

Université de Paris-Est

**THÈSE**

pour obtenir le grade de  
Docteur de l'Université de Paris-Est

Champ disciplinaire : Systèmes d'Information

présentée et soutenue publiquement par

**Abdelbasset GUEMEIDA**

le 16 octobre 2009

**Contributions à une nouvelle approche  
de Recherche d'Information**

basée sur la métaphore de l'impédance et illustrée sur le domaine de la santé

**Directrice de thèse : Mme Gabriella SALZANO**

**Codirecteur : Mr Olivier CURÉ**

**Jury**

<b>Mr Bernard DOUSSET</b>	Professeur U. Paul Sabatier Toulouse (Rapporteur)
<b>Mr Michel LAMURE</b>	Professeur U. Lyon 1 (Rapporteur)
<b>Mme Myriam LAMOLLE</b>	Maître de Conférences U. Paris 8 - IUT de Montreuil (Examinatrice)
<b>Mr Jean Paul RUDANT</b>	Professeur U. Paris Est (Examineur)
<b>Mr Christian BOURRET</b>	Maître de Conférences U. Paris Est (Invité)
<b>Mme Gabriella SALZANO</b>	Maître de Conférences HDR, U. Paris Est (Directrice de thèse)



*Je dédie cette thèse  
à l'âme de mon père,  
à ma mère,  
à toute ma famille,  
à mon épouse  
et à tous mes amis*



# Avant Propos

## *Structure du document*

Ce mémoire est réalisé par l'insertion d'articles et de rapports de recherche, copubliés avec R. Jeansoulin, Directeur de Recherche aux CNRS et membre du laboratoire LabInfo de l'Université Paris Est et avec Gabriella Salzano, ma directrice de thèse. Plus précisément :

- Le Chapitre 1 correspond à un rapport de recherche<sup>1</sup> qui présente « NARI, une Nouvelle Approche de Recherche d'Information basée sur la métaphore de l'impédance, la qualité et les métadonnées, structurée par la dimension géographique et illustrée sur le domaine de la santé ».
- Le Chapitre 2 correspond à un rapport de recherche<sup>2</sup> « Technologies du Web sémantique et outils associés : état de l'art et choix, pour NARI ».
- Les Chapitres 3, 4 et 5 correspondent à des articles de recherche, publiés dans des actes de congrès :
  - Chapitre 3<sup>3</sup> : « Interopérabilité entre systèmes d'information géographiques et de santé : une approche basée sur les métadonnées », Proceedings of Géomatique'2006, Montréal, Canada, 25-26 octobre 2006
  - Chapitre 4<sup>4</sup> : « Qualité de l'information géographique : Approche basée sur les besoins préliminaires », Actes du Colloque International de Géomatique et d'Analyse Spatiale, SAGEO'2007, Clermont-Ferrand, 18, 19 Juin 2007
  - Chapitre 5<sup>5</sup> : « Quality-aware Agents for e-Government Information Systems Architecture », Proceedings of 3rd International Conference on e-Government, University of Quebec at Montreal, Canada on the 27-28 September 2007

---

<sup>1</sup> Auteurs : G. Salzano et A. Guemeida

<sup>2</sup> Auteur : A. Guemeida

<sup>3</sup> Auteurs : G. Salzano et A. Guemeida

<sup>4</sup> Auteurs : A. Guemeida, R. Jeansoulin, G. Salzano

*Mes contributions spécifiques à ces travaux*

Tout en ayant contribué globalement à l'ensemble de cette production scientifique, j'ai focalisé mes recherches sur des aspects de la conception de NARI concernant notamment l'analyse et le choix des approches d'intégration de données et des langages de modélisation des connaissances, la conception de l'ontologie d'application et de l'architecture logicielle globale, le choix des outils et l'implémentation sur des exemples.

---

<sup>5</sup> Auteurs : A. Guemeida, R. Jeansoulin, G. Salzano

## Remerciements

Je tiens à remercier sincèrement toutes les personnes avec lesquelles j'ai travaillé et accompli les travaux présentés dans ce mémoire.

Je dis un GRAND MERCI

à *Gabriella Salzano*, pour son encadrement, ses conseils et son implication personnelle tout au long de mon parcours,

à *Olivier Curé*, pour son co-encadrement qui m'a beaucoup apporté,

à *Robert Jeansoulin*, qui a été à l'origine de la métaphore de l'impédance, développée dans les travaux de recherche conjoints,

à *Bernard Dousset et Michel Lamure*, pour m'avoir fait l'honneur d'accepter le travail de rapporteurs en dédiant une partie de leur temps à mes activités,

à *Myriam Lamolle et Jean Paul Rudant*, pour leur disponibilité à accepter de participer au jury,

à *Robert Eymard*, pour son soutien et ses orientations,

à *Christian Bourret*, qui m'a accueilli à l'IFIS (Institut Francilien d'Ingénierie des Services).

Je tiens aussi à exprimer ma gratitude

*aux doctorants du laboratoire S3IS* pour leur aide et encouragements

*à tous les collègues et tout le personnel* de l'université Paris-Est, que serait trop long de remercier nominativement...

Sur le plan plus personnel, un Merci, qui ne pourrait pas être plus grand, à *Messaoud, Aziz, Abdelghani et Malik*.



## Résumé

Les récentes évolutions dans les technologies de l'information et de la communication, avec le développement de l'Internet, conduisent à l'explosion des volumes des sources de données. Des nouveaux besoins en recherche d'information émergent pour traiter l'information en relation aux contextes d'utilisation, augmenter la pertinence des réponses et l'usabilité des résultats produits, ainsi que les possibles corrélations entre sources de données, en rendant transparentes leurs hétérogénéités.

Les travaux de recherche présentés dans ce mémoire apportent des contributions à la conception d'une Nouvelle Approche de Recherche d'Information (NARI) pour la prise de décision. NARI vise à opérer sur des grandes masses de données cataloguées, hétérogènes, qui peuvent être géo référencées. Elle est basée sur des exigences préliminaires de qualité (standardisation, réglementations), exprimées par les utilisateurs, représentées et gérées à l'aide des métadonnées. Ces exigences conduisent à pallier le manque de données ou leur insuffisante qualité, pour produire une information de qualité suffisante par rapport aux besoins décisionnels. En utilisant la perspective des utilisateurs, on identifie et/ou on prépare des sources de données, avant de procéder à l'étape d'intégration des contenus.

L'originalité de NARI réside dans la métaphore de l'écart d'impédance (phénomène classique lorsque on cherche à connecter deux systèmes physiques hétérogènes). Cette métaphore, dont R. Jeansoulin est à l'origine, ainsi que l'attention portée au cadre réglementaire, en guident la conception. NARI est structurée par la dimension géographique (prise en compte de divers niveaux de territoires, corrélations entre plusieurs thématiques) : des techniques d'analyse spatiale supportent des tâches de la recherche d'information, réalisées souvent implicitement par les décideurs. Elle s'appuie sur des techniques d'intégration de données (médiation, entrepôts de données), des langages de représentation des connaissances et des technologies et outils relevant du Web sémantique, pour supporter la montée en charge, la généralisation et la robustesse théorique de l'approche. NARI est illustrée sur des exemples relevant de la santé.

**Mots clés :** Recherche d'Information, impédance, qualité des données, besoins préliminaires, métadonnées, information géographique, standardisation, applications en santé.

## Abstract

The recent developments in information and communication technologies along with the growth of the Internet have led to the explosion of data source volumes. This has created many growing needs such as in information retrieval to: treat the information according to its usage context, to increase the relevance of answers and the usability of results, and to increase the potential correlations between results, which can be done by making the heterogeneities and source distribution transparent.

Our contributions consist in designing a NARI (New Approach to Information Retrieval) for decision-making. NARI is designed to operate on large amounts of catalogued and heterogeneous data that can be geo-referenced. It is based on quality preliminary requirements expressed by users, which are represented and managed using metadata. These requirements lead to the lack of data or their insufficient quality to produce information with a sufficient quality in relation to decision-making needs. Using the users' perspective, we identify and/or prepare the data sources, before integration step processing.

NARI's originality relies on the metaphor of the impedance mismatch (classical phenomenon when we try to connect two physical heterogeneous systems), due to R. Jeansoulin. This metaphor, as well as the attention paid to regulatory framework (standardization), guides the design of NARI.

The geographical dimension structures NARI, taking into account various territorial levels, correlations between several themes. Thus, it takes advantage of spatial analysis techniques, by automating information retrieval tasks, often implicitly made by policy makers. NARI is based on data integration techniques (mediation, data warehouses), knowledge representation languages and a set of Semantic Web technologies and tools, adapted to support the scalability, robustness and generalization theory of the approach. NARI is illustrated on examples relevant to the health domain.

**Keywords:** Information Retrieval, impedance, data quality, early requirements, metadata, geographic information, standardization, health applications.

# Table des matières

<b>Introduction générale.....</b>	<b>1</b>
1. Le contexte des recherches.....	3
2. Les travaux de recherche présentés dans ce mémoire.....	3
3. Illustrations sur le domaine de la santé .....	4
4. Organisation du mémoire .....	5
<b>Chapitre 1 : Une Nouvelle Approche de Recherche d'Information .....</b>	<b>7</b>
1. Introduction .....	9
2. Contexte et objectifs généraux de ces recherches .....	10
3. Choix d'un champ disciplinaire d'illustration : la santé.....	16
4. Etat de l'art.....	21
5. L'approche NARI : synthèse, analyse et discussion.....	44
6. Perspectives de recherche et Conclusions .....	60
7. Références .....	65
<b>Chapitre 2 : Technologies du web sémantique et outils associés : état de l'art et choix pour NARI .....</b>	<b>73</b>
1. Introduction .....	75
2. Objectifs technologiques de NARI .....	76
3. Présentation des technologies du web sémantique.....	77
4. Evaluation des technologies par rapport aux objectifs.....	95
5. Conclusions .....	101
6. Références .....	102
<b>Chapitre 3 : Interopérabilité entre systèmes d'information géographiques et de santé : une approche basée sur les métadonnées .....</b>	<b>107</b>
1. Introduction .....	109
2. Démarche systémique, développement durable et interopérabilité.....	110
3. Facteurs de complexité de l'interopérabilité de S.I. coopératifs à large échelle.....	112
4. Une approche d'interopérabilité basée sur les métadonnées .....	114
5. Infrastructure technologique .....	117
6. Illustration de l'approche .....	123

7. Conclusions .....	126
8. Bibliographie .....	127

**Chapitre 4 : Qualité de l'information géographique : approche basée sur les besoins préliminaires .....** **129**

1. Introduction .....	131
2. Interopérabilité de systèmes d'information et impédance .....	132
3. Une approche en trois étapes.....	136
4. Une vue intégrée en trois étapes, guidée par les besoins .....	139
5. Application de l'approche à un exemple .....	146
6. Conclusions .....	150
7. Références .....	150

**Chapitre 5: Quality-aware Agents for e-Government Information Systems Architecture .....** **153**

1. Introduction .....	155
2. Critical requirements for G2G systems in health emergencies .....	156
3. The Integration of Information and the Impedance Metaphor .....	158
4. Ontology and sequence structure for information integration .....	160
5. Cooperative multi-agent patterns .....	162
6. Global architecture .....	164
7. Illustration .....	169
8. Conclusions .....	171
9. References .....	172

# Liste des tables

## Chapitre 1

<b>Table 1 :</b> Exemples de problématiques décisionnelles en santé.....	19
<b>Table 2 :</b> Objectifs spécifiques pour la conception de NARI.....	45
<b>Table 3 :</b> Association entre les objectifs spécifiques de l'approche NARI et les aspects traités dans les articles.....	46
<b>Table 4 :</b> Exemples de décomposition et transformation d'une requête.....	54
<b>Table 5 :</b> Exemples de classifications de requêtes .....	55

## Chapitre 2

<b>Table 1 :</b> Quelques constructeurs des LD.....	81
<b>Table 2 :</b> Dénotation de l'expressivité des LD .....	81
<b>Table 3 :</b> Correspondances entre les LD et OWL .....	82

## Chapitre 3

<b>Table 1 :</b> Correspondances entre risques et sources de données .....	124
<b>Table 2 :</b> Territoires géographiques et départements .....	124
<b>Table 3 :</b> Catalogues et sources .....	124
<b>Table 4 :</b> Sources de données .....	124
<b>Table 5 :</b> Critères et seuils .....	124

## Chapitre 4

<b>Table 1 :</b> Expressivité des Logiques de Description .....	144
<b>Table 2 :</b> Jeu de données.....	149

## Chapitre 5

<b>Table 1:</b> Query-Answering System steps on a simple example.....	158
<b>Table 2:</b> Examples of treatments specifications .....	165
<b>Table 3:</b> Query classifications, expressed with rules.....	166
<b>Table 4:</b> Query classifications, expressed with DL.....	166

# Liste des figures

## Chapitre 1

<b>Figure 1</b> : Schématisation des objectifs de l'approche NARI.....	15
<b>Figure 2</b> : Roue de Deming .....	30
<b>Figure 3</b> : Médiation de bases de données hétérogènes .....	36
<b>Figure 4</b> : Architecture d'un entrepôt de données.....	37
<b>Figure 5</b> : Cadre de référence DWQ des métadonnées d'entrepôt de données.....	39
<b>Figure 6</b> : Vue d'ensemble de l'architecture technique INSPIRE.....	43
<b>Figure 7</b> : Exemples d'hétérogénéité affectant les données de santé et les données géographiques.....	48
<b>Figure 8</b> : Schéma de l'évolution des objectifs de recherche.....	59

## Chapitre 2

<b>Figure 1</b> : Première version du Semantic Web Layer Cake .....	77
<b>Figure 2</b> : Les deux approches combinant les ontologies et les règles.....	87
<b>Figure 3</b> : Définition du langage DLP comme intersection de LD et PL.....	88
<b>Figure 4</b> : Infrastructure technologique à trois niveaux .....	97

## Chapitre 3

<b>Figure 1</b> : Diagramme UML du domaine applicatif.....	116
<b>Figure 2</b> : Infrastructure technologique à trois niveaux .....	117
<b>Figure 3</b> : Ontologie au niveau application .....	119

## Chapitre 4

<b>Figure 1</b> : Les principaux lieux de confrontation d'impédance.....	136
<b>Figure 2</b> : Le modèle de référence d'INSPIRE.....	139
<b>Figure 3</b> : Vue intégrée des aspects d'existence, qualité et contenu des données.....	140
<b>Figure 4</b> : Architecture à trois niveaux.....	141
<b>Figure 5</b> : Modèle UML de l'ontologie d'application .....	147

**Chapitre 5**

**Figure 1:** Organizational model for a Query-Answering System in an e-Government context ..... 157

**Figure 2:** Application Ontology for a Query-Answering system..... 161

**Figure 3:** Sources Searcher composition..... 163

**Figure 4:** Sources Searcher along the theme dimension. .... 164

**Figure 5:** QAS global architecture ..... 164



# **Introduction générale**



## 1. Le contexte des recherches

Les énormes progrès réalisés depuis quelques années dans les technologies de l'information et de la communication conduisent à l'émergence d'organisations en réseau et de nouvelles applications qui requièrent l'échange, voir le partage, de données issues de multiples systèmes, conçus a priori de façon autonome et pour des objectifs spécifiques. Dans les cas les plus complexes, comme en épidémiologie (grippe aviaire, virus H1N1, ..., par exemple), un pilotage international de ces applications est nécessaire. Ces changements technologiques et organisationnels s'accompagnent de la construction progressive de cadres réglementaires complexes, auxquels participent les normes et les standards, informatiques et sectoriels. Ainsi par exemple, pour relier les données environnementales à des données d'une autre nature (administratives, de santé, économiques, sociales, ...) on peut désormais s'appuyer sur la Directive européenne INSPIRE (INSPIRE, 2008). Cette directive concerne la diffusion de données et de services géographiques et a pour objectif de rendre accessibles toutes les données géographiques publiques, à un coût minimal ou nul. Elle contribue à interpréter l'information comme un "bien public" et à impulser de nouveaux rapports entre l'offre de données et la demande d'informations, voire de connaissances.

Les nouveaux systèmes de Recherche d'Information, caractérisés par des moyens puissants (moteurs de recherche, interfaces conviviales, graphiques, "géographiques"), doivent répondre à des usages de plus en plus diversifiés, aussi bien dans des contextes collectifs d'entreprise qu'individuels. Des nouvelles problématiques sont liées à l'explosion des volumes des sources de données : (i) traiter l'information en relation au contexte d'utilisation, en filtrant les sources, pour augmenter la pertinence des réponses, et en adaptant les sorties pour augmenter l'usabilité des résultats produits ; (ii) augmenter les corrélations possibles, en rendant transparentes les hétérogénéités et la distribution des sources.

## 2. Les travaux de recherche présentés dans ce mémoire

Les travaux de recherche présentés dans ce mémoire apportent des contributions à la conception d'une Nouvelle Approche de Recherche d'Information (NARI) pour la prise de décision. NARI opère sur des grandes masses de données cataloguées, hétérogènes, qui peuvent être géo référencées. Elle est basée sur des exigences préliminaires de qualité, liées à des obligations réglementaires et aux processus de standardisation, sectoriels et informatiques.

Ces exigences sont exprimées par les utilisateurs et représentées et gérées à l'aide des métadonnées. Elles conduisent à adapter le choix des données en fonction du type de décision et à pallier le manque de données ou leur insuffisante qualité, pour produire une information de qualité suffisante par rapport aux besoins décisionnels (Décideur public, Directeur Régional des Affaires Sociales, Directeur d'hôpital, médecin prescripteur, ...). Ces besoins sont déterminés par les cadres réglementaires dans lesquels les utilisateurs réalisent leurs recherches d'information.

L'originalité de NARI réside dans la métaphore de l'écart d'impédance. L'écart d'impédance est un phénomène classique lorsqu'on cherche à connecter deux systèmes physiques hétérogènes. La métaphore de l'impédance est due à R. Jeansoulin et développée la première fois dans l'article conjoint qui fait l'objet du Chapitre 4. En établissant des analogies entre les systèmes physiques et les systèmes d'information, elle conduit à la reformulation des objectifs d'une recherche d'information en cas d'échec et à une démarche (NARI) qui vise à concilier d'une part les ressources informationnelles disponibles et d'autre part la perspective des utilisateurs. NARI s'attaque prioritairement à l'activité d'identifier et/ou préparer des sources de données à intégrer. On réalise des analyses qualitatives des problèmes d'hétérogénéité et de qualité de données et on apporte des éléments préliminaires de conception de modules logiciels de pre-intégration, qui sont illustrés sur des exemples.

NARI est structurée par la dimension géographique, qui permet de prendre en compte divers niveaux de territoires, de corréler plusieurs thématiques pour la prise de décision. Ainsi NARI tire profit des techniques d'analyse spatiale, en automatisant des tâches du processus d'aide à la décision qui sont souvent réalisées implicitement par les décideurs.

NARI s'appuie sur des techniques d'intégration de données, des langages de représentation des connaissances et des technologies et outils relevant du Web sémantique, adaptés à supporter la montée en charge, la généralisation et la robustesse théorique de l'approche.

### 3. Illustrations sur le domaine de la santé

L'analyse des exigences et la conception de NARI sont illustrées à l'aide d'exemples du domaine de la santé. L'analyse de ce domaine est profitable à la conception, car elle structure les besoins de la recherche d'information :

1. *générer de l'information* via la construction de documents à partir de l'intégration de sources de données hétérogènes ou via la présentation de données extraites d'un entrepôt de données
2. *générer de la connaissance*, dans un but d'aide à la décision (planification de ressources, optimisation de coûts, maximisation du service rendu).
3. *évaluer les pratiques et les politiques de santé*, afin d'améliorer globalement les systèmes de santé.

De plus, qu'il s'agisse d'études d'épidémiologie, d'analyse d'accès aux soins, de planification sanitaire, de gestion de risques ayant un impact sur la santé, on est confronté aux problèmes de volumes et d'hétérogénéité des sources de données disponibles, d'autant plus que les données doivent être couplées avec des données géographiques ou environnementales et des données concernant les populations.

Les spécificités du domaine sont liées premièrement au poids économique du secteur (en France, par exemple, de l'ordre de 10% du PIB), au cadre réglementaire, aux diverses couvertures géographiques des structures de soins et aux nombreuses contraintes, notamment liées à l'identification des patients et au secret médical.

## 4. Organisation du mémoire

Ce mémoire est organisé de la façon suivante :

- Le Chapitre 1 décrit le contexte global des recherches et leurs objectifs, ainsi que le champ disciplinaire d'illustration : la santé. Il présente un état de l'art succinct et la démarche de conception, avant d'apporter une description synthétique de NARI. Celle-ci sert de base à une discussion sur le positionnement, l'originalité et les limites actuelles de NARI, pour terminer avec des perspectives de recherche et des conclusions sur les contributions de ces recherches.
- Le Chapitre 2 focalise sur les aspects technologiques et d'implémentation de NARI. Après avoir défini des objectifs technologiques en liaison avec les objectifs généraux introduits dans le Chapitre 1, il analyse des technologies du web sémantique, avec leurs rôles, avantages et limites et il évalue ensuite ces technologies et les outils associés. Il présente ensuite l'architecture globale retenue pour NARI.

- Les Chapitres 3, 4 et 5 détaillent des aspects spécifiques de NARI :
  - Le Chapitre 3 analyse des Systèmes d'Information coopératifs en santé, reliés avec des SIG, et aboutit au choix d'une technique d'intégration de données de type LAV (Local As View) basée sur les métadonnées. Celles-ci permettent d'identifier et d'évaluer les sources d'informations disponibles pour la recherche d'information, en fonction de critères de qualité spécifiques aux contextes d'intégration. Cet article pose les bases de la conception de l'ontologie d'application et de l'infrastructure technique qui seront développées dans les chapitres suivants.
  - Le Chapitre 4 introduit la métaphore de l'écart d'impédance due à R. Jeansoulin pour établir des analogies entre d'une part les systèmes physiques et d'autre part les systèmes producteurs et utilisateurs de données. Ces analogies conduisent à reformuler le problème de la recherche d'information et à interpréter les problèmes d'hétérogénéité et de qualité de données comme des *résistances* aux flux de données entre systèmes d'information et à leur interopérabilité. On apporte aussi des éléments préliminaires à la conception globale de NARI, par sa structuration en étapes et aspects.
  - Le Chapitre 5 approfondit l'ontologie d'application et donne les lignes générales de modules logiciels, qui peuvent être interprétés comme des *réactances* pour réduire les écarts d'impédance. Nous déterminons les langages de représentation des connaissances aptes à supporter ces modules, nous enrichissons les classifications des requêtes et complétons l'architecture globale de NARI.

