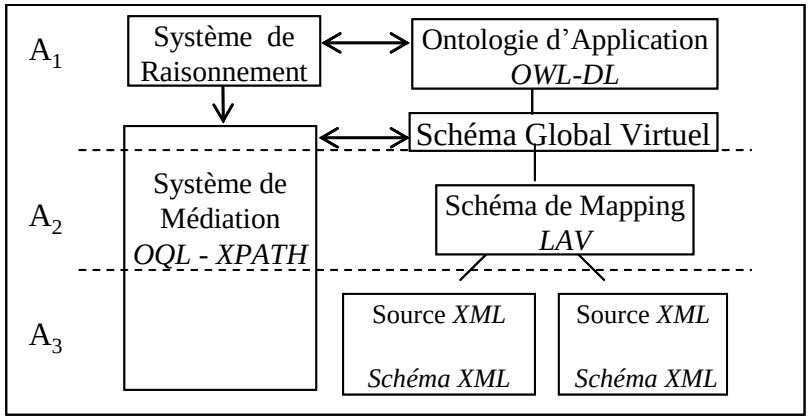

Figure 2. Architecture à trois niveaux

Le niveau global A1 représente le niveau applicatif, dans lequel sont formalisés les besoins pour la prise de décision. Ce niveau contient une ontologie d'application et un schéma global virtuel, à partir duquel les requêtes sont formulées.

Le système de raisonnement permet de déduire de nouvelles connaissances en s'appuyant sur les connaissances disponibles et sur les définitions de nouveaux concepts. Il travaille sur une ontologie d'application et poursuit deux objectifs :

R1 – déterminer les catalogues et sources répondant aux étapes 1 et 2.

R2 – qualifier les applications cible en fonction des catalogues et sources déterminées en (R1).

Le système de médiation comprend: (i) un schéma global, (ii) un ensemble de sources de données contenant des données réelles (sources locales), et (iii) un ensemble de relations entre le schéma global et les sources locales. Il facilite l'accès transparent des utilisateurs à des ressources hétérogènes et reparties. Il s'appuie sur la définition de vues, pour simuler un environnement global, centralisé et homogène, au travers duquel on interroge les sources de données locales. Notre approche généralise les approches de médiation basées sur les schémas [12]: en imposant des contraintes au niveau applicatif, en fonction des contextes et en s'appuyant sur les métadonnées, on peut réduire l'indétermination des interrogations de sources reparties et augmenter la cohérence et performance des interrogations. Dans ce domaine, la "personnalisation" de l'information, de nombreux verrous scientifiques et techniques sont encore à relever [7].

4.1 Description du niveau global

Ontologie d'application : elle peut être interprétée comme une spécialisation d'une ontologie de l'approche et d'une ontologie du domaine [17]. On démarre par un nombre limité de ceux-ci, extraits des deux premiers niveaux, V_1 et V_2, du modèle des classes. Pour représenter la qualité du processus décisionnel et la qualité requise pour les sources de données, on dérive ensuite des nouvelles classes et les classes complémentaires.

Ainsi, on peut définir les sources requises et contenues dans un catalogue comme sources de données "nécessaires et disponibles" **rads(T)** ("required and available data set"). Si de plus, ces sources vérifient les critères de qualité, elles sont alors des sources de données "qualifiées" **qds(T)** ("qualified data sets"). La vérification des critères de qualité s'appuie sur l'analyse des métadonnées décrivant les sources contenues dans les catalogues.