# **Chapitre 1 : Une Nouvelle Approche de Recherche d'Information**

- basée sur la métaphore de l'impédance, la qualité et les métadonnées
- structurée par la dimension géographique
- illustrée sur le domaine de la santé

# Table des matières

<b>1. Introduction .....</b>	<b>9</b>
<b>2. Contexte et objectifs généraux de ces recherches.....</b>	<b>10</b>
2.1 Evolutions liées aux Sciences et Techniques de l'Information et de la Communication.....	10
2.2 Complexité du pilotage et du cadre réglementaire des nouvelles applications.....	11
2.3 Evolutions des problématiques en Recherche d'Information .....	13
2.4 Objectif général de ces recherches .....	14
<b>3. Choix d'un champ disciplinaire d'illustration : la santé.....</b>	<b>16</b>
3.1 Facteurs de complexité et spécificités .....	16
3.2 Recherche d'Information en santé .....	17
3.3 Problématiques en Recherche d'Information en santé pour NARI.....	20
<b>4. Etat de l'art .....</b>	<b>21</b>
4.1 Information Géographique, Géomatique et SIG .....	23
4.2 Qualité des données.....	24
4.3 Métadonnées.....	27
4.4 Vue systémique et Métaphore de l'impédance .....	32
4.5 Intégration .....	35
4.6 Démarches de conception de type CSCW.....	41
<b>5. L'approche NARI : synthèse, analyse et discussion .....</b>	<b>44</b>
5.1 Démarche de conception de NARI et objectifs spécifiques.....	45
5.2 Vue synthétique de l'approche.....	49
5.3 Analyse des caractéristiques visées pour NARI.....	52
5.4 Concordances et originalités de NARI par rapport à d'autres approches .....	56
<b>6. Perspectives de recherche et Conclusions .....</b>	<b>60</b>
6.1 Perspectives de recherche.....	60
6.2 Conclusions sur ces recherches .....	64
<b>7. Références .....</b>	<b>65</b>

# 1. Introduction

Ce rapport de recherche présente des travaux de recherche réalisés depuis 2005 sur une Nouvelle Approche de Recherche d'Information (NARI). Bien que des aspects spécifiques de ces travaux aient fait l'objet de présentations à divers colloques, le recul pris par rapport à des échéances spécifiques et des objectifs partiels, nous a conduit à la construction et à la présentation du contexte général dans lequel s'inscrit l'approche NARI, avec un état de l'art succinct mais global, la démarche de conception ainsi qu'une analyse critique sur l'ensemble de ces recherches.

Schématiquement, NARI est une approche de Recherche d'Information, opérant sur des grandes masses de données cataloguées, hétérogènes, qui peuvent être géo référencées<sup>6</sup>. Visant à supporter la prise de décision, NARI est basée sur des exigences de qualité (réglementations, standardisation), exprimées par les utilisateurs, représentées et gérées à l'aide des métadonnées. L'originalité de NARI réside dans la métaphore de l'impédance, qui en guide la conception. Cette métaphore, due à R. Jeansoulin, a été développée la première fois dans l'article conjoint présenté dans le Chapitre 4. La conception de NARI s'appuie sur des techniques d'intégration, sur des langages de représentation des connaissances et sur des technologies et outils relevant du Web sémantique. NARI est illustrée à l'aide d'exemples du domaine de la santé.

Le contenu de ce rapport est le suivant. Après avoir décrit le contexte global de ces recherches et des objectifs généraux (§ 2), on motive le choix d'un champ disciplinaire d'illustration, la santé, en présentant les problématiques de recherche retenues (§ 3). Dans le § 4, on présente ensuite un état de l'art, qui est au même temps succinct et global, car il vise à couvrir les directions caractérisant NARI : Information Géographique et SIG, qualité des données, métadonnées, métaphore de l'impédance et intégration, démarche de conception. Le Chapitre 2 contient un état de l'art détaillé concernant les langages, les technologies et les outils du Web sémantique.

Le § 5 détaille la démarche de conception de NARI et fournit une description synthétique de NARI<sup>7</sup>. Celle-ci sert de base à l'analyse globale des propriétés visées et à une discussion sur le

---

<sup>6</sup> Avec le terme « données géoréférencées » nous désignons des données concernant des « objets géographiques » référencés par des systèmes directs ou indirects (adresses postales, par exemple). (Cfr. §4.1.1).

<sup>7</sup> Les articles présentés dans les Chapitres 3, 4 et 5 en détaillent des aspects spécifiques.

positionnement, l'originalité et les limites actuelles de NARI. Avant de conclure, nous présentons des perspectives de recherche à court et moyen terme (§ 6).

## 2. Contexte et objectifs généraux de ces recherches

### 2.1 Evolutions liées aux Sciences et Techniques de l'Information et de la Communication

Les énormes progrès réalisés depuis quelques années dans deux domaines technologiques, l'information et la communication, ont conduit, avec l'explosion de l'Internet, à une véritable "révolution numérique". Cette révolution dépasse largement le champ technologique et se manifeste avec des profonds changements organisationnels et notamment avec l'essor d'une société en réseau. Les changements de la production et de la diffusion de l'information affectent les entreprises, dont les périmètres deviennent globaux, mais aussi les Etats et les administrations, dans leurs relations avec les citoyens et les administrés. En s'appuyant sur les technologies de l'information et de la communication, de nouvelles interprétations de l'information s'affirment : elles sont liées au concept de "bien public" et s'insèrent dans de nouveaux rapports entre l'offre de données et la demande d'informations, voire de connaissances (Curien et Muet, 2004).

Ces évolutions techniques supportent et accélèrent des découvertes scientifiques majeures, comme par exemple celles concernant la structure de l'ADN et la génétique. En effet, la révolution numérique entraîne aussi une révolution cognitive : d'une part, au travers du Web, on réalise toutes les fonctions "classiques" de la gestion de l'information, de la publication à la recherche d'information ; de plus, en tirant profit des puissances des ordinateurs et des réseaux, ainsi que des capacités de modélisation des connaissances, le Web sémantique vise la construction et la gestion de connaissances individuelles et collectives. (Berners-Lee *et al.*, 2001 ; Thuraizingham *et al.*, 2002 ; Salaün, 2004 ; O'Reilly, 2005)

Ces connaissances concernent plusieurs disciplines : les sciences de l'information et de la communication (STIC), les sciences humaines, sociales et de gestion, les sciences géographiques. Celles-ci sont fortement impliquées, notamment pour les thématiques concernant la spatialité des sociétés en liaison avec les environnements et les territoires. Le domaine de la santé publique, par exemple, nécessite l'acquisition de données sur le territoire,

son occupation, ses ressources, ses habitants et son utilisation. Elle nécessite, au même temps, de systèmes puissants et performants pour analyser ces données (Proulx *et al.*, 2007).

De nombreux domaines sont impactés par ces évolutions, comme par exemple : (i) les secteurs administratifs : environnement, défense nationale et protection civile, équipement, agriculture ; (ii) les collectivités territoriales : aménagement et urbanisme, risque et environnement, transports, services d'urgence et de secours, gestion des déchets et des ressources naturelles ; (iii) les secteurs commerciaux : banques, assurances, immobilier, géomarketing notamment en tourisme. Parmi des champs où des applications très récentes ont été développées en France, on citera : (i) la formation, avec la construction d'un espace universitaire pédagogique médicale (l'Université Virtuelle Médicale Francophone)<sup>8</sup>, (ii) le champ de la santé et du social, avec le remboursement des frais médicaux via la Carte Vitale<sup>9</sup>.

## **2.2 Complexité du pilotage et du cadre réglementaire des nouvelles applications**

Selon les pays et parfois aussi à l'intérieur d'un même pays, les évolutions liées aux STIC se déploient avec des rythmes différents, liés aux niveaux de maturité des systèmes informatiques existants (fracture numérique).

Les nouvelles applications bâties pour des organisations en réseau requièrent l'échange, voir le partage, de données issues de multiples systèmes, conçus a priori de façon autonome et pour des objectifs spécifiques. Pour ces applications, le *pilotage* constitue un facteur critique.

La complexité du pilotage croît avec la décentralisation de l'organisation en réseau, comme l'illustre par exemple le Dossier Médical Personnel (DMP). Le DMP répond à la nécessité d'un meilleur partage des informations médicales entre tous les acteurs des systèmes de santé, en palliant le cloisonnement des systèmes et leurs hétérogénéités. Le DMP s'inscrit dans le processus de diffusion des technologies de l'information et de la communication et a été mis en œuvre par exemple depuis 2005 par le NHS (National Health Services)<sup>10</sup> britannique. En France, il s'inscrit dans un corpus juridique complexe, qui garantit la protection des données médicales personnelles (identifiant anonyme, cryptage des données et des opérations sur les données). Beaucoup d'obstacles doivent être surmontés pour articuler un grand nombre de

---

<sup>8</sup> Université Virtuelle Médicale Francophone : <http://www.umvf.prd.fr/>

<sup>9</sup> Assurance Maladie - Sécurité Sociale : <http://www.ameli.fr/>

<sup>10</sup> National Health Services (NHS) : <http://www.nhs.uk>

systèmes très hétérogènes (systèmes hospitaliers, informatique professionnelle libérale, ...) <sup>11</sup>. De nombreux autres exemples, et notamment la T2A (Tarification à l'Activité, citée dans le § 6.1.1.2), peuvent illustrer les impacts des cadres réglementaires sur les systèmes d'information de santé.

Un pilotage international est nécessaire pour les systèmes à couverture mondiale. Ainsi par exemple, en s'appuyant sur des structures nationales, voire régionales, plusieurs pays coopèrent, au sein de l'OMS, Organisation Mondiale de la Santé <sup>12</sup>, pour renforcer les réseaux de surveillance sanitaire et de lutte contre les épidémies (VIH, SRAS, ...). Ces réseaux doivent couvrir aussi bien les zones rurales, que les zones urbaines, car celles-ci facilitent la diffusion des épidémies, à cause de facteurs divers (fortes concentrations de personnes, populations hétérogènes, mobilité des personnes, ...).

Les changements induits par la "révolution numérique" s'accompagnent de la construction progressive d'un *cadre réglementaire complexe* <sup>13</sup>. *L'information géographique* d'une part structure ce cadre, par la prise en compte de divers niveaux de territoires, et d'autre part permet de corréliser plusieurs thématiques.

Au niveau mondial, par exemple, le RSI (Règlement International Sanitaire) <sup>14</sup>, entré en vigueur en juin 2007, stipule que tout événement constituant une urgence de santé publique et posant un problème international, doit être notifié à l'OMS, qui assure la qualité des moyens mis en œuvre pour contenir l'épidémie et informe l'ensemble des pays. Ces dispositions concernent également les voyages et les transports internationaux. Ainsi le cadre réglementaire en santé doit être corrélé à terme avec la *directive européenne INSPIRE* <sup>15</sup>.

La Directive INSPIRE, entrée en vigueur en mars 2007, concerne la diffusion de données et de services géographiques. Elle a pour objectif de rendre accessibles toutes les données géographiques publiques, à un coût minimal ou nul. Elle fixe des règles pour établir "l'infrastructure d'information géographique dans la Communauté Européenne, aux fins des politiques environnementales communautaires et des politiques ou activités de la

---

<sup>11</sup> DOOR Jean-Pierre "Le DMP : un choix nécessaire pour la qualité des soins à l'heure du numérique", Janvier 2008, Documents d'information de l'Assemblée nationale, n° 659, <http://www.assemblee-nationale.fr/13/rap-info/i0659.asp>.

<sup>12</sup> Organisation Mondiale de la Santé (OMS) : <http://www.who.int/fr/>

<sup>13</sup> Les normes et standards, discutés dans le § 4.3.2.2 et dans le Chapitre 2, participent aussi de ce cadre réglementaire.

<sup>14</sup> OMS - Règlement sanitaire international : <http://www.who.int/csr/ihr/fr/index.html>

Communauté susceptibles d'avoir une incidence sur l'environnement". L'infrastructure d'interopérabilité portée par INSPIRE doit suivre un calendrier progressif, comprenant des étapes de spécification et de mise en œuvre devant s'achever en 2014.

Des évolutions réglementaires trouvent une forte impulsion dans l'exigence, pour les Etats, d'améliorer la gestion des risques majeurs, (prévention, préparation, réponse, rétablissement) (Guarnieri et Garbolino, 2004). En France, par exemple, l'analyse de la gestion de la récente canicule (2003) ou des différentes vagues de froid, a identifié des problèmes majeurs liés à l'existence et à la qualité des données nécessaires à gérer avec efficacité et rapidité ces événements complexes et critiques. Ces analyses ont abouti à des plans d'action, spécifiant les sources de données nécessaires à la gestion de ces risques, en termes de contenu, fréquence de production et organismes responsables. Ceux-ci appartiennent à différents secteurs d'activité (scientifiques, administratifs) et ont différentes couvertures géographiques<sup>16</sup>.

### **2.3 Evolutions des problématiques en Recherche d'Information**

Les évolutions technologiques et organisationnelles synthétisées ci-dessus s'accompagnent de profondes évolutions en *Recherche d'Information (RI)*, qui se manifestent aussi bien aux niveaux des moyens (moteurs de recherche, interfaces conviviales, graphiques, "géographiques") que des usages, de plus en plus diversifiés, aussi bien dans des contextes collectifs d'entreprise qu'individuels. Les systèmes de RI doivent alors répondre à des nouvelles problématiques, liées à l'explosion des volumes des sources de données : (i) traiter l'information en relation aux contextes d'utilisation, en filtrant les sources, pour augmenter la pertinence des réponses, et en adaptant les sorties pour augmenter l'usabilité des résultats produits ; (ii) augmenter les corrélations possibles, en rendant transparentes les hétérogénéités et la distribution des sources.

Le très récent article paru dans Nature "Detecting influenza epidemics using search engine query data" illustre la puissance des méthodologies de recherche d'information associant moteurs de recherche (Google), Systèmes d'Information Géographique (SIG), modèles de simulation et outils poussés d'analyse statistique. Les auteurs démontrent que, en appliquant cette approche de RI pluridisciplinaire, à partir des requêtes effectuées par les internautes, ils

---

<sup>15</sup> INSPIRE : <http://inspire.jrc.ec.europa.eu/>

<sup>16</sup> 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

calculent des prévisions de grippe aux USA, très fiables et bien plus rapidement que les organismes officiels (Ginsberg *et al.*, 2008).

En s'attaquant aux problèmes d'échange et de partage d'information, les recherches en RI rejoignent celles en *interopérabilité des systèmes d'information*. Démarrées il y a environ 30 ans, celles-ci ont fait évoluer les approches, les technologies et les outils pour la conception et la mise en œuvre des systèmes d'information, aussi bien transactionnels que décisionnels.

L'interopérabilité est définie par l'IEEE (Institute of Electrical and Electronics Engineers, Inc.) comme la "capacité de deux ou plusieurs systèmes ou composants à échanger de l'information et à utiliser l'information qu'ils ont échangées" (t. l.)<sup>17</sup> (IEEE, 1990).

Les dimensions couramment utilisées pour analyser et concevoir des solutions d'interopérabilité entre systèmes sont (Sheth, 1999) : (i) l'autonomie d'organisation, de conception, de communication, d'exécution ; (ii) la distribution des données, des systèmes et de leurs descriptions ; (iii) l'hétérogénéité, aux différents niveaux : des systèmes (d'exploitation, de gestion de bases de données, de communication, ...), de structure, de sémantique des données, des contextes (Wiederhold et Jannink, 1999).

Ces dimensions doivent être interprétées actuellement en relation aux avancées scientifiques, en particulier en médecine et sciences de l'environnement, et technologiques, avec le développement des postes de travail, des réseaux, des moteurs de recherche, devenus aussi géographiques, et du Web sémantique. L'actualité et la complexité des problèmes d'intégration de données multimédia et de la fusion d'informations hétérogènes sont reconnus comme des axes de recherche prioritaires dans la communauté des Bases de Données et de la Recherche d'Information (Abiteboul *et al.*, 2005 ; Halevy *et al.*, 2006)

## 2.4 Objectif général de ces recherches

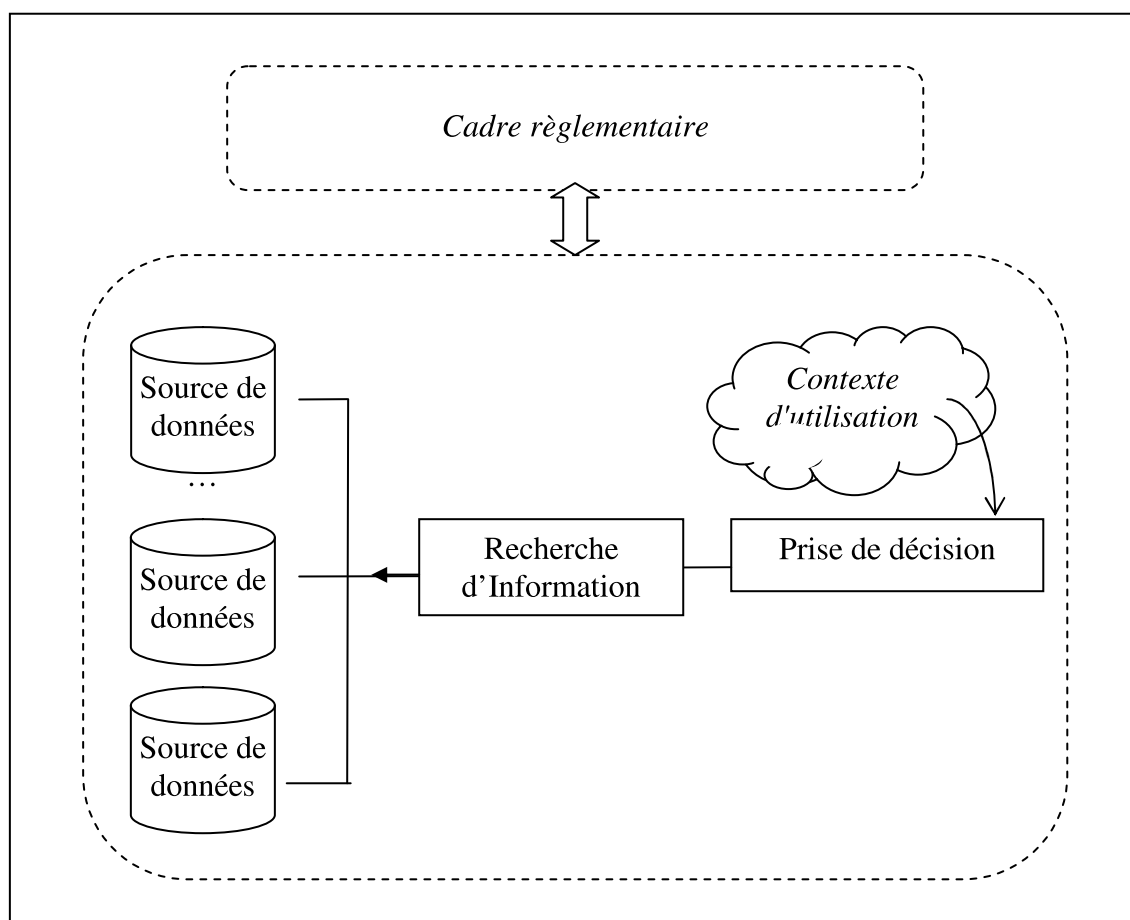
Ces travaux de recherche visent à avoir une "*pertinence sociétale*" : ils prennent appui sur des questionnements relatifs aux transformations de la société, notamment en ce qui concerne la qualité des prises de décision, qui est basée sur la qualité des données disponibles et du processus de RI. Concernant le premier aspect, de multiples organismes publics ou reconnus s'efforcent de réguler progressivement la qualité de la production de données en établissant

---

<sup>17</sup> Le sigle (t. l.), ici et dans la suite du document, indique une "traduction libre" à partir de la source citée.

des cadres réglementaires complexes. Concernant le processus de RI, on requière qu'il soit apte à fournir rapidement des réponses pertinentes, même si approchées, aux questions posées par les décideurs, opérationnels aussi bien que stratégiques.

L'objectif général de ces recherches, schématisé dans la Figure 1, est d'apporter des éléments de conception d'une Nouvelle Approche de Recherche d'Information (NARI), et de les illustrer sur des exemples. Cette approche doit opérer sur des données hétérogènes, géo référencées<sup>18</sup> et cataloguées, produites par des systèmes autonomes, conçus pour des objectifs spécifiques. Elle doit contribuer à la prise de décision et prendre en compte les besoins préliminaires de qualité exprimés par les utilisateurs.



**Figure 1** : Schématisation des objectifs de l'approche NARI.

Avant de présenter succinctement un état de l'art relatif aux thèmes explorés pour la conception de NARI, nous choisissons un champ d'illustration pour l'illustrer.

<sup>18</sup> Le géoréférencement des données peut être indirect, basé par exemple sur les adresses postales (cfr. §4.1.1)

### 3. Choix d'un champ disciplinaire d'illustration : la santé

La santé constitue un champ d'illustration privilégié pour ces recherches, de par la multitude des facteurs de complexité, la spécificité de certains d'entre eux et la diversité des besoins décisionnels (§ 3.1). Dans ce champ d'application nous identifions trois principaux axes de recherche, en étroite liaison avec les récentes évolutions scientifiques et technologiques (§ 3.2).

#### 3.1 Facteurs de complexité et spécificités

La RI en santé opère sur des sources produites par des *systèmes très hétérogènes*.

Les systèmes hospitaliers occupent une place prépondérante parmi les systèmes producteurs de données de santé. Initialement dédiés essentiellement aux aspects administratifs, les SI hospitaliers sont développés de façon de plus en plus modulaire pour répondre aux demandes des différents services (gestion administrative, dossier patients, actes biologiques, dossier infirmier, gestion des prescriptions, ...) (Verdier et Ouziri, 2002)

Les systèmes dits de 'médecine de ville', très répandus en France à différence d'autres pays, compte tenu de l'organisation des soins, visent essentiellement la facilité d'utilisation dans le suivi des dossiers patients (diagnostics et prescriptions) et les fonctions liées à la gestion des cabinets (rendez-vous, fiscalité, ..). Ils intègrent très faiblement des données externes, en provenance des centres publics de soins et des laboratoires. Ces systèmes informatiques de santé (hospitaliers et de ville) s'appuient sur des centaines de logiciels et systèmes propriétaires.

Depuis les derniers 10 ans, les Systèmes d'Information (SI) de santé ont beaucoup évolué (Haux, 2006), en termes de couverture (SI d'hôpitaux, de réseaux de soins régionaux, nationaux voir globaux), finalités (recherche clinique ou épidémiologique, planification, gestion de risques), fonctionnalités (ubiquité, mobilité, acquisition massive de données par des capteurs), qualité (processus de normalisation, protocoles d'échanges).

Plusieurs *facteurs de complexité* sont associés à la Recherche d'Information en santé : le nombre de structures et d'acteurs impliqués, leur distribution géographique, les volumes des données (énormes, car touchant par pays la totalité des populations), le nombre croissant et l'évolution des spécialisations médicales, en liaison aux progrès médicaux (notamment en

génétique et santé environnementale), la multitude et l'hétérogénéité des classifications médicales, la mobilité des patients.

*Les spécificités* du domaine sont liées premièrement au poids économique du secteur (en France, par exemple, de l'ordre de 10% du PIB). Elles sont liées aussi au cadre réglementaire, particulièrement complexe, surtout depuis la très récente loi sur les "Droits des malades et la qualité du système de santé", (2002)<sup>19</sup> qui impose des "contraintes particulières qui obligent à prévoir des fonctionnalités et des modes de gestion adaptés. La question du secret médical, et la sensibilité particulière des informations sont au centre de cette problématique" (Villac, 2004).

Ce cadre réglementaire émane de plusieurs organismes et comprend une multiplicité de codes (santé publique, déontologie, hospitalisation à domicile, modalités de prises en charge, ...) valides à des niveaux de couverture divers, international ou national, avec des codes spécifiques, relevant des communes, des collectivités territoriales, des régions, notamment pour l'Hospitalisation à Domicile. Il doit se composer avec les réglementations régissant les organismes publics et privés.

### **3.2 Recherche d'Information en santé**

La Recherche d'Information en santé s'attaque progressivement à des besoins décisionnels, liés aux évolutions décrites ci-dessus. (Bounekkar et Duru, 2008) structure cette recherche en trois axes. Pour souligner le caractère innovant des recherches en informatique décisionnelle en santé, nous illustrons ces axes par des projets, qui sont 'pionniers'<sup>20</sup> et récents en même temps :

- (i) *Générer de l'information* via la construction de documents à partir de l'intégration de sources de données hétérogènes ou via la présentation de données extraites d'un entrepôt de données.

(Salzano et Bourret, 2004) analysent les premiers projets de dossier médical partagé, aboutissant à des résultats opérationnels. Ces projets concernent par exemple la France, avec des structures récentes de réseaux de soins, en certaines régions (Franche

---

<sup>19</sup> Loi n° 2002-303 du 4 mars 2002 relative aux droits des malades et à la qualité du système de santé, JORF du 5 mars 2002 page 4118, <http://www.legifrance.gouv.fr/>

<sup>20</sup> Des références supplémentaires analysées dans les articles joints (Chapitres 3, 4 et 5) attestent de l'actualité et de la complexité croissante de ces recherches.

Comté, Rhône Alpes, Nord pas de Calais), la Suisse (projet Diogène), l'Italie (projet de télémédecine à l'hôpital de Pise), l'Europe (projet PICNIC (Professional and Citizen Networks for Integrated Healthcare). Le système GENNERE (a Generic Epidemiological Network for Nephrology and Rheumatology) basé sur une expérimentation française dans le domaine des maladies rénales, a été étendu à plusieurs spécialités médicales et à des pays, comme la Chine, avec des modes de prise en charge spécifiques.

- (ii) *Générer de la connaissance*, dans un but d'aide à la décision (planification de ressources, optimisation de coûts, maximisation du service rendu).

Le premier entrepôt de données couvrant en France l'ensemble des assurés a été élaboré pour le projet SNIIR-AM (Système National d'Information Inter-Régimes de l'Assurance Maladie) (Nakache, 2003). Il permet d'analyser par région les interventions de l'Etat dans le financement de l'offre des soins, en corrélant les prescriptions médicales et les remboursements. Au niveau européen, le programme européen SCALE, a été, au niveau de la Commission européenne, l'un des premiers à associer des sources de données concernant la santé, l'environnement et le droit. Par une approche de "médecine par l'évidence", il a contribué à améliorer la stratégie européenne dans ces domaines, et minimiser les effets nocifs de la pollution sur la santé. Les principes de précaution de SCALE, "based on Scientific evidence, focused on Children, meant to raise Awareness, improve the situation by use of Legal instruments and ensure a continual Evaluation of the progress made", sont retenus dans le système européen REACH<sup>21</sup>, entré en vigueur en juin 2007. Ce système d'enregistrement, d'évaluation et d'autorisation des substances chimiques, a pour objectif de rendre ces substances plus sûres pour la santé humaine et l'environnement.

- (iii) *Evaluer les pratiques et les politiques de santé*, afin d'améliorer globalement les systèmes de santé.

Ces recherches explorent les aspects d'organisation des structures de soins, de prise en charge des patients (processus), des stratégies de soins, des aspects économiques, la dimension environnementales (Bonnevay et Lamure, 2002). Un portail<sup>22</sup> recueille des

---

<sup>21</sup> Système REACH :

<http://www.euractiv.com/fr/environnement/revision-politique-substances-chimiques-reach/article-120270>

<sup>22</sup> UMIT University for Health Sciences, Medical Informatics and Technology, Amsterdam, Inventory of evaluation studies : <http://evaldb.umat.at/index.htm>

études d'évaluation en santé, corrélées à la planification, à l'introduction, au développement d'approches et de technologies de l'information en santé. Il montre l'évolution des approches de 1982 à 2005. L'importance et la complexité de l'évaluation se sont accrues en France, grâce à la loi du 4 mars 2002 déjà citée, sur les "Droits des malades et la qualité du système de santé", et à l'étranger grâce à des mouvements analogues. Ces mouvements consacrent "le patient au centre du dispositif de santé", apportent une interprétation très large, pluridisciplinaire, de la santé et conduisent au développement d'organisations en réseau. Un meilleur partage de l'information facilite la coordination et la continuité des soins, ainsi que de nouvelles relations entre un patient, devenu acteur de sa santé, et les différents praticiens.

Les SI sont donc évalués par rapport à un ensemble de rôles : ils doivent non seulement gérer l'information (la stocker et la transmettre) mais aussi améliorer les relations et les compétences des acteurs (patients, praticiens, institutionnels), dans des organisations en constante évolution, en tenant compte des temps d'apprentissage et d'adaptation.

La diversité des scénarios de prise de décision est en relation avec la responsabilité des acteurs. Le tableau ci-dessous, inspiré de (Buthion et Flory, 2002), synthétise des exemples de problématiques décisionnelles en santé en liaison avec les responsabilités des acteurs.

Type de décideur	Exemples de problématique décisionnelle
Décideur public	Efficacité des thérapeutiques dans le cadre d'une analyse d'intérêt de santé publique
Directeur Régional des Affaires Sociales	Comparaison des budgets en fonction de la gravité des pathologies traitées
Directeur d'hôpital	Comparaison des coûts internes de fonctionnement par rapport aux attributions budgétaires
Chef de service	Gestion du budget par rapport aux prestations réalisées
Prescripteur	Complément d'information économique pour le choix entre plusieurs thérapeutiques

**Table 1 :** Exemples de problématiques décisionnelles en santé, extraite de (Buthion et Flory, 2002)

### 3.3 Problématiques en Recherche d'Information en santé pour NARI

En guise de conclusion sur le choix de la santé comme champ d'illustration pour NARI, nous synthétisons des problématiques de recherche d'information en santé auxquelles NARI vise de contribuer.

Qu'il s'agisse d'études d'épidémiologie, d'analyse d'accès aux soins, de planification sanitaire, de gestion de risques ayant un impact sur la santé, on est confronté aux *problèmes de volumes et d'hétérogénéité des sources de données disponibles*, qui ont été produites dans des buts différents de ceux de l'étude (administratifs, de gestion, réglementaires). En France, ces données sont contenues par exemple dans les bases de données hospitalières (PMSI<sup>23</sup>), les schémas régionaux de l'offre de soins (SROS<sup>24</sup>), les fichiers des ressources des établissements (répertoire FINISS<sup>25</sup>), les fichiers des professionnels de santé (ADELI<sup>26</sup>)<sup>27</sup>.

Ces données doivent être couplées avec des *données géographiques ou environnementales* (infrastructures de transport, météorologie, cartes des risques, hydrologie, ...) et des *données concernant les populations*, produites par l'INSEE<sup>28</sup>, complexifiant davantage les problèmes de volumétrie et d'hétérogénéité.

Comme le fait ressortir le rapport sur l'Enquête sur le croisement de données dans le champ santé environnement<sup>29</sup>, menée dans le cadre du Plan Nationale de Santé Environnementale (PNSE), même si chaque étude a des objectifs spécifiques, toutes les étapes préalables au croisement de données sont caractérisées par "le temps passé à identifier les sources de données et à appréhender les données qu'elles contiennent, les modalités d'accès fondées sur le relationnel et la négociation, l'optimisation ou le contournement des limites des données disponibles parfois complexes, ...".

---

<sup>23</sup> PMSI : <http://www.le-pmsi.fr/>

<sup>24</sup> SROS : <http://www.sante.gouv.fr/htm/dossiers/sros/presros.htm>

<sup>25</sup> FINISS : <http://finiss.sante.gouv.fr/index.jsp>

<sup>26</sup> ADELI : <http://www.sante-sports.gouv.fr/dossiers/sante/adeli/repertoire-adeli.html>

<sup>27</sup> Ces exemples sont généralisables à plusieurs autres pays (Canada, Europe).

<sup>28</sup> INSEE : <http://www.insee.fr/fr/default.asp>

<sup>29</sup> Agence Française de sécurité sanitaire de l'environnement et du travail (AFSSET), en partenariat avec l'Institut français de l'environnement (IFEN) : "Enquête sur le croisement de données dans le champ santé environnement", <http://www.afsset.fr/index.php?pageid=2051&parentid=523>

Dans ce contexte, les problématiques de recherche auxquelles NARI s'attaque sont :

- chercher l'information à partir de cette énorme masse de données hétérogènes disponibles
- cibler les données pertinentes le plus rapidement possible
- adapter le choix des données en fonction du type de décision
- pallier le manque de données ou leur insuffisante qualité, pour produire une information de qualité suffisante par rapport aux besoins décisionnels

## 4. Etat de l'art

La conception de NARI s'appuie sur une démarche de "*recherche appliquée*" en bases de données et ingénierie des systèmes d'information, qui intègre plusieurs méthodes, techniques, et outils.

Les ressources informationnelles accumulées (ou accessibles) par les entreprises<sup>30</sup> ne sont plus rares, mais au contraire surabondantes. "Le problème est de la trouver, de la sélectionner, de l'interpréter, ..., de faire émerger du sens de ces grandes masses d'information, exploiter celles qui leur sont propres ou bien celles qui, rapprochées de leurs savoir faire spécifiques, peuvent donner des atouts supplémentaires" (Charlet *et al.*, 2001). Ces demandes ont donné naissance à différentes formes de *Systèmes décisionnels*. Ces systèmes sont complémentaires des systèmes transactionnels, qui privilégient les fonctions de stockage, accès, mise à jour des données et leur intégrité (Grundstein et Rosenthal-Sabroux, 2001). Ils ont pour objectifs une exploitation efficace, performante et "intuitive" (adaptée à l'utilisateur final), de cette énorme masse de données hétérogènes et distribuées. Après les SIAD (Systèmes Interactifs d'Aide à la Décision), et les EIS (Executive Information Systems), plus récemment les systèmes de médiation et surtout les entrepôts de données (§ 4.5.1.2) associent des systèmes d'information géographique (SIG, § 4.1) pour exploiter la dimension géographique de l'information et apporter de la valeur ajoutée à l'information. Même dans les contextes où les sources de données sont identifiées et cataloguées, la multiplicité de ces sources et les usages possibles de ces données, posent des problèmes de pertinence et d'optimisation, analogues à ceux de la RI sur le Web (Salaün, 2004 ; O'Reilly, 2005).

---

<sup>30</sup> Le terme "entreprise" est utilisé en sens large et désigne ici et dans la suite une administration, une entreprise industrielle, une société opérant dans les services, ...

Dans la suite de ce paragraphe, on présente un état de l'art dans les directions caractérisant NARI : apports de l'Information Géographique et des SIG, qualité des données et métadonnées, métaphore de l'impédance et démarche de conception de systèmes coopératifs.

Bien que très succinct<sup>31</sup> et prenant en compte le champ d'illustration de la santé, cet état de l'art a pour objectif de témoigner de la complexité des problèmes posés, compte tenu de la multiplicité des champs disciplinaires impliqués et des perspectives possibles (de l'organisation à la technique). Un état de l'art sur les technologies du web sémantique et les outils associés, en relation à NARI, est présenté dans le Chapitre 2.

Plus précisément, le plan de ce paragraphe est le suivant :

- 4.1 Information Géographique, Géomatique et SIG
  - 4.1.1 Généralités
  - 4.1.2 Essor des SIG en liaison avec la santé
- 4.2 Qualité des données
  - 4.2.1 Exigences de qualité
  - 4.2.2 Qualité externe
- 4.3 Métadonnées
  - 4.3.1 Définitions générales et interprétations des métadonnées
  - 4.3.2 Métadonnées et standardisation en santé et géographie
    - 4.3.2.1 Vue synthétique du processus de standardisation
    - 4.3.2.2 Métadonnées en santé et géographie
- 4.4 Vue systémique et Métaphore de l'impédance
  - 4.4.1 Vue systémique de NARI
  - 4.4.2 Métaphore de l'impédance
    - 4.4.2.1 Concepts généraux
    - 4.4.2.2 Travaux de recherche concernant l'écart d'impédance
- 4.5 Intégration
  - 4.5.1 Formes d'intégration
    - 4.5.1.1 Médiation
    - 4.5.1.2 Entrepôt de données
  - 4.5.2 Qualité externe, métadonnées et intégration
  - 4.5.3 Étapes et techniques d'intégration
    - 4.5.3.1 Étapes d'intégration
    - 4.5.3.2 Techniques de description des schémas
  - 4.5.4 Modèles de données semi-structurées et aspects technologiques
- 4.6 Démarche de conception de type CSCW
  - 4.6.1 Aspects méthodologiques
  - 4.6.2 Pertinence de l'approche CSCW par rapport à NARI

---

<sup>31</sup> Des références complémentaires sont discutées dans les articles joints (Chapitres 3, 4 et 5).

## 4.1 Information Géographique, Géomatique et SIG

### 4.1.1 Généralités

Un *objet géographique* est une représentation abstraite d'un phénomène réel lié à un lieu ou à une zone géographique spécifique. Les objets géographiques sont instanciés par des données géographiques, et on désigne avec le terme de "série de données géographiques" un ensemble identifiable de données géographiques"<sup>32</sup>. Les coordonnées géographiques des objets peuvent être déterminées par des systèmes de référence directs (par exemple, télédétection, campagnes de mesures, ...) ou des systèmes indirects, comme par exemple les adresses postales, qui relient les objets géographiques entre eux.

Les *SIG (Systèmes d'Information Géographique)* sont des systèmes informatiques permettant, à partir de multiples sources, de gérer, analyser et présenter l'information géographique. Les composantes principales des SIG sont des SGBD, des logiciels de cartographie et des puissantes fonctions d'analyse. Les principales fonctions des SIG, regroupées sous le terme de '5A', sont l'Abstraction (modélisation), l'Acquisition, l'Archivage, l'Affichage et l'Analyse (Bordin, 2002). Ces fonctions sont utilisées sous différentes formes, en fonction des objectifs, par exemple (Proulx *et al.*, 2007) :

- (i) des données détaillées et définies à très haute résolution pour les chercheurs en sciences de l'environnement ;
- (ii) des données fortement agrégées, accompagnées de représentations graphiques et des formats compatibles avec des puissants outils statistiques pour les planificateurs ou évaluateurs ;
- (iii) des données simplement agrégées, accompagnées d'outils de visualisation et de systèmes d'alerte pour des événements dangereux pour la santé (pollution de l'air, inondations, ...) pour d'autres utilisateurs.

La dimension géographique structure la *géomatique*, discipline désignée au Québec comme « ayant pour objet la gestion des données à référence spatiale et qui fait appel aux sciences et aux technologies reliées à leur acquisition, leur stockage, leur traitement et leur diffusion »<sup>33</sup>.

<sup>34</sup>. Les problématiques couvertes par la géomatique dépassent les aspects de localisation des

---

<sup>32</sup> Ces définitions suivent la directive INSPIRE :

<http://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=OJ:L:2007:108:0001:0014:FR:PDF>

<sup>33</sup> Cfr l'article sur la Géomatique, de Patrice Langlois : <http://www.hypergeo.eu/spip.php?article68>

<sup>34</sup> Les communautés de recherche en géomatique approfondissent des méthodes, techniques, outils et applications, lors de congrès, nationaux (comme SAGEO (Spatial Analysis and GEomatics) en France et

objets géographiques : (i) au travers des analyses spatiales, elle s'intéresse aussi aux objets avec lesquels les objets géographiques rentrent en interaction et des types d'interactions ; (ii) la prise en compte de la dimension temporelle permet d'étudier l'évolutions des phénomènes qui impactent les objets géo référencés ; (iii) elle contribue à l'aide à la décision, par des études de simulation de phénomènes et des études d'impact.

#### **4.1.2 Essor des SIG en liaison avec la santé**

La dimension géographique prend une importance croissante dans le foisonnement des recherches sur les 'nouveaux' Systèmes informatiques, qui sont pervasifs, sur le web, collaboratifs (Sèdes, 2007)<sup>35</sup>. Elle structure fortement la RI en santé, en particulier pour les applications décisionnelles. Elle est nécessaire pour localiser des ressources et organiser les services, sur la base de l'exploration de corrélations spatiales et thématiques, guider la collecte d'informations, leur évaluation et diffusion, générer des nouvelles connaissances et améliorer la prise de décision en prenant en compte des spécificités liées à la couverture géographique (Vigneron et Tonnellier, 1999).

Les champs d'application des SIG en santé concernent particulièrement les études d'épidémiologie, en particulier pour les pays en voie de développement, et les études concernant les accès aux soins et la planification des ressources. En corrélation avec ces champs applicatifs, les développements plus récents, sont liés respectivement à la télédétection, et aux nouvelles techniques de cartographie et à Internet (Bédard *et al.*, 2003 ; Goodchild, 2005).

## **4.2 Qualité des données**

### **4.2.1 Exigences de qualité**

La définition et la formalisation de critères de qualité, aptes à lier l'information recherchée et le contexte d'interrogation, améliorent le processus de Recherche d'Information, depuis

---

Géomatique au Canada), et internationaux (ACM GIS, EuroGeographics, ISSDQ (International Symposium on Spatial Data Quality). L'ACI (ACI : <http://cartography.tuwien.ac.at/ica/>) (Canada) a créé un Groupe de travail sur les technologies géospatiales libres et open source, et la revue ISI : Ingénierie des Systèmes d'Information (Hermès), dédiera un numéro spécial (2009) à la thématique des Systèmes d'information et géolocalisation.

<sup>35</sup> Les congrès de la communauté informatique en SI et BD (AIM, ER, VLDB, ...) dédient des workshops à la dimension géographique.

l'expression des besoins des utilisateurs et/ou des applications jusqu'à la production d'une information le plus possible pertinente et exploitable.

Le champ des recherches sur la qualité des systèmes est très étendu et interdisciplinaire. Il concerne les différentes phases du processus métier et du développement informatique : acquisition des données, modélisation, stockage, intégration, diffusion, recherche, analyse. Les problèmes concernent la qualité des données, des modèles conceptuels, des processus. (Batini et Scannapieco, 2006) présente un panorama des approches et techniques pour modéliser les dimensions de la qualité des données, identifier des indicateurs, proposer des métriques et des méthodes d'évaluation, pour des systèmes à large échelle.

Les critères de précision, complétude, cohérence logique et sémantique, fraîcheur sont étudiés dans (Bouzeghoub et Peralta, 2004), qui propose une taxonomie des approches par rapport à ce dernier critère. Cette taxonomie prend en compte (i) la dynamique des changements des données (rare / fréquente); (ii) le type d'intégration (virtuelle / matérialisée) ; les règles de synchronisation entre les requêtes et les données (modes pull/push).

(Akoka *et al.*, 2007) développent un meta-modèle qui relie (i) les dimensions de qualité, (ii) les facteurs qui la composent, (iii) les métriques associées à chaque facteur et (iv) les méthodes de mesure qui peuvent être associées à ces métriques. Par exemple, on peut associer au critère 'précision syntaxique', de la dimension 'précision des données' une métrique basée sur le pourcentage de valeurs non-conformes syntaxiquement à un modèle de référence. L'évaluation de cette mesure pourrait s'appuyer sur diverses méthodes, basées par exemple sur la proportion de valeurs identiques ou sur le coût des opérations nécessaires pour rendre identiques des valeurs différentes.

Ce métamodèle est utilisé pour évaluer les dimensions et les critères significatifs pour divers scénarios, ainsi que les interdépendances entre les dimensions de qualité, notamment dans deux systèmes multisources, portant respectivement sur des données de santé et des données géographiques. Dans le scénario de *santé*, pour le développement du dossier médical partagé, les critères retenus concernent l'identité du patient et donc l'élimination des doublons, les données manquantes et la fraîcheur des données des dossiers de soin. Les critères retenus dans le scénario de *géographie* portent sur la complétude des données, la précision sémantique et la cohérence.

Concernant les *données géographiques*, leur précision géométrique est de grande importance, de même que les précisions sémantiques et temporelles et la 'généalogie' des données. Ce critère décrit les différentes phases et transformations subies par les données, de l'acquisition par différents moyens (campagnes de terrain, télédétection, numérisation des cartes, ...), jusqu'à la diffusion (Gutiérrez et Servigne, 2007).

Le choix des critères, les poids assignés à chaque critère, ainsi que le type de seuil (déterministe ou flou) et les valeurs, dépendent de la spécificité des projets et des points de vue qui doivent être pris en compte. Par exemple, les recherches d'itinéraires pour rejoindre des centres de soins s'appuient sur des critères différents, selon que ces itinéraires soient destinés aux pompiers (voies rapides de communication par la route) ou à des personnes pouvant utiliser des transports en commun.

#### **4.2.2 Qualité externe**

Le point de vue exploré dans nos recherches concerne plus particulièrement la qualité "externe". Analysé par (Devillers et Jeansoulin, 2005), ce concept concerne la qualité des données attendues pour un usage donné, pour répondre aux besoins exprimés explicitement ou implicitement par des groupes d'utilisateurs. En contraposition, la qualité définie "interne", concerne les données produites (généalogie et modèles appliqués, ...) et leur distance par rapport aux spécifications initiales.

La qualité externe, appelée aussi '*fitness for use*', inclue des aspects contextuels, donc variables, qui rendent plus complexe la formalisation de la qualité des données. Bien que s'appuyant sur les standards, les approches d'évaluation de la qualité externe sont très difficiles à cerner. Des facteurs de complexité sont :

- (i) la subjectivité des analyses des risques associés à la non qualité,
- (ii) les "distances", parmi le monde "réel" et les modèles,
- (iii) les interactions (production, usage) que les utilisateurs (producteurs de données, utilisateurs) réalisent avec les modèles et le monde réel, en fonction de leur rôle.

Dans le champ de *l'information géographique*, (Gervais, 2004) propose des approches et des outils pour visualiser de façon très intuitive des propriétés géographiques des territoires. Ces approches sont destinées à des décideurs opérationnels, pour les affranchir des interprétations de seuils de valeurs d'admissibilité pour différents critères de qualité. (Devillers et Beard,

2005) relie les propriétés de qualité externe, performance et adaptabilité des systèmes logiciels, dans le concept de "quality aware" : un SIG "quality aware" est défini comme un système 'permettant d'intégrer différentes techniques de documentation, gestion et utilisation des informations sur la qualité, dans le but d'améliorer leur fonctionnement'. (Stein *et al.*, 2009) traite l'ensemble des problèmes de qualité de données géo référencées, pouvant intervenir en fouille de données et génération de connaissances, depuis l'acquisition des données à leurs utilisations.

Les modèles de données semi-structurées (§ 4.5.4) contribuent largement à la Recherche d'Information à partir d'entrepôts de données ou au travers de l'Internet, à la qualité externe de l'information, et globalement à sa corrélation aux contextes. Par exemple, pour répondre aux besoins des utilisateurs interrogeant un entrepôt de documents multimédia, (Amous *et al.*, 2004) proposent une approche dans laquelle les documents multimédia sont vus comme des documents semi-structurés, dont les éléments sont alimentés par des sources hétérogènes. Les documents attendus sont construits dynamiquement, en partant d'une description des contenus des documents interrogés.