En utilisant ces classes, un système cible est dit "décrit" **DT** ("described target"), si toutes les sources demandées sont disponibles, et "bien décrit" **WDT** ("well described target"), si de plus, toutes ses sources vérifient les critères de qualité.
Les formules correspondantes à ces concepts sont :

$rads(T) = rids(T) \cap uds(T)$ [3]

$qds(T) = \{ds \mid ds \in rads(T)$ et ds satisfait les critères$\}$

$DT = \{T \mid rids(T) \subseteq rads(T)\}$

$WDT = \{T \mid rids(T) \subseteq qds(T)\}$

Elles introduisent des relations d'ordre parmi les classes Cible (Target) et Sources de données, aux deux niveaux d'existence et qualité: la position d'une cible dépend de la position de toutes les sources de données associées avec elle.
Ces relations d'ordre peuvent être affinées, par l'introduction de classes intermédiaires, basées sur les correspondances, les critères de qualité et des pourcentages.
Schéma Global : est donné par l'ontologie de domaine [18]. Celle-ci, développée indépendamment des sources de données, fournit une vue unifiée pour formuler des requêtes au niveau global. Dans notre cas, il s'agît d'un schéma orienté-objet, décrivant des concepts, munis d'attributs typés et reliés par des relations binaires. Ce modèle conceptuel, virtuel, peut être implémenté pour réaliser des vérifications syntaxiques et lexicales sur les requêtes globales. Un concept clé, comme le Territoire, est commun aux deux ontologies, d'application et de domaine.
Formalismes pour la représentation des connaissances
Les règles sont utilisées pour répondre à l'objectif R1 du système de raisonnement.
Le formalisme des Logiques de Description (LD) [6] est utilisé pour répondre à l'objectif R2 du système de raisonnement**.**

5 Choix techniques

Les choix techniques concernent l'implémentation des expressions formulées au dessus de l'ontologie d'application, en logique de description et par le langage des règles [2], ainsi que le système de médiation.
En correspondance des LD
L'ontologie d'application est implémentée avec OWL DL [16], un sous langage de OWL basé sur $\mathcal{SHOIN(D)}$, auquel on peut associer des services de raisonnement, car il est décidable.
En effets, pour formuler les concepts associés à l'ontologie d'application, l'expressivité de LD requise est $\mathcal{ALCOIN(D)}$. $\mathcal{N}$ est nécessaire pour formuler des expressions $\mathcal{CWA}$, avec des constructeurs qui limitent les cardinalités. Or, $\mathcal{ALCOIN(D)}$ est un sous-ensemble de $\mathcal{SHOIN(D)}$ [3], et pas de $\mathcal{SHIF(D)}$, auquel, dans la famille des langages OWL, est associé le sous langage OWL Light.
Le code OWL-DL peut être obtenu soit en appliquant les correspondances entre les syntaxes des constructeurs LD et d'OWL DL, soit par des outils graphiques comme Protégé.
On utilise l'éditeur d'ontologies Protégé [11] pour définir la base de connaissances et Pellet [15] comme système de raisonnement.
Protégé est un environnement de développement open source pour construire des ontologies décrites en OWL DL et des systèmes à base de connaissances, supportant $\mathcal{SHOIN(D)}$. Développé à la Stanford University, il possède une architecture très extensible et peut être utilisée en conjonction avec des systèmes de raisonnement, au travers d'une interface standardisée, développée par le DL Implementation Group (DIG).

Pellet est un système d'inférence open-source capable de vérifier la consistance d'une ontologie ainsi que de classifier automatiquement ses concepts et instances. Il implémente un algorithme de tableau optimisé pour la LD $\mathcal{SROIQ}(\mathcal{D})$.

5.1 En correspondance des règles

Pour les formules non exprimables en OWL DL, on a besoin d'utiliser un langage de règle afin de formuler ces connaissances.
Plusieurs langages de règles ont été proposés, cependant le seul langage basé sur OWL est SWRL[1] (Semantic Web Rule Language). SWRL a été soumis comme proposition au W3C en 2004 combinant le langage RuleML avec OWL. SWRL est considéré comme une approche homogène combinant OWL avec les règles. Néanmoins, il n'existe pas, pour l'instant, un raisonneur capable de supporter et de raisonner simultanément sur les deux formats. La solution est d'utiliser à la fois un raisonneur LD et un système de règle avec exécution interactive. Le système de règle choisi est Jess, un système performant, pour lequel on dispose d'outils adéquats pour l'environnement de développement Protégé. SwrlJessTab est un plug-in qui permet de formuler des règles en SWRL et de réaliser un pont entre l'ontologie OWL+SWRL vers Jess et vice versa.