### **4.3 Métadonnées**

Les métadonnées structurent largement les approches, dites 'préemptives', pour modéliser, mesurer, contrôler et améliorer la qualité des données dans les bases de données relationnelles (Berti-Equille, 2004). Après avoir donné des définitions générales et une interprétation des métadonnées dans le champ d'application, nous analysons les métadonnées en relation avec l'intégration et avec le processus de standardisation (§ 4.3.2.2).

#### **4.3.1 Définitions générales et interprétations des métadonnées**

Les métadonnées, définies littéralement comme "données au dessus des données" dépassent désormais leur définition. L'ISO (International Organization for Standardization) les définit comme des "données utilisées pour décrire des ressources. Dans ce sens plus général, une ressource peut représenter des données, des documents, des pièces de musées ou n'importe quelle autre classe d'objets dont la description est exigée pour certains objectifs" (t.l., ISO/IEC N1257, 2005). La directive INSPIRE renforce la portée des métadonnées, définies comme des

"informations décrivant les séries et services de données géographiques et rendant possible leur recherche, leur inventaire et leur utilisation"<sup>36</sup>.

Les analyses des rôles mutuels des métadonnées, des contextes et des ontologies et leurs apports au partage de données entre systèmes distribués, au travers de plusieurs contextes et domaines d'activités, démarrées avec (Kashyap et Sheth, 1999), se développent aussi en liaison avec les exigences de qualité.

Les métadonnées sont nécessaires aux services de catalogage, localisation, interrogation et diffusion de ressources d'information, définis globalement 'services de médiation', et les verrous scientifiques à dépasser sont en relation avec les différents types de services (Libourel, 2003).

### **4.3.2 Métadonnées et standardisation en santé et géographie**

#### **4.3.2.1 Vue synthétique du processus de standardisation**

Les normes et standards sont essentiels pour atteindre une intégration durable de données multipartenaires, même si ils ne peuvent pas la garantir par eux-mêmes. En effet, le partage d'information passe par l'harmonisation de plusieurs niveaux : des protocoles d'échange, aux types et structures de documents, aux contenus et classifications, en plus des niveaux sous-jacents et surtout du niveau externe, par exemple pour la facilité d'utilisation et les sécurités. Les standards contribuent aussi à développer réellement une culture commune intra- et inter-organisations, sur les niveaux cités ci-dessus, et à faire avancer les débats entre les acteurs (fournisseurs, utilisateurs, administrateurs, ...).

Au niveau international, de très nombreux organismes et communautés participent aux activités de standardisation. Les normes émanent d'organismes officiels, mandatés ou reconnus, comme l'Organisation Internationale de Normalisation (ISO)<sup>37</sup>, au niveau mondial, et le Comité européen de normalisation (CEN)<sup>38</sup>, au niveau européen, avec de nombreux comités techniques organisés par discipline (Computer Sciences, Geographic Information, Health Information, ...). Des instances nationales animent des activités de standardisation et harmonisation et servent de relais vers ces organismes internationaux. Dans ce même sens

---

<sup>36</sup> Cette définition suit la directive INSPIRE :

<http://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=OJ:L:2007:108:0001:0014:FR:PDF>

<sup>37</sup> ISO : [www.iso.org/iso/fr/home.htm](http://www.iso.org/iso/fr/home.htm)

opèrent des très larges consortiums et groupements, comme le W3C (World Wide Web Consortium)<sup>39</sup>, en ce qui concerne une architecture pour le Web sémantique et les Web services<sup>40</sup>, et l'Object Management Group (OMG)<sup>41</sup>, en ce qui concerne les modèles et développements orientés-objet, ou le Open Geospatial Consortium (OGC), promoteur de standards d'implémentation techniques.

La culture autour des standards se diffuse largement grâce à l'action de très larges consortiums ou d'organismes internationaux, comme l'ACM (Association for Computing Machinery), l'IEEE (Institute of Electrical and Electronics Engineers, Inc.), ou l'IFIP (International Federation for Information Processing). Les grands industriels s'impliquent dans le processus de standardisation, en développant des produits compatibles avec les standards les plus répandus. Oracle<sup>42</sup> en est un exemple, avec les supports de modèles de données objets et semi-structurés.

A côté des standards officiels, d'autres spécifications sont devenues des "standards de facto" : c'est le cas par exemple en santé avec Health Level 7 (HL7)<sup>43</sup>, concernant au départ les échanges de messages de données médicales, ou avec le Digital Imaging and Communications in Medicine (DICOM)<sup>44</sup>, pour le traitement et les échanges d'images digitales.

Pour l'élaboration d'applications coopératives sur des domaines d'activité transversaux, comme en santé ou en géographie, un des problèmes majeur est d'identifier et combiner les standards les plus adéquats, compte tenu de la richesse mais aussi de la fragmentation et de l'hétérogénéité de la production, qui aboutit à des classifications des dimensions de la qualité souvent différentes.

La norme ISO 19126, "Information technology. Software product evaluation. Quality characteristics and guidelines for their use", insiste notamment sur l'attention devant être apportée aux critères non fonctionnels et en particulier à l'usabilité.

Du point de vue gestionnaire, la gestion de la qualité fait l'objet de la série de normes ISO 9000. Celles-ci définissent la qualité comme "l'aptitude d'un ensemble de caractéristiques

---

<sup>38</sup> CEN : <http://www.cen.eu>

<sup>39</sup> W3C : <http://www.w3.org/>

<sup>40</sup> Les avancées du W3C et les travaux de recherche sur le Web sémantique font l'objet du Chapitre 2

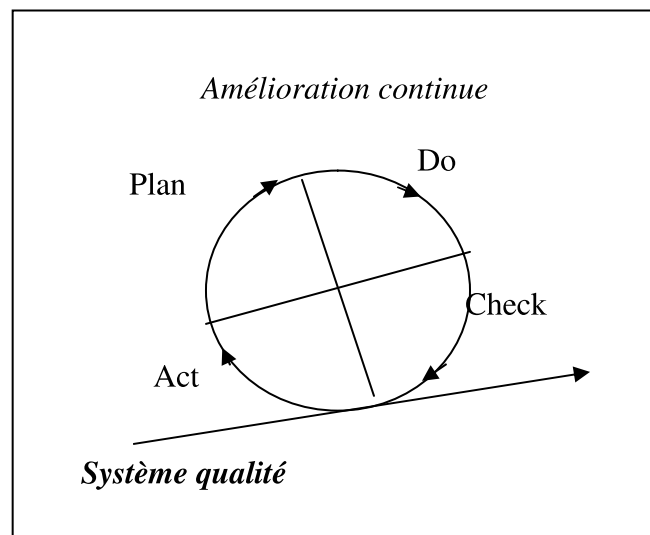
<sup>41</sup> OMG : <http://www.omg.org/>

<sup>42</sup> Oracle : <http://www.oracle.com>

<sup>43</sup> HL7, Health Level 7 : <http://www.hl7.org/>

intrinsèques à satisfaire des exigences"<sup>45</sup>. Cette norme s'applique aussi bien à des produits industriels, qu'aux services et aux administrations. Une approche orientée processus préconise un ensemble d'actions visant une amélioration continue de la qualité généralement représentées sous la forme de la "roue de Deming" (Figure 2), qui représente le paradigme "Plan - Do - Check - Act". Chaque quadrant de la roue caractérise les actions d'une étape :

- *Plan* définit les objectifs à atteindre et planifie la mise en oeuvre d'actions pour les atteindre,
- *Do* correspond à la réalisation de ces actions,
- *Check* consiste à faire des bilans par rapport à l'atteinte des objectifs fixés, aux résultats obtenus, aux dysfonctionnements rencontrés,
- *Act* correspond à la recherche des axes d'amélioration.



**Figure 2** : Roue de Deming

(Bourret *et al.*, 2002) adapte ce paradigme aux Systèmes d'information de Santé, et le relie à la mise en place des organisations en réseaux, soulignant la complexité de l'évaluation ("Check").

<sup>44</sup> DICOM : [http:// medical.nema.org/](http://medical.nema.org/)

<sup>45</sup> Définition ISO 9000 de la "qualité" : <http://www.pme.gouv.fr/essentiel/vieentreprise/qualite/annexes.htm#a2>

#### 4.3.2.2 Métadonnées en santé et géographie

La description des métadonnées de la Dublin Core Metadata Initiative (ISO 15836)<sup>46</sup> est très largement utilisée. A la fois très générale et simple, elle regroupe en tout 15 éléments en trois classes : contenu, propriétés intellectuelles et instance particulière. Dans le cadre des règles d'implémentation de la directive INSPIRE, cette norme a été considérée avec une attention particulière, pour sa capacité à disséminer les métadonnées à différentes communautés, et en particulier dans l'e-gouvernement, et à effectuer des recherches 'de base' sur des ressources géospatiales.

(INSPIRE-2008) analyse les éléments de la norme Dublin Core et les affinements nécessaires pour représenter les métadonnées concernant les données spatiales et les services. A côté de nombreuses concordances possibles, certains concepts, comme la résolution spatiale, ne sont pas supportés dans DC, et d'autres, comme la date de la saisie de la métadonnée ou le point de contact, requièrent la construction d'une ressource supplémentaire.

Dans certaines disciplines les métadonnées sont très répandues, pour décrire les différents "objets du métier". On citera, par exemple :

- en santé, les Health care client identification et Health care provider identification Data Set Specifications, standards enregistrés en mai 2005 par le National Health Information Group (NHIG)<sup>47</sup>, en Australie
- en géographie, la norme ISO 19115<sup>48</sup>, qui est la plus suivie actuellement pour décrire des données géographiques de tout type.

(Servigne *et al.*, 2005) analyse en détail l'état de l'art et les projets de standards concernant les métadonnées et la qualité, pour souligner (page 243) que "le point de vue le plus répandu et utilisé est actuellement celui du producteur de données, que ceux-ci soient institutionnels ou occasionnels. ... Le concept de 'fitness for use', bien qu'il ne soit pas récent, n'est pas réellement mis en œuvre... Les outils, voir les normes d'évaluation, sont encore à définir, mais ce nouveau challenge est une étape incontournable dans l'avenir des Systèmes d'Information Géographique".

---

<sup>46</sup> DC, Dublin Core Metadata Initiative: <http://dublincore.org/>

<sup>47</sup> NHIG, National Health Information Group : <http://www.ahic.org.au/nhig>

<sup>48</sup> ISO 19115 : <http://www.iso.org/iso/en/CatalogueDetailPage.CatalogueDetail?CSNUMBER=26020>

## 4.4 Vue systémique et Métaphore de l'impédance

### 4.4.1 Vue systémique de NARI

Une *vue systémique* globale conduit à considérer la Recherche d'Information en relation avec les fonctions qui vont de la collecte de données à l'utilisation des données et leur diffusion. Ainsi, en adoptant une perspective d'ingénierie des systèmes logiciels, NARI doit interagir avec les systèmes producteurs et les systèmes utilisateurs des données, et comprendre les règles propres aux métiers et les contraintes organisationnelles (légales, sociales, économiques, ...) de ces organisations (productrices et utilisatrices). En particulier, nous nous intéressons aux objectifs des utilisateurs.

En tant que système, NARI peut être vu comme un "ensemble organisé de composantes intercorrélées qui accomplissent ensemble des tâches pour atteindre un objectif commun" (t.l., Sommerville, 2004, p. 21)<sup>49</sup>. Il s'agit ainsi d'un système "sociotechnique", à quatre niveaux, associés respectivement aux processus métier, aux logiciels applicatifs, aux logiciels de support et au matériel. En général, les changements apportés à un niveau du système se répercutent largement sur les niveaux adjacents. Ce phénomène est particulièrement critique, car beaucoup de systèmes (aussi bien "producteurs" que "utilisateurs"), dans les grands organismes et surtout dans les administrations, sont des systèmes propriétaires ("legacy systems").

### **Pertinence de la vue systémique pour le champ d'illustration**

La directive INSPIRE, contribue, à notre avis à introduire une vision systémique reliant les systèmes de production de données et les systèmes décisionnels, notamment au travers de deux dimensions :

- Les données, car elle s'applique à l'ensemble des données géographiques sous forme électronique. 34 thèmes sont regroupés en trois catégories de données géographiques : (1) référentiels de coordonnées, systèmes de maillage géographique, dénominations géographiques, unités administratives, ...; (2) altitudes, occupation de terres, géologie, ... ; (3) unités statistiques de diffusion ou d'utilisation en liaison avec d'autres informations statistiques, usage des sols, *santé et sécurité des personnes*, services d'utilité publique et services publics, zones de gestion, ...

---

<sup>49</sup> Les objectifs de NARI sont décrits dans le § 5.1.

- Les acteurs, car elle concerne (i) toutes les autorités publiques, aux niveaux des gouvernements et des administrations publiques nationales, régionales ou locales, ainsi que (ii) toutes les personnes physiques et morales exerçant des fonctions d'administration publique (tâches, activités, services) en rapport avec l'environnement ou (iii) ayant des responsabilités ou fournissant des services en rapport avec l'environnement.

## 4.4.2 Métaphore de l'impédance

### 4.4.2.1 Concepts généraux

Le concept d'écart d'impédance est classique en physique, pour étudier les problèmes qui se posent lorsqu'on essaye de brancher dans un seul « circuit » deux systèmes hétérogènes, qu'ils soient électriques, hydrauliques ou mécaniques.

Dans les circuits physiques, la notion d'impédance est liée à des grandeurs physiques, Intensité et Voltage, et à des propriétés, Résistance et Réactance.

L'ajustement de l'écart d'impédance nécessite des interactions entre les deux systèmes, une sorte de courant alternatif, et de mécanismes divers. Ces mécanismes sont choisis en fonction des objectifs, qui peuvent être :

- (i) *égalisation* des impédances des deux systèmes,
- (ii) *pontage*, consistant à viser des accords sur des portions fines des systèmes,
- (iii) *transformations*, par des décompositions, agrégations, désagrégations, filtrages.

Le Chapitre 4 décrit en détail la métaphore de l'impédance, dont R. Jeansoulin est à l'origine, pour explorer les analogies entre :

- a. d'une part les problèmes et solutions d'interopérabilité entre deux systèmes physiques
- b. d'autre part les problèmes et solutions d'interopérabilité entre systèmes producteurs de données et systèmes utilisateurs de ces données.

Guidés par cette métaphore, on analyse des questions liées aux écarts d'impédance, sur les deux aspects clés de l'hétérogénéité et de la qualité des données. Ces analyses préliminaires, de type qualitatif, conduisent à reformuler le problème de la recherche d'information, à définir une démarche en trois étapes pour NARI et concevoir des modules logiciels pour réduire ces écarts dès les toutes premières étapes (§5.1).

Les mécanismes pour mettre en œuvre la réduction de l'écart d'impédance ont un coût, qui se justifie selon les contextes par la criticité des besoins d'usabilité et les volumes des données disponibles. Nous pourrions aborder ces justifications, ainsi que la conception des modules de coordination de NARI, esquissés dans l'article présenté dans le Chapitre 5, dans des travaux de recherche futurs, portant sur des aspects quantitatifs de l'impédance.

#### **4.4.2.2 Travaux de recherche concernant l'écart d'impédance**

A notre connaissance, l'écart d'impédance a été peu traité en informatique, sans faire référence explicite aux similitudes entre des grandeurs physiques et logiques et les mécanismes associés.

Un bien connu « impedance mismatch » existe entre bases de données relationnelles et orienté-objets (Ambler, 2001). Au niveau des systèmes d'information, (Castro *et al.*, 2002) qualifie de 'impedance mismatch' l'écart entre le système 'nominal' et l'environnement opérationnel d'un Système d'information : le système nominal est déterminé en fonction d'un objectif, dont les concepts diffèrent de ceux de l'environnement opérationnel. Il propose une approche guidée par les objectifs, basée sur la première expression des besoins par les divers acteurs. L'accent de ces travaux est mis sur la définition des besoins de qualité, en les transformant du niveau organisationnel au niveau opérationnel. NARI étend l'utilisation de la métaphore pour analyser des hétérogénéités, dans les champs de l'information géographique et de la santé, entre les données produites et les données souhaitées pour des contextes spécifiques.

Les travaux de J. Levesque (Levesque, 2007) apportent une analyse approfondie des coûts liés à l'adaptation des demandes des utilisateurs en fonction des ressources géo spatiales disponibles, ainsi que des stratégies possibles pour réduire ces coûts.

## 4.5 Intégration

Dans le but de construire une vue systémique de plusieurs systèmes autonomes, nos recherches focalisent sur *l'intégration*<sup>50</sup>, qui est une forme d'interopérabilité, pour générer des systèmes multi-sources<sup>51</sup> qui s'accorde avec la métaphore de l'impédance et la vision systémique de NARI<sup>52</sup>.

D'une façon générale, l'intégration combine différents éléments d'un système dans une relation qui fonctionne comme un tout.

### **Pertinence de l'intégration par rapport au champ d'illustration**

L'intégration s'accorde, à notre avis, avec les préconisations de la directive INSPIRE, qui définit l'interopérabilité comme la possibilité de combiner des séries de données géographiques et de faire interagir des services, sans intervention manuelle répétitive, de telle façon que le résultat soit cohérent et la valeur ajoutée des séries et des services de données soit renforcée. L'infrastructure d'interopérabilité préconisée par INSPIRE implique la mise en œuvre de règles<sup>53</sup> concernant notamment un cadre commun international pour identifier des objets géographiques (définis dans le § 4.1.1). Ce cadre doit être de référence pour les moyens d'identification nationaux.

L'intégration s'accorde aussi avec les besoins de pilotage des applications décisionnelles en santé et avec les spécificités de l'information de santé (confidentialité, criticité, ...). Un problème majeur de l'intégration en santé est lié à l'identification des patients et des structures. Ces identifications sont nécessaires à la construction de vues cohérentes, complètes et agrégées, des données de santé (au niveau des individus, des praticiens, des structures de soins et des services de soins réalisés, ...), destinées aux usages les plus diversifiés.

---

<sup>50</sup> Dans ce rapport on ne discute pas d'autres formes d'interopérabilité, basées sur un couplage plus faible entre les systèmes existants et le système cible, par exemple basées sur des services d'échanges.

<sup>51</sup> Ici et dans la suite, le terme « système multi-source » représente un système obtenu à partir de plusieurs systèmes hétérogènes et distribués.

<sup>52</sup> D'autres formes de systèmes multi-sources, jugées pas pertinentes avec les recherches ici présentées, sont les systèmes de réplication et les systèmes Peer-to-Peer.

<sup>53</sup> IRD-INSPIRE

[http://www.ird.fr/informatique-scientifique/methodo/standards/legislation/directives\\_europeennes/inspire/](http://www.ird.fr/informatique-scientifique/methodo/standards/legislation/directives_europeennes/inspire/)

### 4.5.1 Formes d'intégration

Les architectures d'intégration et les taxonomies de solutions prennent en compte plusieurs critères, comme le type de couplage entre les schémas locaux et le schéma global, la matérialisation de ces schémas, le nombre de sources locales (Elmargamid *et al.*, 1998).

Nous synthétisons ci-dessous deux approches d'intégration de données pour générer des systèmes de RI multi-sources, la médiation et les entrepôts de données, en présentant succinctement des aspects en relation avec les objectifs généraux de nos recherches.

#### 4.5.1.1 Médiation

La médiation est une forme d'intégration où le schéma global n'est pas matérialisé. Elle est adaptée aux contextes où la préparation des réponses aux interrogations est construite à la demande et le nombre de sources est important, comme dans le cas des recherches au travers de l'Internet.

Une requête formulée à partir du schéma externe est décomposée en requêtes partielles. Celles-ci sont envoyées à d'autres médiateurs ou réparties entre plusieurs sources de données et le résultat final est obtenu par fusion des résultats partiels. Des extracteurs ou traducteurs (*wrappers*) au dessus des sources de données traitent au préalable les problèmes d'hétérogénéité physique (Figure 3).

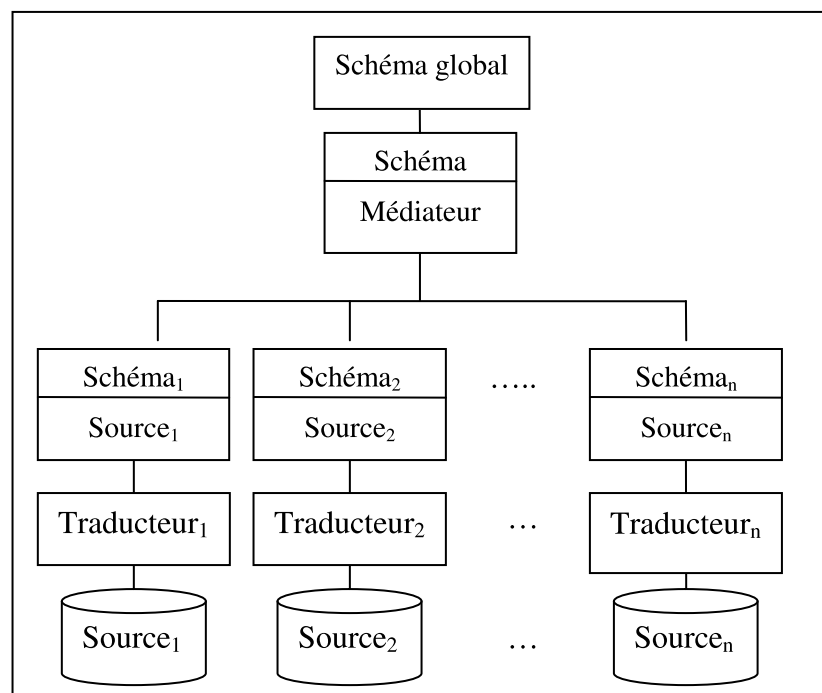
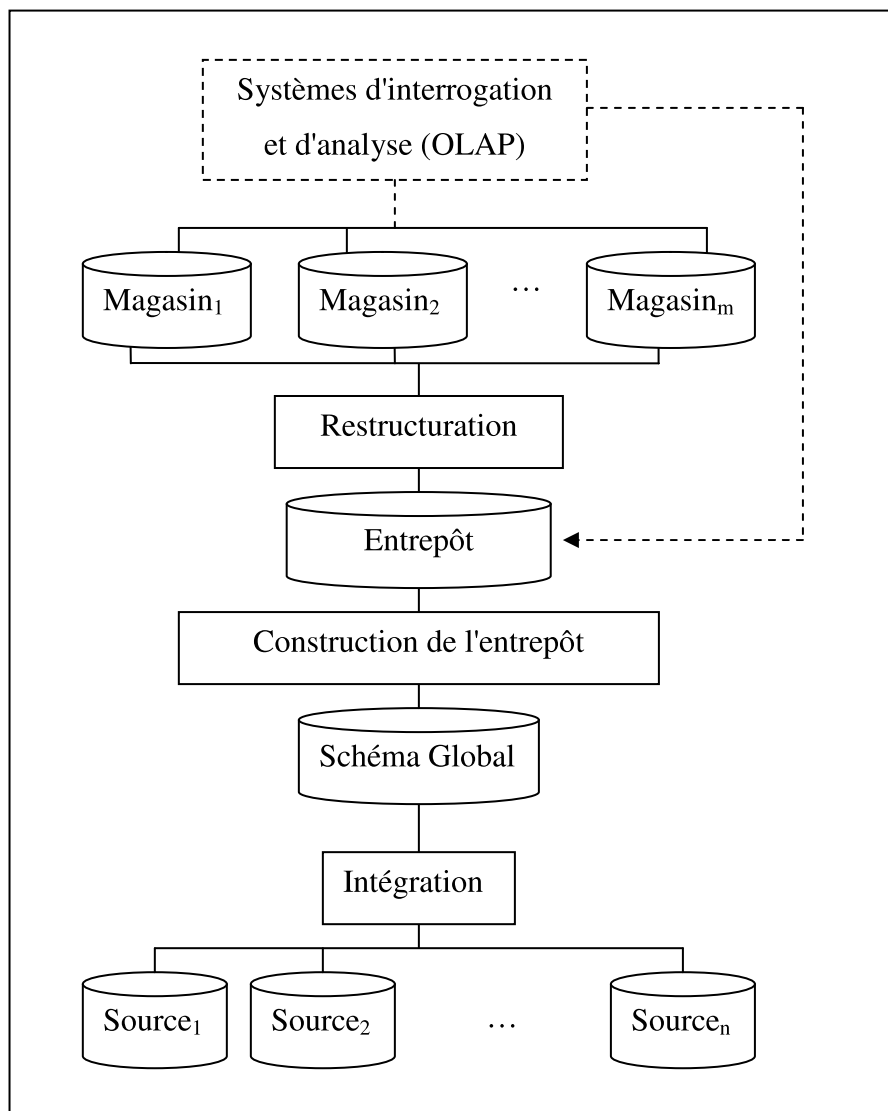


Figure 3 : Médiation de bases de données hétérogènes

#### 4.5.1.2 Entrepôts de données

Les entrepôts de données apportent aux entreprises une solution orientée vers l'aide à la décision, pour exploiter de façon efficace et performante d'énormes masses de données, hétérogènes et distribuées. Un entrepôt de donnée (ED) est défini comme une "collection de données intégrées, orientées sujet, non volatiles, historisées, résumées et disponibles pour l'interrogation et l'analyse" (Inmon, 2002).

Son architecture est schématisée dans la Figure 4.



**Figure 4 :** Architecture d'un entrepôt de données, inspirée de (Cauvet et Rosenthal-Sabroux, 2001)

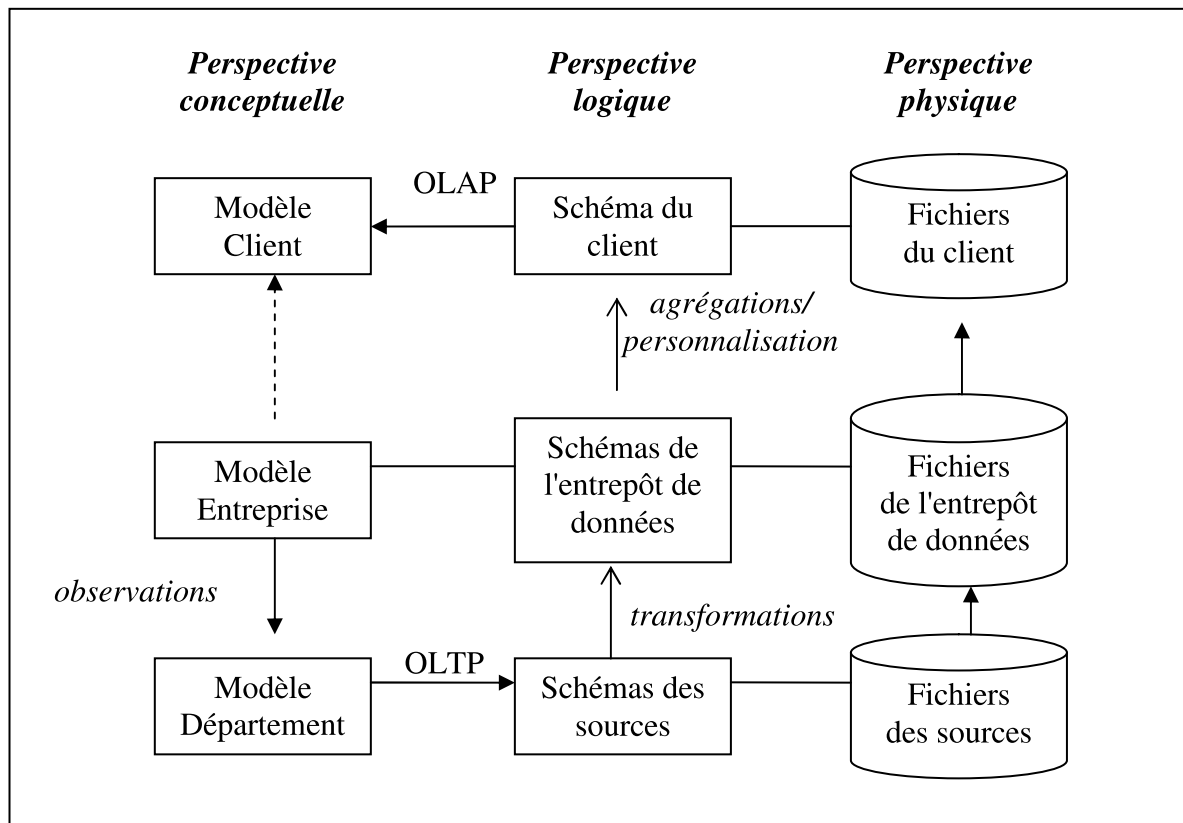
L'alimentation et la mise à jour des données de l'ED se font au travers d'extractions à partir des bases de productions et de saisies de données extérieures à l'entreprise, concernant par exemple la législation.

La construction d'un ED, comme des systèmes de médiation, s'appuie sur des paradigmes et outils spécifiques (Kimball, 2002 ; Proulx *et al.*, 2007). Les étapes générales d'intégration sont décrites dans le § 4.5.3.1.

L'entrepôt de données et les magasins (*datamarts*) répondent à des exigences qui se situent respectivement au niveau global, de l'entreprise, et aux niveaux de classes de décideurs ou d'usages particuliers. Ils sont exploités par des systèmes d'analyses de données en ligne (OLAP), interactifs et performants. Des techniques de plus en plus évoluées (*datamining* ou fouille de données), intégrant des analyses spatiales, sont désignées avec le terme SOLAP (Spatial On line Analytical Processin) (Bédard *et al.*, 2007).

#### **4.5.2 Qualité externe, métadonnées et intégration**

Nous interprétons la qualité externe par rapport au cadre formel des métadonnées des entrepôts de données défini dans le projet européen Esprit DWQ (Data Warehouse Quality) (Jarke *et al.*, 2000). DWQ analyse la qualité des entrepôts de données selon trois perspectives (conceptuelle, logique et physique) et trois niveaux (Opérationnel, Entreprise et Utilisateur final). Il relie la qualité des données, et ses mesures, aux objectifs des différents acteurs, situés à chaque intersection perspectives/niveaux. Par exemple, les préoccupations du concepteur ou de l'administrateur de l'entrepôt se situent essentiellement au niveau de la qualité des schémas (justesse, complétude, ...) et de leurs évolutions, tandis que les critères concernant les utilisateurs finaux se rapportent principalement à l'usabilité, vue en termes d'accessibilité (disponibilité du système, sécurité et droits d'accès) et d'utilité (pertinence, fraîcheur, interprétabilité des données).



**Figure 5 :** Cadre de référence DWQ des métadonnées d'entrepôt de données, adapté à partir de (Jarke *et al.*, 2000)

Ce métamodèle de référence a été généralisé par le modèle (Akoka *et al.*, 2007), qui est discuté dans le § 4.2.1.

Dans le § 5.4 Discussion, nous positionnons NARI par rapport aux techniques d'intégration et à ce cadre de référence.

### 4.5.3 Étapes et techniques d'intégration

#### 4.5.3.1 Étapes d'intégration

Les étapes de l'intégration de bases de données, communes aux systèmes de médiation et aux entrepôts de données, ont pour objectif de construire une vue unifiée des données présentes dans les sources locales. Pour les solutions d'intégration virtuelle, ces étapes sont décrites par (Parent et Spaccapietra, 1996). Schématiquement :

- la pré-intégration définit les objectifs de l'intégration, la stratégie et l'architecture d'intégration, et spécifie les schémas des bases locales dans un langage facilitant les étapes suivantes
- la comparaison des schémas recherche les correspondances entre les schémas locaux.
- la mise en conformité des schémas analyse les conflits induits par les correspondances et propose, si possible, des solutions pour les résoudre (Kim *et al.*, 1993).
- la fusion des schémas et la restructuration ont pour objectif de construire un schéma externe intégré.

(Jarke *et al.*, 2000) spécifient et complètent ces étapes pour la construction d'entrepôts de données.

L'enrichissement sémantique et les avancées des modèles de données orientés-objet conduisent à approfondir ces étapes et à une plus grande complexité des architectures (Papazoglou *et al.*, 2000).

M. Lenzerini (2005) présente une perspective théorique de l'intégration de données par rapport aux traitements de requêtes, à l'inconsistance entre sources de données et aux inférences sur les requêtes.

#### **4.5.3.2 Techniques de description des schémas**

Les deux approches majeures pour décrire le schéma global et son intégration avec les schémas locaux des sources, appelées GAV (Global As View) et LAV (Local As View), sont étudiées et comparées dans (Halevy, 2001).

L'approche GAV définit le schéma global comme une vue globale sur les schémas sources. Utilisée par exemple dans le projet TSIMMIS (Chawathe *et al.*, 1994), elle est particulièrement adaptée si l'on suppose vérifiée l'hypothèse du monde fermé, selon laquelle d'une part toutes les sources sont connues au moment de la définition du schéma global et d'autre part l'ensemble des données interrogées correspond à l'union des données dans les sources.

Dans l'approche LAV le schéma global décrit le contexte du système cible, indépendamment des sources de données disponibles. Celles-ci sont intégrées progressivement et interprétées

comme vues sur le schéma global, comme par exemple dans le prototype STyX (Fundulaki *et al.*, 2002).

L'approche LAV est la plus pertinente pour NARI, car elle facilite l'intégration progressive de sources de données (environnementales, administratives, scientifiques) en fonction de besoins décisionnels inopinés.

#### **4.5.4 Modèles de données semi-structurées et aspects technologiques**

Les approches d'intégration sont souvent couplées aux modèles de données semi structurées, car ces modèles n'imposent pas de structure a priori dans le schéma, qui est instancié par les données elles-mêmes. La grande souplesse de ces structures est un avantage particulièrement précieux pour interroger ou générer des données issues du Web, en composant des structures pas complètement connues au départ (Abiteboul *et al.*, 1999).

Les modèles de données semi-structurées sont supportés par XML<sup>54</sup>, désormais reconnu comme le meilleur modèle semi-structuré d'échange et d'intégration de données, qui s'accompagne de très nombreux outils. XQuery<sup>55</sup>, qui s'affirme comme un standard d'interrogation du W3C<sup>56</sup> et XSLT, langage de transformation et présentation de structures, facilitent l'échange et la transformation de données semi-structurées.

Les technologies qui se rapportent à ces modèles ont connu un essor particulier ces dernières années. Elles contribuent au développement des recherches sur le Web sémantique et la gestion des connaissances. Le Chapitre 2, dédié à ces technologies, les analyse et précise les choix technologiques retenus dans NARI.

### **4.6 Démarches de conception de type CSCW**

#### **4.6.1 Aspects méthodologiques**

Les recherches en *ingénierie des besoins* visent à concilier dans le "world system", le monde réel vu comme système, les deux sous-systèmes "subject world" et "usage world", contenant respectivement un ensemble de modèles (de classes, d'activités, ...) et les activités de multiples agents (Rolland et Prakask, 2000), (DeRosa *et al.*, 2008). Dans (Freitas *et al.*, 2002)

---

<sup>54</sup> XML : <http://www.w3.org/XML/>

<sup>55</sup> XQuery : <http://www.w3.org/XML/Query/>

<sup>56</sup> W3C, World Wide Web Consortium : <http://www.w3.org/>

les objectifs organisationnels des systèmes décisionnels sont formalisés en un ensemble de questions à différents niveaux de responsabilité (national, régional, local) pour aboutir à la spécification des vues correspondantes. Cette approche est utilisée dans (Nakache, 2003) pour la conception de l'entrepôt de données de l'Assurance Maladie.

Les approches de type CSCW (Computer Supported Cooperative Work), initiées par (Grudin, 1994), combinent plusieurs approches d'ingénierie des besoins (guidés par les scénarios, les objectifs, les modèles, les agents, ...). Elles conduisent à la construction d'architectures globales et technologiques pour des systèmes coopératifs, pluridisciplinaires. (Pohl et Haumer, 1997 ; Hichey *et al.*, 1999 ; Bardram, 2000 ; Romano *et al.*, 2002 ; Darses *et al.*, 2004) décrivent et classifient des scénarios de collaboration pour guider le recueil des besoins et l'analyse des contraintes dans la conception des systèmes coopératifs<sup>57</sup>.

#### **4.6.2 Pertinence de l'approche CSCW par rapport à NARI**

Les approches CSCW sont très pertinentes par rapport à la recherche d'information en santé, car elles supportent des espaces de coopération pouvant être : (a) les mêmes, (b) différentes mais prédictibles, (c) différents et non prédictibles. Ces trois espaces de coopération se manifestent par exemple (i) parmi des spécialistes de la même discipline médicale ; (ii) parmi des spécialistes de disciplines médicales différentes ou parmi des médecins et des gestionnaires ; (iii) lors de la diffusion de l'information au grand public.

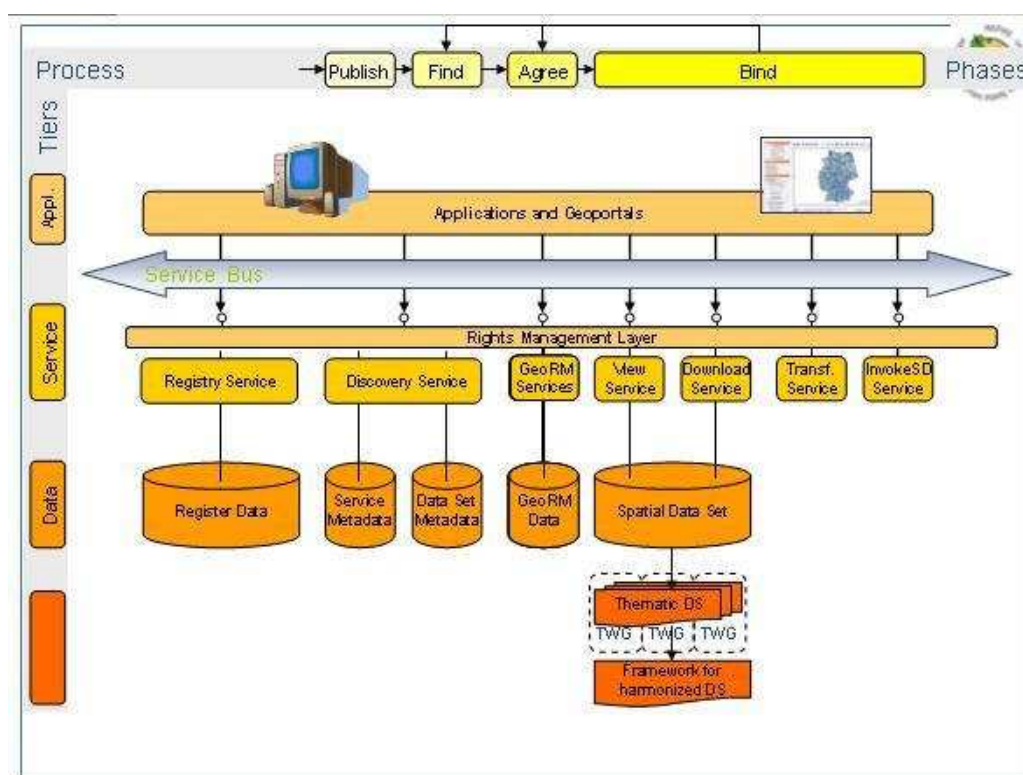
Les espaces de coopération considérés par NARI concernent des systèmes de production de données de santé et des systèmes utilisateurs au sein de services de l'Administration (par exemple, préfectures, services d'urgences sanitaires) ou d'organismes de veille travaillant pour ces services.

---

<sup>57</sup> L'intérêt pour le CSCW et pour le "groupware", qui dénote ses aspects plus techniques, ne cesse d'augmenter, comme le témoignent les multiples communautés de recherche (en bases de données, agents, ...) impliquées dans les systèmes organisationnels innovants ainsi que la diversité des applications. La synergie entre ces communautés est illustrée par l'organisation de "fédérations de conférences", comme OTM "OnTheMove" Federated Conferences and Workshops, qui regroupe plusieurs conférences internationales (CoopIS - Conference on Cooperative Information Systems, DOA - International Symposium on Distributed Objects and Applications, GADA - International Symposium on Grid computing, high-performance and Distributed Applications et ODBASE - International Conference on Ontologies, DataBases and Applications of Semantics). La conférence annuelle HICSS (Hawaii International Conference on System Sciences, qui a atteint en 2009 sa 42-ème édition), organise des larges sessions dédiées aux systèmes collaboratifs, en particulier en e-gouvernement et santé.

Les finalités des coopérations peuvent être aussi bien liées à des activités de production, comme dans les workflows, qu'à la prise de décision. Elles sont utilisées dans des applications de téléassistance et téléconsultation, pour utiliser de multiples systèmes en réseaux et faire appel aux meilleurs spécialistes quelle que soit la localisation du malade. En e-santé, des aspects critiques du travail coopératif se matérialisent dans la complexité de la conception du dossier médical personnel électronique, de systèmes d'alerte, et plus généralement de systèmes d'aide à la décision (Beuscart *et al.*, 2002), (Haux, 2006).

L'architecture INSPIRE pour le partage de données et services géographiques, représentée dans la Figure 6, peut être aussi interprétée dans le cadre d'une approche CSCW.



**Figure 6 :** Vue d'ensemble de l'architecture technique INSPIRE<sup>58</sup>

Cette architecture s'appuie sur un modèle "publier - trouver - accorder - lier", qui vise à supporter le travail coopératif entre les différents acteurs opérant sur des données et des services géographiques. Cette architecture est structurée en quatre niveaux : applications,

<sup>58</sup> INSPIRE Network Services Architecture

[http://inspire.jrc.ec.europa.eu/reports/ImplementingRules/network/D3\\_5\\_INSPIRE\\_NS\\_Architecture\\_v3-0.pdf](http://inspire.jrc.ec.europa.eu/reports/ImplementingRules/network/D3_5_INSPIRE_NS_Architecture_v3-0.pdf)

gestion des droits, services, données. Les données et services, décrits respectivement dans les Data Set Metadata et Service Metadata, et publiés dans les Register Data, sont découverts par des services ad hoc (Discovery services). Les données sont visualisées sous forme d'images avec des représentations cohérentes et un filtrage possible des données (View services) ; elles peuvent être transformées (changement de format, de géométrie, de langue, de structure de données, par des Transformation services) et ensuite téléchargées (Download services).

## 5. L'approche NARI : synthèse, analyse et discussion

Dans ce paragraphe, nous analysons l'approche désignée synthétiquement par *NARI* (*Nouvelle Approche de Recherche d'Information, basée sur la métaphore de l'impédance, la qualité et les métadonnées, structurée par la dimension géographique*).

Nous présentons la démarche de conception de NARI, avec des objectifs spécifiques (qui détaillent l'objectif global énoncé dans le § 2.4). Nous indiquons ensuite comment ces objectifs ont été progressivement atteints. Des aspects spécifiques de NARI étant détaillés dans les publications jointes dans les Chapitres 3, 4 et 5 de ce mémoire, nous présentons ici seulement une vue synthétique et globale de NARI, accompagnée d'une analyse des propriétés visées. Cette analyse bénéficie du recul acquis sur l'ensemble des travaux et prépare à la discussion sur ces recherches, et aux perspectives de recherches futures (§6).

Plus précisément, le plan détaillé de ce paragraphe est le suivant :

- 5.1 Démarche de conception de NARI et objectifs spécifiques
  - 5.1.1 Démarche de conception
  - 5.1.2 Obtention progressive des objectifs spécifiques
    - 5.1.2.1 Niveau « intégration de données »
    - 5.1.2.2 Niveau « identification / préparation de sources de données »
- 5.2 Vue synthétique de l'approche
  - Exemples d'illustration
- 5.3 Analyse des caractéristiques de NARI
  - 5.3.1 Progressivité
  - 5.3.2 Adaptabilité au contexte d'utilisation
    - 5.3.2.1 Fonctions d'adaptabilité et modèles
    - 5.3.2.2 Conception des fonctions d'adaptabilité
  - 5.3.3 Traçabilité via les classifications de requêtes
- 5.4 Concordances et originalités de NARI par rapport à d'autres approches
  - 5.4.1 Concordances
  - 5.4.2 Originalités
  - 5.4.3 Atteinte des objectifs et limites actuelles de l'approche NARI

## 5.1 Démarche de conception de NARI et objectifs spécifiques

### 5.1.1 Démarches de conception

Pour minimiser le risque lié à la non qualité du processus décisionnel supporté par la recherche d'information, nous nous appuyons sur le modèle de développement de logiciels RUP (Rational Unified Process)<sup>59</sup>, basé sur la notion de risque. Ce modèle, générique, en combine plusieurs autres. Itératif, il identifie quatre phases dans le cycle de vie des systèmes : inception, élaboration, construction, transition. Chaque phase réalise un ensemble d'activités et une activité peut se dérouler pendant plusieurs phases. Les activités macroscopiques sont : recueil des besoins, analyse, conception, implémentation, tests, déploiement et gestion de projet.

Selon ce modèle, les étapes en "amont" nécessitent le plus d'attention et de bonnes pratiques : (i) identification des acteurs impliqués et leurs rôles ; (ii) recueil et analyse des besoins et contraintes liés aux activités, ainsi que des besoins non fonctionnels, tels que l'usabilité, la performance et la sécurité ; (iii) élaboration de l'architecture globale. En focalisant sur les étapes en amont, nous choisissons de :

1. bâtir les modèles conceptuels de l'approche NARI elle-même et du domaine d'application, respectivement sur une ontologie d'application et une ontologie de domaine, construits à partir de l'analyse des exigences de cas concrets
2. bâtir l'architecture globale de NARI, en choisissant les techniques, technologies et outils les plus adaptés, identifiés par rapport à l'état de l'art (§ 4)

Ainsi, nous déclinons l'objectif général de nos recherches (§ 2.4) dans des objectifs spécifiques, détaillés dans la Table 2.

OS1	<b>Analyser des scénarios de RI en santé</b> représentatifs de cas d'échecs ou d'évolutions significatives, avec une perspective d'ingénierie des besoins. Concevoir une ontologie de domaine, focalisée sur les données (hétérogénéité, qualité)
OS2	<b>Proposer des éléments de conception de NARI</b> , concernant : <ul style="list-style-type: none"><li>▪ l'ontologie d'application et la spécification des étapes</li><li>▪ l'architecture globale : niveaux, composantes fonctionnelles, approches techniques</li></ul>
OS3	<b>Proposer des éléments de conception de NARI</b> , concernant les langages de formalisation des connaissances et les outils associés.