5.2 En correspondance de la technique d'intégration

Les principales approches, pour relier le schéma global aux sources locales, sont : *Global As View* "GAV" et *Local As View* "LAV" [12].
Notre vue intégrée suit l'approche LAV, qui décrit les sources locales comme des vues à partir du schéma global. La priorité est donnée à la construction de ce schéma pour prendre en compte des besoins et contraintes exprimés au niveau global. Dans notre contexte, ceux-ci concernent les concepts, objets et relations ainsi que des niveaux de qualité nécessaires pour supporter la prise de décision.
Une approche GAV, au contraire, aurait déterminé le schéma global et l'expression des besoins seulement à partir des schémas des sources de données disponibles. Ceci contredit la priorité que nous assignons à l'expression des besoins au niveau global, afin de garantir les évolutions des contextes (internationalisation, pluridisciplinarité, avancée des connaissances imposant des nouveaux critères de qualité, …). Ainsi, l'approche LAV a été préférée à l'approche GAV, même si celle-ci apparaît plus naturelle et sûrement plus simple à implémenter.
Le niveau de médiation A2 de la figure 2 est décrit en suivant les principes d'une technique d'intégration [1]. Un ensemble de règles ('mapping rules') relient le niveau global et le niveau local. Ces règles expriment des correspondances entre les chemins conceptuels du niveau global et des chemins dans les schémas des sources locales.
Les requêtes sont formulées sur le schéma global dans une variante du langage OQL. Si la totalité de l'information recherchée ne peut pas être obtenue à partir d'une seule source, alors la requête est décomposée en un ensemble de sous requêtes 'locales'. Chaque sous requête est exécutée par un système local, pour fournir des résultats partiels, fusionnés ensuite.
Le niveau local A3 de la figure 2 comprend les sources de données locales, stockées dans des catalogues et décrites par leurs schémas. Les schémas sont complétés par un ensemble opportun de métadonnées, correspondant aux critères de qualité (étape 2). Seulement les sources de données répondant directement ou indirectement aux requêtes globales, sont retenues. On ignore les autres sources de données, qui ne répondent pas aux buts de la cible. La section suivante décrit ce processus.

[1] http://www.w3.org/Submission/SWRL/

Au niveau local, on utilise (i) XML Schéma pour représenter les schémas des sources de données, incluant les métadonnées et (ii) XQuery pour exécuter les requêtes sur les sources locales.

6 Illustration de l'approche sur un exemple

Requêtes Type

Des exemples de requêtes pour évaluer l'existence et la qualité de sources de données, couvrant les territoires concernés, sont :

Etape 1 - Q1: Quelles sont les cibles décrites sur un territoire géographique ?

Etape 2 - Q2: Quelles sont les cibles bien décrites sur ce territoire ?

Dans [10] on détaille le traitement de l'étape 3 (requêtes sur le contenu), comme par exemple "Quel est le nombre de personnes âgées dépendantes, dans les départements inclus dans un territoire géographique donné ?".

Ontologie d'application

La figure 3 présente le modèle UML de l'ontologie proposée au niveau application. Pour certains rôles, on utilise des éléments de métadonnées définis par le vocabulaire Dublin Core. Ce modèle est approprié dans les contextes où les sources sont clairement identifiées, comme dans des plans nationaux d'urgences climatiques [14].