**Table 2** : Objectifs spécifiques pour la conception de NARI.

<sup>59</sup> RUP, Rational Unified Process : <http://www-01.ibm.com/software/awdtools/rup/>

### 5.1.2 Obtention progressive des objectifs spécifiques

L'obtention des objectifs s'est faite par étapes, qui ont conduit à approfondir successivement deux différents niveaux :

- intégration de données
- identification / préparation des sources de données à intégrer.

L'article présenté dans le Chapitre 3 focalise sur le premier niveau, tandis que les articles présentés dans les Chapitres 4 et 5 focalisent sur le deuxième niveau. La progression dans la conception de ces deux niveaux est synthétisée dans la table ci-dessous, où nous associons les objectifs spécifiques aux aspects traités plus particulièrement par les articles. Nous décrivons cette progression dans les paragraphes suivants.

Objectif spécifique	Aspects traités dans les articles
OS1- (i) Analyser des scénarios de RI en santé. (ii) Concevoir une ontologie de domaine, focalisée sur les données (hétérogénéité, qualité).	(i) Tout le long des articles (ii) Esquisse d'ontologie de domaine : données hétérogènes, distribuées, géo-référencées, cataloguées (Ch. 3 et 4). Elaboration de deux exemples simples : le premier est présenté dans les Ch. 3 et 4 et le deuxième dans le Ch. 5.
OS2 - Proposer des éléments de conception de NARI, concernant l'ontologie d'application, les étapes et l'architecture globale	- Esquisse d'ontologie d'application et d'architecture globale d'intégration (Ch. 3) - Introduction de la métaphore de l'impédance et approfondissement de l'ontologie d'application et de l'architecture globale (Ch. 4 et 5)
OS3 - Proposer des éléments de conception de NARI, concernant les langages de formalisation des connaissances et les outils associés	Chapitre 2 et tout le long des chapitres suivants. Techniques d'intégration et outils associés (Ch. 3) Logiques de description et outils associés (Ch. 4) Langages à base de règles et outils associés (Ch. 5)

**Table 3 :** Association entre les objectifs spécifiques de l'approche NARI et les aspects traités dans les articles.

#### 5.1.2.1 Niveau « intégration de données »

Dans l'article présenté dans le Chapitre 3 nous analysons des systèmes coopératifs couplant santé et géographie et participant des objectifs de développement durable. Nous menons cette analyse avec une approche systémique. Elle conduit à identifier deux fonctions essentielles de l'approche, qui se situent au niveau de l'intégration des schémas et des contenus : (i) Interroger les sources locales à partir d'une vue globale, transversale ; (ii) Évaluer la qualité des sources locales par rapport au contexte applicatif.

Nous identifions trois aspects critiques pour la conception, liés respectivement à la complexité du cadre normatif, aux volumes des sources de données et à l'hétérogénéité de ces sources.

Nous réalisons aussi des choix structurants concernant l'architecture globale, les langages de représentations des connaissances, basés sur les logiques de description (Calvanese *et al.*, 2004) et les outils associés.

L'architecture globale comprend à ce stade un niveau médiateur et un niveau local :

- (i) le premier contient une ontologie de domaine, spécifiée indépendamment des sources de données disponibles selon une approche LAV (Local As View) ; un schéma global, généré à partir de cette ontologie, utilisé pour formuler les requêtes suivant une technique académique de médiation (Amann *et al.* 2002) ; une description des sources de données cataloguées enrichie par les métadonnées décrivant les critères de qualité ;
- (ii) Au niveau local, on suppose que les sources de données soient disponibles au format XML.

Ces choix sont illustrés sur un exemple très simplifié de gestion de risques, avec des interrogations concernant (i) l'existence/absence de sources de données vérifiant des critères de qualité et (ii) nécessitant l'intégration de données à partir de plusieurs sources de données.

### **5.1.2.2 Niveau « identification / préparation des sources de données »**

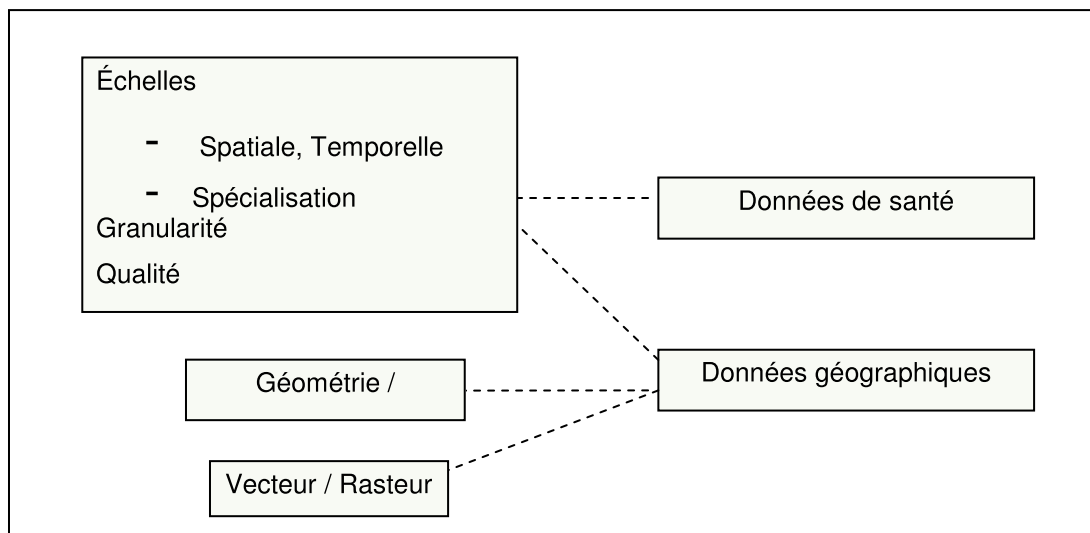
#### *Introduction de la métaphore de l'impédance*

L'article présenté dans le Chapitre 4 développe une idée originale de R. Jeansoulin en introduisant la métaphore de l'impédance.

Cette métaphore permet d'aborder l'analyse de l'écart entre les données obtenues et les données attendues dans un système de RI qui relie un ensemble de systèmes producteurs de données et un ensemble de systèmes utilisateurs de ces données. Elle conduit à reformuler l'expression des besoins préliminaires des utilisateurs en adaptant les questions en cas d'échec. On anticipe le plus possible l'analyse et la résolution des écarts d'impédance entre les systèmes. Ces écarts sont dus à des problèmes d'hétérogénéité et de qualité, interprétés comme des *résistances* à la Recherche d'Information. La Figure 7 schématise les problèmes d'hétérogénéité qui peuvent affecter les données de santé et les données géographiques.

La métaphore de l'impédance conduit aussi à une réduction progressive des écarts. Nous introduisons trois étapes de la Recherche d'Information : (1) Catalogues, (2) Sources, (3) Données, et nous nous intéressons à trois aspects : (i) existence de données pertinentes pour les besoins ; (ii) qualité des données suffisante pour ces besoins ; (iii) cohérence des données elles-mêmes et la capacité à les traiter. On assimile à des *réactances* les modules qui traitent de ces aspects aux différentes étapes.

La géolocalisation, qu'elle soit directe ou indirecte, basée sur les adresses postales (cfr. § 4.1.1) est essentielle pour traiter les questions d'existence de sources de données : en effet, en enrichissant la description des objets par des coordonnées géographiques, il est possible ensuite de tirer profit des techniques et des outils d'analyse spatiale pour trouver des données en fonction de ces coordonnées.



**Figure 7 :** Exemples d'hétérogénéité affectant les données de santé et les données géographiques

### *Approfondissements*

Dans l'article présenté dans le Chapitre 5, rédigé conjointement avec R. Jeansoulin, l'analyse d'applications d'e-gouvernement conduit à développer davantage la métaphore de l'impédance. Nous structurons et classifions divers modules logiciels qui réduisent l'écart d'impédance entre les données produites et les données attendues :

- nous focalisons sur les aspects concernant l'existence et la qualité de catalogues et de sources de données pertinentes pour les requêtes.

- nous approfondissons la conception de l'ontologie d'application, centrée sur les territoires, en introduisant des relations géographiques et de similarité pour tirer profit des analyses spatiales,
- nous apportons des éléments supplémentaires pour concevoir les modules opérant sur les requêtes (décompositions, transformations, applications de modèles de similarité, coordination)
- en plus des logiques de description, nous utilisons les règles, pour la formalisation des connaissances
- nous affinons la mesure de la qualité externe des données, par rapport à l'usabilité : au lieu des classifications binaires (oui/non) décrites dans les Chapitres 3 et 4, nous établissons des classifications en plusieurs niveaux.

## 5.2 Vue synthétique de l'approche

On suppose connu un ensemble de multiples sources de données, cataloguées, hétérogènes et géo-référencées. On formule des requêtes à partir du schéma global du domaine, construit indépendamment des sources, par une approche de type Local As View. Les requêtes sont des conjonctions d'atomes, du type :

$$Q(x_1, x_2, x_3) := \text{Personne\_Agée\_Dépendante}(x_1) \\ \text{ET Département}(x_2) \text{ ET Année } (x_3) \text{ ET Habite } (x_1, x_2, x_3) \\ \text{où } X(x_1, x_2, \dots, x_n) \text{ est une variable dont l'utilisateur veut obtenir des instances.}$$

Une requête est complétée par des contraintes sur la qualité des données, du type  
C : Fraîcheur des données  $\geq 1998$

Afin de réduire l'écart parmi les sources disponibles et les vues souhaitées par l'utilisateur, l'approche NARI s'appuie sur la métaphore de l'impédance, détaillée dans le Chapitre 4. Cette métaphore suggère des analogies entre l'interopérabilité de systèmes physiques et l'interopérabilité de systèmes d'information et contribue à définir des éléments de conception pour les composantes de NARI.

On considère trois aspects de la recherche d'information :

- (RI1) existence de données pertinentes dans les catalogues ;

- (RI2) qualité des sources de données, en corrélation avec les contraintes associées aux requêtes ;
- (RI3) contenu des sources de données, pour une utilisation consistante et effective des données.

Les traitements des deux premiers aspects, (RI1) et (RI2), concernent le niveau « identification / préparation des données ». Ils s'appuient exclusivement sur les métadonnées des catalogues et des sources de données, qui apportent respectivement des descriptions générales (recouvrement, formats, ...) sur la pertinence des sources par rapport aux critères d'interrogation, ainsi que leur localisation et qualité, et plus fines sur le contenu des sources.

Les traitements du troisième aspect (RI3) concernent le niveau « intégration de données ».

L'architecture globale de l'approche, synthétisée dans le § 4.2 du Chapitre 2 et détaillée dans Chapitre 5, comprend trois niveaux. Le niveau "application" contient :

- une base de connaissances, bâtie sur l'ontologie d'application, avec des classes correspondantes aux dimensions Thème, Espace et Temps des requêtes. Des relations sémantiques, spatiales et temporelles permettent d'explorer des solutions au manque de données ou à la faible qualité des données, de même que l'application de modèles de similarité thématiques et spatiales
- un système de raisonnement, pour déterminer des instances de spécialisations des classes
- un système à base de règles, pour inférer de nouveaux catalogues et sources de données associés à une requête et classer des requêtes par rapport à des critères de qualité.

Les niveaux médiateur et local sont détaillés dans le Chapitre 3.

Les ontologies sont spécifiées avec le formalisme de la Logique de Description (DL, *SHOIN(D)*) (Calvanese *et al.*, 2004). L'implémentation technique s'appuie sur les systèmes Protégé (Knublauch *et al.*, 2004) et Pellet (Sirin *et al.*, 2007).

### **Exemples d'illustration**

Sans rentrer dans les détails, on synthétise des exemples d'illustration de NARI. Il s'agit d'exemples mettant en évidence le manque de données relevant du domaine de la santé, sur

des territoires, et/ou leur faible qualité par rapport au critère de la fraîcheur des données, requis pour la prise de décision.

Les articles des Chapitres 3 et 4 présentent un exemple visant à déterminer la disponibilité et la qualité des sources nécessaires à la gestion de certains risques (canicule, vagues de froid, inondations). Ces analyses s'appuient essentiellement sur les métadonnées. Pour chaque source de données, on fixe le seuil acceptable pour le critère de fraîcheur. On détermine pour chaque risque :

- (i) si chaque source définie comme nécessaire est disponibles ou pas
- (ii) si chaque source nécessaire et existante vérifie ou pas les critères de qualité.

Ces critères sont utilisés pour classer les risques (décrits ou pas) et la qualité de la description (satisfaite ou pas).

L'article du Chapitre 5 illustre NARI avec un exemple qui corrèle sur des territoires géographiques des demandes de services, par exemple pour des Personnes Agées Dépendantes (ODP), avec l'offre de services (hôpitaux, lits, ...). Les données concernant les populations sont connues sur chaque département de chaque région et chaque année. Par contre, sur une région donnée (R2), et une année (2005), les données concernant les ODP sont connues sur un département (D21) et absentes sur un autre (D24). En correspondance de la requête « Donner le nombre de ODP sur D24 pour 2005 », NARI :

- détecte, à l'aide des métadonnées, l'absence des sources de données requises sur la région
- explore l'existence de sources de données relatives à un thème plus général que « ODP » sur la même région
- explore l'existence de sources de données relatives au thème « ODP » sur un département inclus dans la région

Comme les résultats de ces explorations sont positifs, NARI calcule un résultat approché pour la requête, en appliquant un modèle de similarité sur la « population » entre D24 et D21 et le rapport de proportionnalité entre ODP et « population » connu sur D21 ; enfin on classe la requête d'origine comme partiellement satisfaite et la requête approchée comme complètement satisfaite.

### **5.3 Analyse des caractéristiques visées pour NARI**

On analyse ici les principales caractéristiques de l'approche NARI, à savoir : progressivité, adaptabilité au contexte de l'utilisation et traçabilité.

#### **5.3.1 Progressivité**

De par la complexité et les volumes des données en santé, la progressivité est reconnue comme une exigence prioritaire en recherche d'information (Brunie *et al.*, 1998). Une démarche progressive basée sur les métadonnées, comme dans l'exemple MDWEB (Desconnets *et al.*, 2003), s'adapte aussi aisément aux normes géographiques ISO 19113-115 et à la directive européenne INSPIRE (INSPIRE, 2008) pour la diffusion des données environnementales.

Dans l'approche NARI, l'exigence de progressivité est supportée par les traitements sur les métadonnées (RI1 et RI2), avant d'accéder aux données elles-mêmes (RI3). La progressivité de NARI n'est pas nécessairement séquentielle, car les explorations des métadonnées et des données se font en interaction avec l'utilisateur, qui oriente, poursuit ou arrête le processus d'adaptabilité, décrit ci-dessous.

#### **5.3.2 Adaptabilité au contexte d'utilisation**

L'expression "adaptabilité au contexte d'utilisation" est la traduction de l'expression "context awareness", qui caractérise la capacité d'un système à fournir des informations et des services pertinents pour l'utilisateur, en fonction de son contexte. Schématiquement, on identifie trois aspects de l'adaptation au contexte, concernant respectivement la présentation, les traitements et le stockage de l'information. (Chaari *et al.*, 2007) définissent le contexte comme "l'ensemble de paramètres externes à l'application et pouvant influencer sur le comportement de l'application en définissant des nouvelles vues sur ses données et ses services". Ces auteurs mettent en œuvre et illustrent cette conceptualisation de l'adaptabilité sur une application en néphrologie pour le suivi de patients dans le cadre de l'hospitalisation à domicile.

La complexité de l'adaptabilité de contenu, services et présentation aux contextes médicaux de différents pays est illustrée par exemple par le système décisionnel en épidémiologie SIMS-REIN (Ben Saïd *et al.*, 2005).

### 5.3.2.1 Fonctions d'adaptabilité et modèles

Dans l'approche NARI, on vise deux types de fonctions d'adaptabilité :

- (A1) construire, en cas d'échec d'une requête, des requêtes approchées, valides pour le contexte de l'utilisateur ;
- (A2) guider globalement la recherche d'information, depuis la formalisation de la requête jusqu'à la préparation et l'ordonnancement des résultats et l'évaluation de leur usage effectif.

Le support de l'ensemble des objectifs nécessiterait plusieurs modèles aptes à représenter :

- les utilisateurs (MA1) ; les ressources (MA2) ; les producteurs (MA3), ainsi que la proximité (ou la distance) entre l'utilisation et la production de données via les ressources (utilisateur <-- ressources --> producteur) ;
- le domaine d'application (MA4) ;
- la gestion des requêtes (MA5) (formulation, décomposition, transformation, coordination des décompositions et transformations, préparation des résultats) ;
- les interactions entre les utilisateurs et le système de RI (MA6) et entre les producteurs et le système de RI (MA7).

Dans NARI, (MA4) et (MA2) sont représentés respectivement par l'ontologie du domaine et par les schémas XML des ressources, augmentés des métadonnées. Des éléments de conception pour (MA1), (MA5) et (MA6) sont identifiés ci-dessous.

### 5.3.2.2 Conception des fonctions d'adaptabilité

Les fonctions (A1) de l'adaptabilité sont traitées par les modules de décompositions et transformations des requêtes. Ces modules, spécifiés en DL, sont chargés des actions suivantes : (i) la recherche, selon des dimensions spécifiques, de catalogues (resp. sources) utilisables pour une requête ; (ii) la transformation d'une requête ayant généré un échec en une requête approchée ; (iii) la coordination de ces activités.

<i>Nom</i>	<i>Input</i>	<i>Output</i>	<i>Traitement</i>
UCS <sub>X<sub>1</sub></sub>	Q Catalogue	UC <sub>X<sub>1</sub></sub>	UC <sub>X<sub>1</sub></sub> (C) ← Catalogue(C) ∧ subject(C,x <sub>1</sub> )
Transformer <sub>N<sub>1</sub> X<sub>1</sub></sub>		Q', x' <sub>1</sub> , UC <sub>X<sub>1</sub></sub>	UC <sub>X<sub>1</sub></sub> (C) ← Catalogue(C) ∧ subject(Q,x <sub>1</sub> ) ∧ subject(C,x' <sub>1</sub> ) ∧ RelationSémantique(x <sub>1</sub> ,x' <sub>1</sub> )

**Table 4 :** Exemples de décomposition et transformation d'une requête Q (N1 = niveau Catalogues et X1=dimension thématique).

La Table 4 illustre la spécification de deux modules, UCS<sub>X<sub>1</sub></sub> et Transformer<sub>N<sub>1</sub>X<sub>1</sub></sub> dont les fonctions sont respectivement : recherche de Catalogues Utilisables (UC) selon la dimension X<sub>1</sub> (thème), et transformation d'une requête Q en cas d'échec au niveau Catalogues, selon cette dimension. Les spécifications sont basées sur l'exploration de l'ontologie de domaine pour rechercher des catalogues dont les métadonnées référencent des thèmes similaires aux thèmes de la requête initiale Q.

Des expressions analogues régissent les spécifications des modules relatifs aux dimensions de recherche 'espace' et 'temps' (i=2 et 3) et le niveau 'Sources' (N<sub>2</sub>).

### 5.3.3 Traçabilité via les classifications de requêtes

La traçabilité contribue à la qualité du processus global d'adaptabilité d'un système de RI. Dans NARI, elle a deux objectifs, orientés respectivement vers des aspects d'évaluation du système et des aspects d'architecture (performances) :

- (T1) évaluer globalement la qualité du processus de RI en santé (production des données, utilisation effective des résultats fournis, description du domaine,...) ;
- (T2) anticiper sur la construction de réponses, en s'appuyant sur des questions similaires déjà traitées, des sources de données approchées déjà préparées.

L'intérêt de (T1) est prouvé par de nombreux efforts aussi bien en santé qu'en géographie, portant sur la normalisation (HL7, ISO 19115), sur les évolutions des thésaurus et la construction d'ontologies.

## Classifications de requêtes

A différence des approches basées sur l'analyse des sessions des utilisateurs (Badue *et al.*, 2005), dans NARI (T1) s'appuie sur des classifications de requêtes qui mémorisent les besoins préliminaires des utilisateurs et les requêtes, exactes ou approchées, qui ont été construites.

Les classifications des requêtes s'appuient sur les modèles des ressources (MA2), du domaine (MA4) et de la gestion des requêtes (MA5), réalisée par le processus d'adaptabilité. Comme illustré dans la Table 5, par rapport à une instance  $x_i$  de la dimension  $X_i$  (thématique, spatiale, temporelle), une requête  $Q$  est classée au niveau  $N_1$  :

- fully described,  $QFD_{N_1X_i}$ , si il existe des catalogues qui couvrent  $x_i$
- partially-described,  $QPD_{N_1X_i}$ , si  $Q$  n'est pas  $QFD_{N_1X_i}$  et s'il existe des catalogues qui couvrent  $x'_i$ , instance de  $X_i$  en relation sémantique, spatiale ou temporelle avec  $x_i$
- un-described, si  $Q$  n'est ni  $QFD_{N_1X_i}$  ni  $QPD_{N_1X_i}$

Des définitions analogues sont valables au niveau ( $N_2$ ) des sources de données.

Nom	Traitement
$QFD_{N_1X_1}$	$Requ\hat{e}te(Q) \wedge subject(Q,x_1) \wedge Catalogue(C) \wedge subject(C,x_1)$
$QPD_{N_1X_1}$	$Requ\hat{e}te(Q) \wedge \neg QFD_{N_1X_1}(Q) \wedge subject(Q,x_1) \wedge Catalogue(C) \wedge subject(C,x'_1) \wedge rS\acute{e}mantique(x_1,x'_1)$
$QUD_{N_1X_1}$	$Requ\hat{e}te \sqcap \neg(QFD_{N_1X_1} \sqcup QPD_{N_1X_1})$

**Table 5 :** Exemples de classifications de requêtes, au niveau  $N_1$  (Catalogues), pour la dimension  $X_1$  (thématique).

Ces classifications jouent le rôle d'annotations sur les requêtes, qui pourraient être exploitées pour optimiser le processus de RI, en proposant à l'utilisateur des requêtes approchées de type "fully described" ou "partially-described". Dans une boucle d'optimisation plus large (cfr. § 4.3.2.1), ces annotations pourraient renseigner les producteurs sur les usages effectifs des données, opérés par des utilisateurs.

## 5.4 Concordances et originalités de NARI par rapport à d'autres approches

Le but de cette discussion est d'établir une analyse critique de NARI. Nous décrivons des concordances et originalités de cette approche, par rapport à d'autres approches de RI et par rapports aux objectifs fixés dans le § 5.1.2. Nous analysons ensuite en quelle mesure les objectifs ont été atteints, et les limites actuelles de NARI.

### 5.4.1 Concordances

Globalement, pour la conception de NARI, nous avons intégré des démarches en ingénierie des systèmes d'information et en base de données présentées dans l'état de l'art (§ 3).

NARI réalise une Recherche d'Information à partir de sources de données hétérogènes, cataloguées et géoréférencées, relatives à plusieurs domaines disciplinaires (par exemple, santé, social, ...), en prenant en compte des critères de qualité. Ces aspects sont représentés au niveau *conceptuel* dans l'ontologie d'application, par des classes en correspondance des thématiques et des différents types de territoires, ainsi que par des relations sémantiques, topologiques et géométriques. Ces concepts et relations permettent d'enrichir la RI par des techniques d'analyse spatiale.

*L'intégration* des sources de données est motivée par des *exigences* législatives et réglementaires, pour construire une vue globale, à partir des descriptions des ressources. Les exigences de *qualité externe* de la Recherche d'Information (manque de données, seuils de qualité non respectés, par rapport à des contextes applicatifs) sont capturées et supportées par les *métadonnées*.

Les *modèles de développement centrés sur les risques* sont interprétés du point de vue de l'utilisateur. On focalise sur les *étapes en amont* de la conception, pour aboutir à des éléments d'architecture globale et d'infrastructure technique. Celle-ci s'appuie sur les langages informatiques de représentation de connaissances et sur les outils associés.

Plus particulièrement, NARI présente des concordances avec d'autres travaux de recherche du même domaine, comme indiqué dans le § 5.2 et le § 5.3, ainsi que dans les Chapitres suivants.

Schématiquement, par rapport à d'autres approches, nous positionnons NARI selon différents aspects :

- elle est *globale*, car elle s'insère dans les trois axes de Recherche d'Information en santé, avec des buts décisionnels, décrits dans le § 3.2. L'adaptation des requêtes a pour but de pallier le manque de sources de données ou leur insuffisante qualité. La classification des requêtes évalue les ressources d'information, dans le but d'optimiser a priori le processus de RI, et de contribuer à améliorer la qualité de la production des données (cfr. Figure 2).
- elle *s'accorde au modèle Inspire "publier - trouver - accorder - lier"*, dans une vue globale et unifiée de systèmes multi-sources et multi-usages, sur les ensembles des acteurs, des besoins informationnels et des ressources. Par rapport à l'architecture Inspire (cfr. Figure 6), les traitements des aspects *existence* et *qualité* de NARI (cfr. RI1 et RI2 dans le § 5.2) se positionnent au niveau des services de découverte, de transformation et d'évaluation, tandis que les traitements concernant les *contenus* (cfr. RI3 dans le § 5.2) sont au niveau des services d'intégration.
- en accord avec le cadre DWQ (cfr. Figure 5), elle *lie les problématiques de la qualité des connaissances et la qualité des données*, dans une chaîne logique : les adaptations des requêtes au niveau des étapes 1 et 2 de NARI peuvent être utiles à la préparation de sources de données spécifiques pour des entrepôts de données, tandis que l'étape 3 concerne la médiation des données. Par rapport à DWQ, nous focalisons sur des aspects conceptuels et organisationnels.
- elle *s'appuie sur des langages de représentation des connaissances et sur les outils du Web sémantique* qui sont les plus adaptés à supporter la montée en charge, la généralisation et la robustesse théorique de l'approche.

#### 5.4.2 Originalités

Les originalités de la conception de NARI consistent, à notre avis, dans deux aspects :

- l'introduction de la métaphore de l'écart d'impédance
- les aspects réglementaires dans la prise de décision en santé, structurée par la prise en compte de la dimension géographique.

La métaphore de l'impédance constitue une clé d'analyse et de conception globale pour NARI. Elle nous amène à :

- adopter une vision systémique de NARI, dont les composantes doivent réduire les écarts entre les systèmes producteurs de données et les systèmes utilisateurs de ces données, et

faciliter les flux entre ces systèmes pour répondre à une recherche d'information avec des buts décisionnels ;

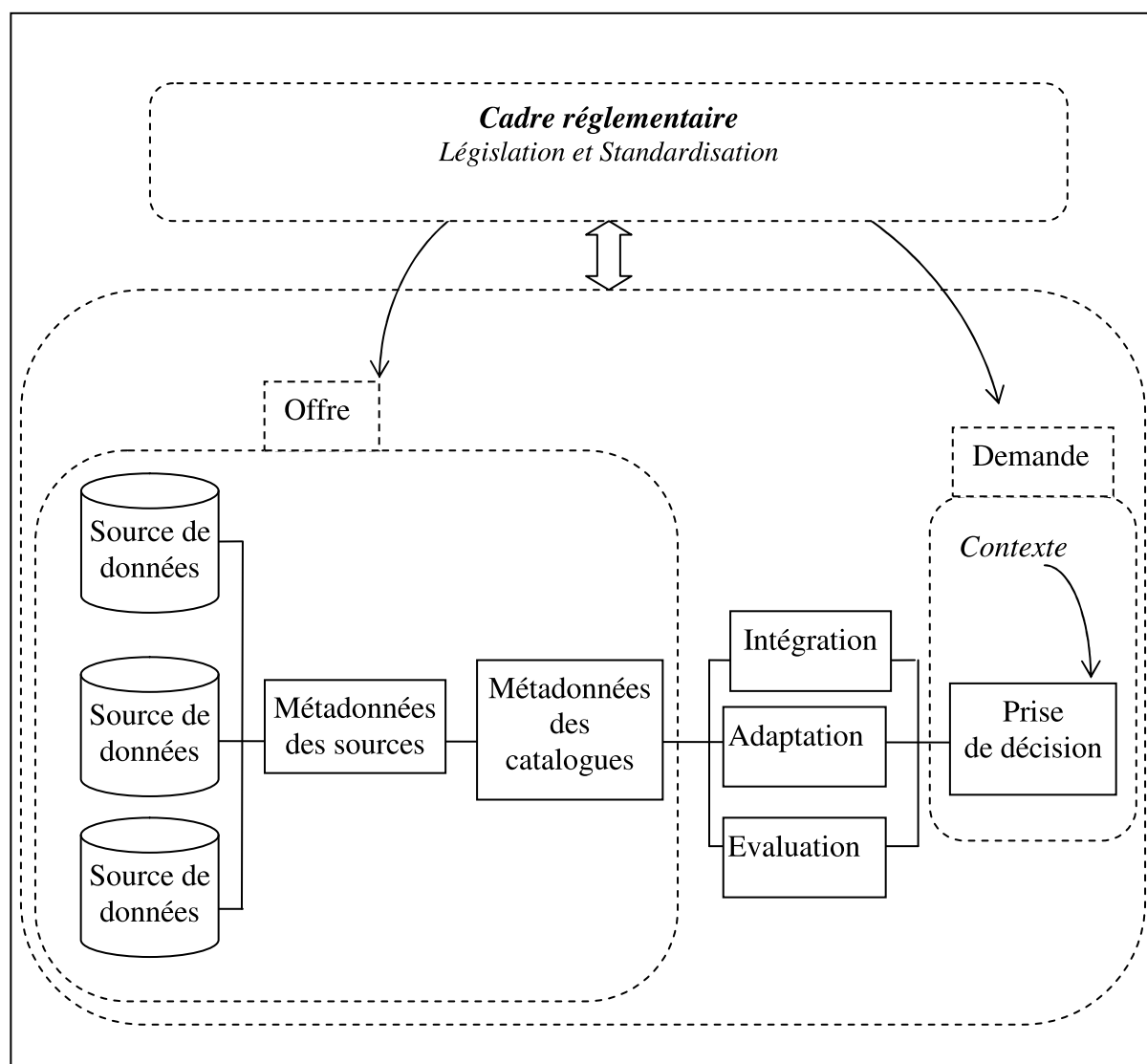
- considérer ces systèmes, ainsi que le domaine d'application, comme profondément structurés par la dimension géographique ;
- proposer les éléments préliminaires d'une démarche qui traite progressivement les différents aspects de la Recherche d'Information (de l'identification / préparation des sources à l'intégration de contenus), et ceux-ci à différents niveaux, à partir des catalogues contenant des sources de données ;
- interpréter les hétérogénéités, le manque de données ou leur insuffisante qualité comme des *résistances* à la Recherche d'Information ;
- indiquer des composantes logicielles (*réactances*) avec différentes finalités (décomposition de requêtes, adaptation, coordination, intégration, ...).

En accord avec ces visions, nous avons choisi :

- une technique d'intégration de type LAV, privilégiant le monde réel et les évolutions croissantes en besoins d'information, plutôt que les données disponibles ;
- des techniques de Recherche d'Information, relevant des bases de données, de la recherche documentaire et de l'analyse spatiale, s'appuyant sur des métadonnées représentatives de la qualité externe ;
- une architecture évolutive, à base de composantes

Pour illustrer la faisabilité conceptuelle de la démarche, nous avons construit des simples exemples.

L'attention portée aux aspects réglementaires (juridiques, de normalisation, ...) nous amène à interpréter les objectifs de nos recherches dans le contexte d'un système visant à *concilier l'offre et la demande*. La Figure 8 complète la Figure 1 : elle explicite cette interprétation et les fonctions de l'approche NARI.



**Figure 8 :** Schéma de l'évolution des objectifs de recherche

### 5.4.3 Atteinte des objectifs et limites actuelles de l'approche NARI

Les recherches présentées dans ce rapport apportent des éléments préliminaires et globaux à la conception de NARI. Elles ont fait l'objet de la publication d'un chapitre de livre (Stein *et al.*, 2009) et dans les actes de congrès et workshops de différentes communautés scientifiques (veille scientifique et technique, géomatique, qualité de l'analyse spatiale, e-gouvernement...). Ces communautés partagent un fort intérêt pour une culture pluridisciplinaire, ancrée territoire, à composante technologique.

Les *limites actuelles* de la conception de NARI, qui méritent d'être améliorées, concernent à notre avis des aspects de robustesse et d'interactivité. Ces aspects sont corrélés.

- La *robustesse* peut être améliorée en approfondissant la métaphore de l'impédance par une formalisation quantitative des problèmes d'hétérogénéité et de qualité (*résistances*) et des coûts des modules activés pour résoudre ces problèmes (*réactances*). Ces aspects sont déterminants pour estimer les bénéfices apportés par NARI. Ils sont importants aussi pour formaliser les stratégies d'adaptabilité et de transformation de requêtes. Ces stratégies concernent les choix multiples, qui peuvent porter sur : les critères de qualité, les modèles de similarité entre territoires, les relations de tout type entre territoires ou thématiques, les dimensions utilisées pour transformer des requêtes. L'utilisation d'un jeu volumineux de données hétérogènes doit mettre en évidence les gains apportés par NARI, par rapport à des approches explorant la globalité des sources sans applications de critères a priori
- *L'interactivité* peut être améliorée par la mise en œuvre de composantes logicielles supportant différents échanges entre les utilisateurs et le système, de l'expression d'une requête à l'acceptation des résultats. Ces échanges sont particulièrement utiles dans des contextes de choix multiples, pour anticiper le plus possible les résultats attendus par les utilisateurs.

Le paragraphe suivant indique des perspectives de recherche visant à améliorer ces aspects et globalement l'approche plus à long terme.

## 6. Perspectives de recherche et Conclusions

La discussion esquissée dans le § 5.4 nous conduit à tracer des perspectives de recherche et à formuler des remarques de conclusion sur l'ensemble de ces recherches.

### 6.1 Perspectives de recherche

Nos perspectives de recherche sont dans un premier temps orientées à améliorer la robustesse et l'interactivité de NARI, et ensuite à enrichir l'approche NARI de fonctionnalités pour l'évaluation.

#### 6.1.1 Robustesse

Pour augmenter la robustesse de l'adaptabilité de NARI, nous envisageons des approfondissements conceptuels, portant globalement sur la métaphore de l'impédance.

Pour analyser des aspects quantitatifs de NARI, nous envisageons d'approfondir notre étude des recherches menées par J. Levesque (Levesque, 2007), sur l'estimation des coûts liés à l'adaptation des demandes des utilisateurs en fonction des données disponibles.

Les approfondissements qualitatifs de NARI concernent plus particulièrement le modèle des utilisateurs et la prise en compte du "zonage" des activités.

#### **6.1.1.1 Modèle de l'utilisateur**

On souhaite améliorer, dès les phases initiales, l'exploration, la sélection des sources et l'ordonnancement des résultats, par des fonctions d'adaptabilité au contexte de l'utilisateur (cfr. (A2), § 5.3.2.1). Cette démarche, basée sur les interactions entre l'utilisateur et le système, est aussi en phase avec les principes du Web 2.0 (O'Reilly, 2005). L'idée est de bâtir un modèle augmenté de l'utilisateur (MA1), à partir d'un modèle générique de l'utilisateur (Zayani *et al.*, 2006). Ce modèle doit permettre d'anticiper sur les usages effectifs des données, en cernant l'activité de l'utilisateur par des caractéristiques comme : localisation, domaine d'activité (santé, social, environnement, juridique,...), type d'activité (pratique médicale, enseignement, recherche, action sociale,...), niveau de l'activité (communal, régional, national, international), niveau du pilotage (stratégique, opérationnel), liens institutionnels (services publics, agences, collectivités, industries,...). Ces connaissances orientent efficacement les recherches et réduisent l'indétermination des résultats. Ceci est le cas, par exemple, pour une recherche sur les personnes âgées dépendantes, liée potentiellement à plusieurs analyses (thérapies gériatriques, localisation des établissements d'Hospitalisation à Domicile, ...).

Concernant l'implémentation du modèle (MA1), on envisage de comparer l'utilisation d'un langage à base de profils, de type Composite Capability/Preference Profile (CC/PP) (W3C, 2004) ou d'une ontologie, pour inférer sur l'exploration du domaine en corrélant les modèles de l'utilisateur (MA1), des ressources (MA2) et du domaine (MA4).

#### **6.1.1.2 "Zonage" des activités**

Le "zonage" des activités de santé évolue continuellement et constitue un aspect critique de la prise de décision en santé, notamment dans les études comparatives des systèmes de santé. Pour ces études, il serait utile d'analyser les problèmes les plus courants liés au manque de

données et de déterminer une classification des adaptations des requêtes selon la dimension géographique, pour automatiser ces traitements.

L'idée est de partir de l'analyse des nouvelles réglementations sur les territoires de santé. Le projet de lois « Hôpital, patients, santé, territoires » (Assemblée Nationale, 2008), illustre des évolutions des territoires de santé : en constatant l'inégalité des Français face à l'accès aux soins, il prône une nouvelle territorialisation de l'aménagement de l'offre de soins. Ainsi, par exemple, les Schémas Régionaux d'Organisation des Soins (SROS), utilisent actuellement jusqu'à cinq niveaux territoriaux, allant du niveau de proximité, pour une offre de soins dits "de premier recours", jusqu'au niveau inter-régional, pour des activités de pointe (greffes, neurochirurgie). La carte des zones de santé a été récemment révisée par le législateur et "le bassin de vie" se substitue désormais au canton, comme niveau de "proximité" pour analyser les besoins de santé et l'offre de soins, notamment en termes d'accessibilité aux soins.

La révision du partitionnement des territoires par rapport aux soins est liée aussi à la mise en place très récente de la T2A (tarification à l'activité), la localisation des centres de soins et l'installation des professionnels de santé libéraux. Des outils cartographiques, comme par exemple C@rtoSanté, produit par les URCAM, permettent de connaître les valeurs de divers indicateurs socio-économiques et de consommation médicale à différents niveaux d'échelle géographique, du canton à la région.

La révision du partitionnement des territoires a un impact économique important. Les aides financières dépendent actuellement de critères, comme la densité des généralistes et leurs activités, mesurées en chiffre d'affaires. Cette révision est complexe, car elle doit prendre en compte les spécificités géographiques des territoires, qui influencent les temps d'accès aux structures de soins (zones accidentées, manquant d'axes rapides de communications, avec de nombreux reliefs, cours d'eau,...).

### **6.1.2 Interactivité**

Pour améliorer l'interactivité, nous souhaitons enrichir techniquement NARI, à différents niveaux, par :

- une interface apte à représenter les échanges dynamiques et interactifs entre NARI et l'utilisateur, depuis la formulation des requêtes jusqu'à la validation des résultats, en passant par l'adaptabilité des requêtes

- une interface cartographique pour visualiser les territoires et les valeurs assumées par les différents critères de qualité, aux niveaux des Catalogues et des Sources de données
- ainsi que par des interfaces qui visualisent différentes structures de données (classifications, thésaurus associés aux thématiques de santé, comme par exemple le thésaurus de la Banque Documentaire de Santé Publique<sup>60</sup>).

### 6.1.3 Extensions

A plus long terme de nombreuses extensions de NARI peuvent être envisagées, et donner lieu à des collaborations avec d'autres équipes de recherche. Elles se situent à différents niveaux :

- au niveau de l'ontologie d'application : (i) prendre en compte le multilinguisme et plus généralement les zones interfrontalières ; (ii) introduire la gestion des versions (des catalogues, des sources de données, des classifications), et des annotations des utilisateurs, pour compléter les classifications actuelles des requêtes. Si elles sont prises en compte par les "régulateurs" et par les producteurs de données, ces annotations peuvent contribuer à l'amélioration continue du système global de l'offre et de la demande d'informations.
- au niveau de l'ontologie du domaine : coupler plusieurs ontologies, associées à diverses spécialités médicales (par exemple toxicologie<sup>61</sup>) et géographiques (cadastre, occupation du sol, ...).
- au niveau de l'architecture globale : introduire des modules de géo-référencement, pour enrichir des données thématiques, utiliser des techniques de workflow pour spécifier des composantes d'architecture, leur composition et coordination
- au niveau technique : s'appuyer sur des outils renseignant sur la qualité des données ("quality aware"), qui, selon (Devilleers *et al.*, 2005), visent à (i) intégrer, mettre à jour et visualiser les métadonnées sur la qualité externe; (ii) avertir les utilisateurs d'éventuels problèmes concernant la qualité des données et l'usage simultané avec d'autres données ; (iii) faire fonctionner différentes composantes en fonction du contexte.

---

<sup>60</sup> BDSP : <http://www.bdsp.ehesp.fr/>

<sup>61</sup> TOXIBASE : <http://www.toxibase.org/>

## 6.2 Conclusions sur ces recherches

Les recherches présentées dans ce rapport ont pour objectif d'apporter des éléments de conception d'une nouvelle approche de Recherche d'Information (NARI) dans des contextes multi-sources, géoréférencées, fortement hétérogènes, où les problèmes de réglementation et de volumes ont une grande importance.

Ces recherches s'inscrivent dans la thématique plus générale de l'interopérabilité des Systèmes d'Information. Le développement et la diffusion des standards, ainsi qu'une plus large disponibilité de données « publiques », témoignent de l'importance de ces thèmes dans de nombreux secteurs d'activités.

Le problème posé et les solutions envisagées sont complexes, car ils concernent les organisations, les modèles, les architectures et les technologies.

Nous avons établi un état de l'art qui, bien que succinct, témoigne de cette complexité et de l'articulation nécessaire entre différentes approches et perspectives, qui relèvent de l'informatique, de la gestion, des SIG. Le fil conducteur choisi pour élaborer cet état de l'art s'appuie sur une vue systémique des systèmes d'information et de la recherche d'information. Il privilégie le point de vue des utilisateurs, comme levier d'amélioration de la qualité globale des organisations. Cette vue systémique est renforcée par les évolutions récentes (formalismes, techniques et outils) de la représentation des connaissances, détaillées dans le Chapitre 2.

Ce même fil conducteur nous a conduit à élaborer la métaphore de l'impédance, avec laquelle nous avons bâti un cadre préliminaire d'analyse et de conception, illustré sur un champ particulièrement complexe, la santé, et nous avons opéré des choix d'architecture.

Achever la conception de NARI est un projet sur le long terme, qui nécessite d'une part plusieurs compétences, aussi bien généralistes que spécialisées, et d'autre part la disponibilité de jeux de données significatifs. Il nous semble intéressant, dans le court et moyen terme, d'approfondir la métaphore de l'impédance, du point de vue quantitatif et qualitatif, et puis d'étendre NARI. Pour cela nous comptons nous appuyer sur des collaborations avec des équipes de recherche, en France et au Canada, qui travaillent sur la recherche d'information et sur la qualité externe des données en santé, en adoptant une vision pluridisciplinaire des Systèmes d'Information.

## 7. Références

- Abiteboul, S., Aurawal, R., Bernstein, B. *et al.* *The Lowell Database Research Self-Assessment*. Communication of the ACM, Vol. 48, n°5, May 2005, pp. 111-118.
- Abiteboul, S., Buneman, P., Suciu, D. *Data on the Web: From Relations to Semistructured Data and XML*. Morgan Kaufmann Publishers, 1999.
- Akoka, J., Berti-Equille, L., Boucelma, O., Bouzghoub, M., Comyn-Wattiau, I., Cosquer, M., Goasdoué, V., Kedad, Z., Peralta, V., Nugier S., Si-Said, S. *A framework for quality evaluation in data integration systems*. In Proceedings of the 9th International Conference on Enterprise Information Systems, 2007 (ICEIS'07).
- Amann, B., Beerl C., Fundulaki I., Scholl M. *Ontology-based integration of xml web resources*. LNCS, Vol. 2342, 2002, pp. 117-131.
- Ambler, S.W. *Agile Modeling: A Brief Overview*. In Evans, France, Moreira & Rumpe (éd.) Proc. of 'Practical UML-Based Rigorous Development Methods' Workshop, UML2001 Conference, Toronto, Canada, 1er Octobre 2001, LNI series, vol. 7, pp. 7-11.
- Amous, I., Jedidi, A., Sèdes, F. *Documents semi-structurés et métadonnées, Contribution à la réingénierie de collections de documents*. In : Calabretto, S., Pinon J.M. (sous la direction) Bases de données semi-structurées - Ingénierie des systèmes d'information - RSTI - Série ISI, Vol. 8/2003, pp 153-172, 2004
- Assemblée Nationale, Comm. des Affaires Culturelles, Familiales et Sociales, Mission d'information sur l'offre de soins sur l'ensemble du territoire, « Rapport d'Information », 30 septembre 2008, <http://www.assemblee-nationale.fr/13/rap-info/i1132.asp>
- Badue C., Barbosa R., Golgher P., Ribeiro-Neto B., Ziviani N. « Basic Issues on the Processing of Web Queries » The 28th International Conference on Information Retrieval (ACM SIGIR'05), Salvador, Brazil, 2005, pages 577-578
- Baeza-Yate R.A., Ribeiro-Neto B., *Modern Information retrieval*. Addison-Wesley Longman Publishing CO., Inc., Boston, 1999
- Bardram, J.E. *Scenario-Based Design of Cooperative Systems. Re-designing an Hospital Information System in Denmark*. Group Decision and Negotiation, 9:237-250, Kluwer Academic Publishers, 2000

- Batini, C., Scannapieco, M., Data Quality: Concepts, Methodologies And Techniques, Springer-Verlag Berlin and Heidelberg GmbH & Co. K, 2006
- Bédard, Y., Gosselin P., Rivest S., Proulx M.-J., Nadeau M., Lebel G., Gagnon M.-F., *Integrating GIS Components with Knowledge Discovery Technology for Environmental Health Decision Support*, International Journal of Medical Informatics, Vol. 70, No. 1, 2003, pp. 79-94.
- Bédard, Y., Rivest, S., Proulx, M.-J. *Spatial On-Line Analytical Processing (SOLAP): Concepts, Architectures and Solutions from a Geomatics Engineering Perspective*, In Robert Wrembel & Christian Koncilia (ed(s)), Data Warehouses and OLAP : Concepts, Architectures and Solutions, Chapitre 13, IRM Press (Idea Group), London, UK, 2007, pp. 298-319.
- Ben Saïd, M., Le Mignot, L., Mugnier, C., Richard, J.B., Le Bihan-Benjamin, C., Jais, J.P., Simonet, A., Guillon, D., Simonet, M., Landais P. *A Multi-Source Information System via the Internet for End-Stage Renal Disease: Scalability and Data Quality*. Stud Health Technol Inform. 2005; 116:994-9, 2005.
- Berners-Lee, T., Handler, J., Lassila, O. The Semantic Web. A new form of Web content that is meaningful to computers will unleash a revolution of new possibilities. In Scientific American, May 17, 2001, [http://www.scientificamerican.com/print\\_version.cfm?articleID=00048144-10D2-1C70-84A9809EC588EF21](http://www.scientificamerican.com/print_version.cfm?articleID=00048144-10D2-1C70-84A9809EC588EF21)
- Berti-Equille, L. *Un état de l'art sur la qualité des données*. In : *Qualité des systèmes d'Information - Ingénierie des systèmes d'information - RSTI - Série ISI*, 9/2004, pp. 117-143.
- Beuscart, R., Zweignebaum, P., Venot, A., Degoulet, P. *Télé médecine et e-santé*. Paris : Springer-Verlag, 2002
- Bonnevay, S., Lamure, M. (sous la direction de), *Santé et Systémique*, 6/2002, Lavoisier, 2002
- Bordin, P., *SIG concepts, outils et données*, Lavoisier, 2002
- Bounekkar, A., Duru, G., (éd.). *Les nouvelles organisations des systèmes de santé, Santé Décision Management*. Volume 11, n° 3-4/2008.