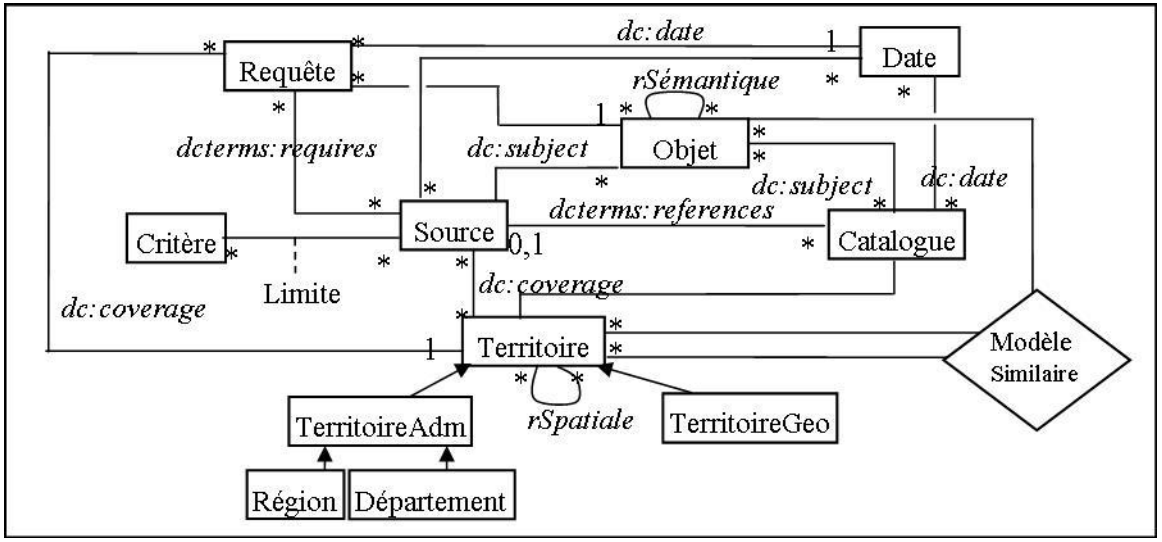

Figure 3 : Ontologie (Diagramme UML)

Données

Les données utilisées vérifient les assertions de la Table 1. Elles concernent des cibles concernant des risques (canicule, vague de froid, …), des ressources (hôpitaux, …), des personnes vulnérables. Ces dernières données sont renseignées sur des territoires administratifs.

Table 1. Données utilisées dans l'exemple

rids(T_1) ≡ {S_1, S_2, S_3, S_4, S_5} rids(T_i) est défini par [1], T_i application cible
rids(T_2) ≡ {S_1, S_2, S_5, S_6, S_7} S_i source de données
rids(T_3) ≡ {S_1, S_2, S_5, S_8}
contient {(TG_1, {TA_1, TA_2}), (TG_2 ,{TA_3, TA_4}) },
avec TG_i territoire géographique, TA_i territoire administratif
contient (P_1, {TA_1, TA_2, TA_3, TA_4}) avec P_i pays
appartient { ({S_1, S_2, S_5, S_8}, C_1), ({S_3, S_4, S_6, S_7}, C_2) }, avec C_i Catalogue,
C_1 de niveau national sur P_1, C_2 de niveau départmental
couvre { (S_1, P_1), (S_2, P_1), (S_5, P_1), (S_8, P_1), (S_3, TA_1), (S_4, TA_2), (S_6, TA_3), (S_7, TA_4) }
F (S_1, 2Y)… F (S_4, 6M) … (F, Y et M indiquant respectivement : critère de Fraicheur, Année et Mois)

Déroulement de l'approche

Définition des concepts

Etape 1 - Q1 :

SourceDisponible ≡ Source ⊓ ∃ appartient.Catalogue

SourceManquante ≡ Source ⊓ ¬ SourceDisponible

TerritoireTG ≡ Territoire ⊓ ∃ contient⁻.{TG}

SourceTG ≡ Source ⊓ ∃ couvre.TerritoireTG

SourceDisponibleTG ≡ SourceDisponible ⊓ SourceTG

CibleDécriteTG ≡ Cible ⊓ ∃ gère.(Risque ⊓ ∃ concerne.{TG}) ⊓ ∀ nécessite.(SourceDisponibleTG ⊔ ¬ SourceTG)

Etape 2 - Q2 :

SourceQualifiée ≡ SourceDisponible ⊓ ∀ vérifie.LimiteCritèreRespecté

SourceNonQualifiée ≡ SourceDisponible ⊓ ¬ SourceQualifiée

SourceQualifiéeTG ≡ SourceQualifiée ⊓ SourceTG

$CibleBienDécriteTG \equiv CibleDécriteTG \sqcap \forall nécessite.(SourceQualifiéeTG \sqcup \neg SourceTG)$

Métadonnées

Au niveau des catalogues, les métadonnées sont décrites en OWL DL. Un fragment de cette description est donné ci-dessous :

```
<owl:Ontology rdf:about=""/>
<owl:Class rdf:ID="Catalogue"/>
<owl:Class rdf:ID="Territoire"/>
<owl:Class rdf:ID="Source"/>
<owl:ObjectProperty rdf:ID="contient">
  <rdfs:domain rdf:resource="#Catalogue"/>
  <rdfs:range rdf:resource="#Source"/>
</owl:ObjectProperty>
<owl:ObjectProperty rdf:ID="couvre">
  <rdfs:domain rdf:resource="#Source"/>
  <rdfs:range rdf:resource="#Territoire"/>
</owl:ObjectProperty>
<owl:DatatypeProperty rdf:ID="DatePub">
  <rdfs:domain rdf:resource="#Source"/>
</owl:DatatypeProperty>
…
```

Résultats

Etape 1 - Q1: $CibleDécriteTG_1 = \{T_1\}$, car toutes les sources de rids(T_1) sont disponibles sur tous les territoires administratifs TA contenus dans TG_1.

Etape 2 - Q2: $CibleBienDécriteTG_1 = \{T_3\}$, car toutes les sources de rids(T_3) sont disponibles et vérifient les critères de qualité, sur tous les territoires TA dans TG_1. T_1 n'appartient pas à cette classe, car la source S_4 ne vérifie pas le critère de *fraîcheur* imposé.

D'autres concepts, basés sur *SourceManquante* et *SourceNonQualifiée*, informent sur les sources demandées et non disponibles ou de qualité insuffisante. On peut donc déclencher la recherche de sources alternatives, moins bien qualifiées, sur ce territoire.

7 Conclusion et perspectives

Les volumes et l'hétérogénéité des sources de données sont des freins considérables à l'intégration de données. Cette intégration est un principe structurant des systèmes d'information à large échelle (e-gouvernement, veille stratégique territoriale, production de services).

Dans ce papier nous avons présenté les lignes générales d'une approche de recherche, sélection et interrogation de sources de données distribuées et hétérogènes. Cette approche vise à répondre à des besoins décisionnels. Elle est structurée en trois niveaux. Nous avons proposé une architecture de médiation pour sa mise en œuvre et illustré cette démarche par un exemple simple, pris dans le domaine de la santé.

En perspective, nous souhaitons affiner cette démarche, en développant deux aspects pour :

- augmenter la traçabilité des choix des sources dans les catalogues
- expliciter les modifications opérées sur les requêtes et sur les sources

Nous préconisons d'introduire des classifications plus fines des catalogues et des sources disponibles, en fonction des requêtes formulées dans les différents contextes d'application. Ces classifications ont pour objectif d'accroître la base de connaissances et de participer à l'évaluation du processus global.

8 Références

[1] AMANN, B., BEERI, C., FUNDULAKI, I. et SCHOLL, M., *Ontology-based integration of xml web resources*, Lecture Notes in Computer Science, vol. 2342, 2002, p 117-131

[2] ANTONIOU, G., DAMÁSIO, C.V., GROSOF, B., HORROCKS, I., KIFER, M., MALUSZYNKI, J. et PATEL-SCNEIDER P.F., *Combining Rules and Ontologies. A survey*, REWERSE, 2005, http://rewerse.net/deliverables/m12/i3-d3.pdf

[3] BAADER, F., HORROCKS, I. et SATTLER, U., *Description Logics as Ontology Languages for the Semantic Web*, Lecture Notes in Artificial Intelligence, vol. 2605, 2005, p 228–248

[4] BÉDARD, Y., DEVILLERS R., GERVAIS M. et JEANSOULIN R., *Towards Multidimensional User Manuals for Geospatial Datasets: Legal issues and their Considerations into the design of a Technological Perspective*. Proceedings of the Third International Symposium on Spatial Data Quality (ISSDQ'04), Bruck an der Leitha, Autriche, 15-17 avril, vol. 28b GeoInfo Series, 2004, p 183-195

[5] BIANCHINIA, D., DE ANTONELLIS, V., PERNICI, B. et PLEBANI, P., *Ontology-based methodology for e-service discovery*, Information Systems 31, 2006, p 361–380

[6] CALVANESE, D., MCGUINNESS, D., NARDI, D., PATEL-SCHNEIDER, P., *The Description Logic Handbook: Theory, Implementation and Applications*, UK, Cambridge Univ. Press, 2004

[7] CNRS AS 97 du département STIC du CNRS, *Médiation via les métadonnées*, 2003, http://www.lirmm.fr/~libourel/MM/MetaMedia.htm

[8] DESCONNETS, J.C., MOYROUD, N. et LIBOUREL, T., *Méthodologie de mise en place d'observatoires virtuels via les métadonnées*, InforSid, Nancy, Juin 2003. Voir démo Mdweb : http://www.mdweb-project.org

[9] DEVILLERS, R. et JEANSOULIN, R. (Eds.), *Qualité de l'information géographique*. Hermès Science, 2005

[10] GUEMEIDA, A., JEANSOULIN, R. et SALZANO, G., «Quality-aware and Metadata-based Interoperability for Environmental Health Information», In Proc. of the 5th International Symposium on Spatial Data Quality ISSDQ'07, Enschede, Netherlands, 13-15 juin 2007

[11] KNUBLAUCH, H., FERGERSON, R.W., NOY, N.F. et MUSEN, M.A, *The Protégé OWL Plugin: An Open Development Environment for Semantic Web Applications*, Lecture Notes in Computer Science, vol. 3298, 2004, p 229-243

[12] LENZERINI, M., *Data integration: A theoretical perspective*, In Proc. of 21st ACM SIGACT-SIGMOD-SIGART Symposium on Principles of Database Systems PODS, Madison, Wisconsin, 03-06 juin 2002, New York, ACM Press, p 233–246

[13] PARENT, C. et SPACCAPIETRA, S., *Advances in Object-Oriented Data Modeling in Database Integration: The Key to Data Interoperability*, p 221-254, Cambridge, The MIT Press, 2000

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

[15] SIRIN, E., PARSIA, B., CUENCA GRAU, B., KALYANPUR, A. et KATZ, Y., *Pellet: A practical OWL-DL reasoner*, Journal of Web Semantics, 2006

[16] SMITH, M.K., WELTY, C. et MCGUINNESS, D., *OWL Web Ontology Language Guide*, Recommendation W3C, 2004, http://www.w3.org/TR/owl-guide

[17] VAN HEIJST, G., SCHREIBER, A. TH. et WIELINGA, B.J., *Using explicit ontologies in KBS development*, International Journal of Human-Computer Studies, 46(2-3), Feb. 1997, p 183-292

[18] VISSER, U., *Intelligent Information Integration for the Semantic Web*, Lecture Notes in Artificial Intelligence, vol. 3159, 2004, p 13-34