- Bourret, C., Salzano, G., Caliste, J.-P. *Nouveaux métiers dans le domaine de la santé : maîtrise de l'information, transversalité des compétences et autres exigences*. In : Actes du Colloque "Les systèmes d'information élaborée", Ile Rousse, France, Octobre 2002.
- Bouzeghoub, M., Peralta, V. *A Framework for Analysis of Data Freshness*. In : Proceedings of IQIS 2004, Paris France, 2004
- Brunie, V., Morizet-Mahoudeaux, P., Bachimont, B. *Separating Textual Contents from Structures for Reading Hypertext Structured Medical Records*. HYPERTEXT '98, Pittsburgh, PA, USA, June 20-24, 1998
- Buthion, V., Flory, A.. *Utilisation des entrepôts de données pour produire des études médico-économiques*. In Bonnevey, S. Lamure, M. (sous la direction de), *Santé et Systémique*, 6/2002, Lavoisier, 2002, pp. 31-50.
- Calvanese, D., McGuinness, D., Nardi, D., Patel-Schneider P. *The Description Logic Handbook: Theory, Implementation and Applications*. UK, Cambridge Univ. Press, 2004.
- Castro, J., Kolp, M., Mylopoulos, J. *Towards requirements-driven information systems engineering: the Tropos project*. *Information Systems*, vol. 27 n°6, 2002, pp. 365-389.
- Cauvet, C., Rosenthal-Sabroux, C. *Ingénierie des systèmes d'information*. Hermès Science Europe Ltd., 2001.
- Chaari, T., Laforest, F., Celentano, A. *Adaptation in Context-Aware Pervasive Information Systems: The SECAS Project*. *International Journal on Pervasive Computing and Communications*, vol 3-4, 2007.
- Charlet, J., Reynaud, C., Teulier, R.. *Ingénierie des connaissances pour les systèmes d'information*. In : Cauvet, C., Rosenthal-Sabroux, C. (sous la direction de) *Ingénierie des systèmes d'information*. Hermès Science Europe Ltd., 2001.
- Chawathe, S., Garcia-Molina, H., Hammer, J., Ireland, K., Papakostantinou, Y., Ullman, J., Widom, J. *The TSIMMIS Project: Integration of Heterogeneous Information Sources*. In: Proceedings of IPSJ Conference, pp. 7-18, Tokyo, Japan, October 1994, <http://www-db.stanford.edu/pub/papers/tsimmis-overview.ps>
- Curien, N., Muet, P.-A. *La société de l'information*. La Documentation française, Paris, 2004 - ISBN : 2-11-005534-0, <http://lesrapports.ladocumentationfrancaise.fr/BRP/044000180/0000.pdf>

Darses, F., Dieng, R., Simone, C., Zaklad, M. *Cooperative Systems Design, Scenario-Based Design of Collaborative Systems*. IOS Press, 2004.

DeRosa, J.K., Rebovich, G., Norman, D. *Le casse-tête des exigences dans l'ingénierie des systèmes complexes*. Génie Logiciel, n° 87, pp. 9-14, Décembre 2008.

Desconnets, J.C., Moyroud, N., Libourel, T. *Méthodologie de mise en place d'observatoires virtuels via les métadonnées*. InforSid, Nancy, Juin 2003. Voir demo Mdweb : <http://www.mdweb-project.org>

Devillers, R., Beard, K. *Communication et utilisation de l'information sur la qualité dans les SIG*. In Devillers, R., Jeansoulin, R. (éd.) *Qualité de l'information géographique*. Hermès Science, 2005

Devillers, R., Jeansoulin, R. (éd.) *Qualité de l'information géographique*. Hermès Science, 2005

Elmargamid, A.K., Rusinkiewicz, M., Sheth, A.P. (éd.). *Management of Heterogeneous and Autonomous Database Systems*. San Francisco, CA: Morgan Kaufmann Publishers, 1998.

Elmasri, R., Navathe, S. *Fundamentals of Database Systems, Fourth Edition*, Pearson, Addison Wesley, 2003

Freitas, G. M., Laender, A. H. F., Campos, M. L., MD2 Getting Users Involved in the Development of Data Warehouse Applications, Caise2002 (2002)

Fundulaki, I., Amann, B., Beerli, C., Scholl, M. *STYX: Connecting the XML World to the World of Semantics (Demo)*. In Proceedings EDBT'2002, 2002.

Gervais, M., *Pertinence d'un manuel d'instructions au sein d'une stratégie de gestion du risque juridique découlant de la fourniture de données géographiques numériques*, Thèse de doctorat, Université Laval, 2004, 347 pages.

Ginsberg, J., Mohebbi, M.H., Patel, R.S., Brammer, L., Smolinski, M.S., Brilliant, L. "Detecting influenza epidemics using search engine query data", *Nature* 457, 1012-1014, 19 February 2009

<http://www.nature.com/nature/journal/v457/n7232/full/nature07634.html>

Goodchild, M.F., 2005, *Geographic Information Science: The Grand Challenges*, <http://www.geog.ucsb.edu/~good/papers/438.pdf>

- Grudin, J. *Computer-Supported Cooperative Work: Its History and Participation*. IEEE Computer, 27, 5, pp. 19-26, 1994
- Grundstein, M., Rosenthal-Sabroux, C., *Vers un système d'information source de connaissances*, In : Cauvet, C., Rosenthal-Sabroux, C. (sous la direction de) *Ingénierie des systèmes d'information*. Hermès Science Europe Ltd., 2001.
- Guarnieri, F., Garbolino, E. *Systèmes d'information et risques naturels*. Presses de l'Ecole des Mines de Paris, 2004
- Gutiérrez C., Servigne S., *Métadonnées et Qualité pour les Systèmes de Surveillance en Temps-Réel*, In Proceedings SAGEO'2007
- Halevy, A.Y. *Answering queries using Views: A survey*. In The International Journal on Very Large Data Bases, Vol. 10, Issue 4, 2001, pp. 270-294.
- Halevy, A., Franklin, M., Maier, D. *Principles of dataspace systems*, PODS '06: Proceedings of the twenty-fifth ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems, pp. 1-9, 2006.
- Haux, R. *Health information systems: past, present, future*. In International Journal of Medical Informatics, Vol. 7, Issue 3-4, 2006, pp. 268-281.
- Hickey, A.M., Dean, D.L., Nunamaker, J.F. *Establishing a foundation for collaborative scenario elicitation*. In Database for Advances in Information Systems, Vol. 30, Issue 3/4, pp. 92-110, summer-fall 1999
- IEEE Standard Computer Dictionary: A Compilation of IEEE Standards Computer Glossaries. New York, NY: 1990
- Inmon, W.H. *Building the Data Warehouse*, 2nd edition, John Wiley & Sons, USA, 1996)
- INSPIRE, 2008 : <http://inspire.jrc.ec.europa.eu/>
- ISO/IEC JTC1 SC32 N1257, *Information Technology - Metadata for technical standards and specification documents*, 2005-03-30
- Jarke, M., Lenzerini, M., Vassiliou, Y., Vassiliadis, P. *Fundamentals of Data Warehouses*, Springer, 2000.
- Kashyap, V., Sheth, A. *Semantic Heterogeneity in Global Information Systems : The Role of Metadata, Context and Ontologies*. In : Papazoglou, M., Schlageter, G. (éd.) *Cooperative Information Systems: Current Trends and Directions*, Academic Press, pp. 139-178, 1998.

- Kim, W., Choi, I., Gala, S., Scheevel, M. *On Resolving Schematic Heterogeneity in Multidatabase Systems*. In : *Distributed and Parallel Databases*, Kluwer Academic Publishers, Vol. 1, n. 3, pp. 251-279, 1993.
- Kimball, R., *The Data Warehouse Toolkit*, John Wiley and Sons, Inc, Second edition, 436 pages, 2002
- Knublauch, H., Ferguson, R.W., Noy, N.F., Musen, M.A. *The Protégé OWL Plugin: An Open Development Environment for Semantic Web Applications*, LNCS, Vol. 3298, 2004, pp. 229-243.
- Lenzirini, M. *Data Integration : A theoretical Perspective*. Course : logic-based information integration, ESSLLI 2005
- Levesque, J. *Evaluation de la qualité des données géospatiales - Approche top-down et gestion de la métaqualité*, Thèse de doctorat Université de Laval, Quebec, Canada, 2007.
- Libourel, T. (animateur) Action Spécifique 97 du département STIC du CNRS. Médiation via les métadonnées. 2003, <http://www.lirmm.fr/~libourel/MM/MetaMedia.htm>
- Nakache, D. *Data Warehouse design and methodology for the French Health Department*. In: Proceedings of JCIS - CS&I, (7th Joint Conference on Information Sciences), 26 -30 septembre 2003, Cary, North Car. 2003.
- O'Reilly, T. *What is Web 2.0. Design Patterns and Business Models for the Next Generation of Software*, 09/30/2005, <http://oreilly.com/web2/archive/what-is-web-20.html>
- Papazoglou, M. P., Spaccapietra, S., Tari, Z. (éd.) *Advances in Object-Oriented Data Modeling*, MIT Press, 2000
- Parent, C., Spaccapietra, S. Intégration de bases de données : panorama des problèmes et des approches. In *Ingénierie des Systèmes d'Information*, Vol. 4, n. 3, p. 333-358, 1996
- Pohl, K., Haumer, P. Modelling Contextual Information about Scenarios. 3rd International Workshop on Requirements Engineering: Foundation for Software Quality, Barcelone, Spain, June 16-17, 1997.
- Proulx M.-J., Bernier, E., Bédard, Y.. *Revue systématique en santé environnementale*, 2007, [http://www.nceh.ca/files/Technologies\\_de\\_la\\_geomatique\\_nov\\_2007.pdf](http://www.nceh.ca/files/Technologies_de_la_geomatique_nov_2007.pdf)
- Rolland, C., Prakask, N. *From Conceptual Modelling to Requirement Engineering*. *Annals of Software Engineering*, 10, pages 151-176, 2000

- Romano, N.C., Chen, F., Nunamaker, J.F. *Collaborative projects management software*. In: Proceedings of the 35th Hawaii International Conference on System Sciences, 2002.
- Salaün Jean-Michel, "Documents et Numérique", in Curien Nicolas et Muet Pierre-Alain "La société de l'information", La Documentation française, Paris, 2004 - ISBN : 2-11-005534-0, 2004, <http://lesrapports.ladocumentationfrancaise.fr/BRP/044000180/0000.pdf>
- Salzano, G., Bourret, C. *Health Networks and global health services. An Information System Viewpoint*, In Actes du congrès ICSSHC2004, Genève, 2004.
- Sèdes, F. *Ingénierie des Systèmes d'Information*, Vol 12, n°2/2007
- Servigne, S., Lesage, N., Libourel, T. *Composantes qualité et métadonnées*. In : Devilliers, R., Jeansoulin, R. (éd.), *Qualité de l'information géographique*. Paris : Hermès Science, 12 octobre 2005, 384 p.
- Sheth, A. *Changing Focus on Interoperability in Information Systems: From System, Syntax, Structure to Semantics*. In: Goodchild, M.F, Egenhofer, M.J., Fegeas R., Kottman, (éd.), *Interoperating Geographic Information Systems*. Etas unis: Kluwer Academic Publishers, février 1999, 536 p.
- Sirin, E., Parsia, B., Cuenca, Grau B., Kalyanpur, A., et Katz, Y. *Pellet: A practical OWL-DL reasoner*. *Web Semantics: Science, Services and Agents on the World Wide Web*, Volume 5, Issue 2, June 2007, pp. 51-53.
- Sommerville, I. *Software Engineering*. 7<sup>ème</sup> Edition. Etas unis : Addison Wesley, 20 mai 2004, 784 p.
- Stein A., Shi W., Bijker, W. (éd.). *Quality Aspects in Spatial Data Mining*. Etas unis : CRC Press, Taylor & Francis Group, 18 septembre 2008, 374 p.
- Thuraisingham, B., Gupta, A., Bertino, E., Ferrari, E. Knowledge Management for Heterogeneous Information Exchange. In : Bestougeff, H., Dubois, J.E., Thuraisingham, B. (éd.), *Heterogeneous Information Exchange and Organizational Hubs*. Netherlands : Kluwer Academic Publishers, juin 2002, 256 p.
- Verdier, C., Ouziri, M. *Intégration et visualisation de documents médicaux partagés. Santé et Systémique*, sous la direction de Bonnevey, S., Lamure, Lavoisier, M., 2002, p. 9-30.
- Vigneron, E., Tonnellier, F. *Géographie de la santé en France*. France : Presses Universitaires de France (PUF), 1 février 1999, 128 p.

Villac, M. *La 'e-santé' : Internet et les TIC au service de la santé*. In : Curien N. et Muet P.-A., *La société de l'information*. Paris : La Documentation française, 2004, p. 277-298.  
<http://lesrapports.ladocumentationfrancaise.fr/BRP/044000180/0000.pdf>

Wiederhold, G., Jannink, J. *Composing Diverse Ontologies*. Prepared for IFIP Working Group on Database 8th Working Conference on Database Semantics (DS-8), Rotorua, New Zeland. Janvier 1999. <http://www-db.stanford.edu/SKC/publications/ifip99.html>

W3C, Composite Capability/Preference Profiles (CC/PP): Structure and Vocabularies 1.0, Recommendation 15 January 2004, <http://www.w3.org/TR/CCPP-struct-vocab/>

Zayani, A., Péninou, A., Canut, M.F., Sedes F. *An adaptation approach: query enrichment by user profile*, In : Proceedings of International Conference on Signal-Image Technology and Internet-Based Systems (SITIS), pages 24–35, 2006.

## **Chapitre 2 : Technologies du web sémantique et outils associés : état de l'art et choix pour NARI**

# Table des matières

<b>1. Introduction .....</b>	<b>75</b>
<b>2. Objectifs technologiques de NARI.....</b>	<b>76</b>
<b>3. Présentation des technologies du web sémantique.....</b>	<b>77</b>
3.1 Extensible Markup Language (XML) et XML Schéma.....	78
3.2 Ressource Description Framework (RDF).....	79
3.3 RDF Schema (RDFS).....	79
3.4 Web Ontology Language (OWL).....	80
3.5 Les règles.....	86
3.6 Combinaison de règles et d'ontologies .....	87
3.7 Les outils du web sémantique .....	92
<b>4. Evaluation des technologies par rapport aux objectifs.....</b>	<b>95</b>
4.1 Formalismes de représentation de connaissances .....	95
4.2 Approche de médiation .....	95
4.3 Architecture globale .....	97
4.4 Choix d'implémentation technique .....	99
<b>5. Conclusions .....</b>	<b>101</b>
<b>6. Références .....</b>	<b>102</b>

# 1. Introduction

Face à la fragmentation et l'hétérogénéité des sources de données, les utilisateurs et les applications à large échelle demandent un accès rapide et uniforme à un ensemble très volumineux des données contenues dans ces sources.

Les systèmes d'intégration qui visent à fournir une vue unifiée de ces données se distinguent selon l'approche d'intégration employée. Dans le Chapitre 1 nous avons présenté NARI, une Nouvelle Approche de Recherche d'Information, opérant sur des grandes masses de données cataloguées, hétérogènes, qui peuvent être et géo référencées<sup>62</sup>. Nous avons illustré NARI à l'aide d'exemples du domaine de la santé.

NARI doit contribuer à la prise de décision et peut être considérée comme une approche « quality aware », car elle anticipe le plus possible la prise en compte des exigences de qualité « externe », appelées aussi « *fitness for use* », exprimées par les utilisateurs. Dans NARI, ces exigences de qualité, complémentaires à celles définies par les producteurs de données (qualité interne), sont représentées et gérées à l'aide des métadonnées. Celles-ci participent largement à l'optimisation du processus d'intégration, en identifiant et préparant des sources de données pertinentes pour l'intégration.

NARI est structurée par la dimension géographique. L'originalité de l'approche réside dans la métaphore de l'impédance, due à R. Jeansoulin et développée dans les articles conjoints, présentés dans les Chapitre 4 et 5. Cette métaphore guide la conception de NARI, qui s'appuie sur des techniques d'intégration, sur des langages de représentation des connaissances et sur des technologies et outils relevant du web sémantique.

Dans ce chapitre nous approfondissons les apports de ces langages, technologies et outils, à la conception de NARI, pour supporter la description des besoins de qualité, la formalisation du contexte applicatif et des fonctionnalités d NARI.

Plus précisément, ce chapitre :

- décrit les objectifs technologiques de NARI, en partant des objectifs généraux présentés dans le Chapitre 1 (§ 2) ;

---

<sup>62</sup> Avec le terme « données géoréférencées » nous désignons des données concernant des « objets géographiques » référencés par des systèmes directs ou indirects (adresses postales, par exemple). (Cfr. Chapitre 1, §4.1.1).

- analyse les technologies du web sémantique, pertinentes pour NARI, avec leurs rôles, avantages et limites (§ 3) ;
- évalue ces technologies et les outils associés par rapport aux objectifs technologiques de NARI et présente l'architecture globale retenue pour NARI (§ 4), avant de conclure.

## 2. Objectifs technologiques de NARI

Les objectifs qui guident la conception de NARI (Chapitre 1, Table 2) sont les suivants :

- OS1 : Analyser des scénarios de RI en santé représentatifs de cas d'échecs ou d'évolutions significatives, avec une perspective d'ingénierie des besoins. Concevoir une ontologie de domaine, focalisée sur les données (hétérogénéité, qualité)
- OS2 : Proposer des éléments de conception de NARI, concernant :
  - l'ontologie d'application et la spécification des étapes
  - l'architecture globale : niveaux, composantes fonctionnelles, approches techniques
- OS3 : Proposer des éléments de conception de NARI, concernant les langages de formalisation des connaissances et les outils associés.

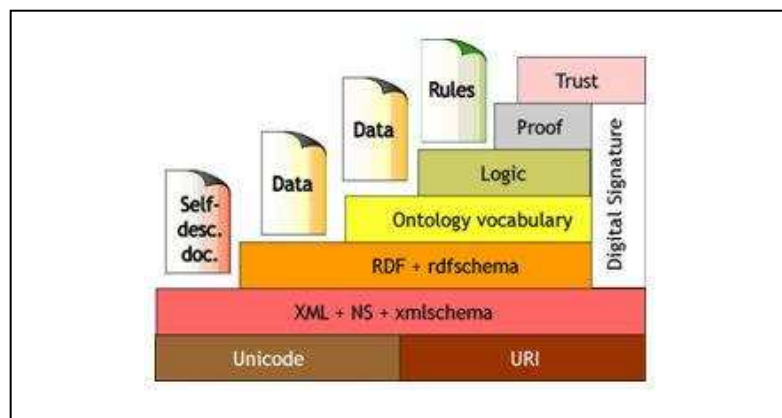
A partir de ces objectifs, nous associons à NARI les objectifs technologiques d'intégration suivants :

- Utilisation d'un seul schéma global, décrivant le domaine d'intérêt et développé indépendamment des schémas des sources locales.
- Possibilité d'ajouter de nouvelles sources de données sans changer le schéma global.
- Formulation des interrogations en termes du schéma global, et construction d'une réponse globale à partir des résultats locaux.
- Modélisation du contexte applicatif et des besoins de qualité requise.
- Recherche du meilleur compromis entre les données et la qualité demandée par rapport aux données disponibles et leur qualité.
- Intégration des données pertinentes répondant aux critères de qualité définis au niveau applicatif

- Identification / préparation de sources des sources de données pertinentes en vue de l'intégration. Réalisation, si nécessaire, de transformations sur la requête et/ou les contraintes et/ou les données.
- Traçabilité des traitements et des transformations effectués.
- Support de grands volumes de données

### 3. Présentation des technologies du web sémantique

L'article (Berners-Lee *et al.*, 2001) introduit le web sémantique comme « une extension du web actuel, dans lequel on donne à une information un sens bien défini pour permettre aux ordinateurs et aux personnes (utilisateurs/opérateurs) de travailler en coopération ». Le web sémantique repose sur un ensemble de composantes qui sont généralement présentées à l'aide d'une architecture que la communauté du web sémantique nomme généralement « Semantic Web Layer Cake », illustrées en figure 1. Parmi ces technologies, le W3C a déjà publié les recommandations pour XML, XML Schema, RDF, RDFS, SPARQL et OWL. Cependant, ces technologies évoluent et sont enrichies de nouvelles fonctionnalités (par exemple SPARQL, OWL2). Actuellement, ce consortium conduit également des travaux sur la standardisation des formalismes de règles et les couches supérieures ne sont pas encore développées.



**Figure 1** : Première version du Semantic Web Layer Cake (Berners-Lee, 2000)

Nous présentons maintenant les différentes couches de cette architecture.

### 3.1 Extensible Markup Language (XML) et XML Schéma

XML (Bray, Paoli *et al.*, 2006) est un langage de balisage extensible permettant d'encoder des données, structurées et semi-structurées, et de spécifier la syntaxe d'autres langages de balisage, c'est donc un métalangage. Il est utilisé, dans le web sémantique, pour la sérialisation de RDF(S) et OWL. XML définit des règles de syntaxe pour organiser des données sous forme d'un arbre étiquetée et constituée d'éléments et d'attributs. Un document XML respectant ces règles est dit *bien formé*. XML a l'avantage d'être indépendant des formats d'affichage et des plateformes informatiques. Il facilite l'échange de données entre applications grâce à son format texte et sa capacité d'encodage des données issues des bases de données relationnelles et objets (Abiteboul *et al.*, 2000).

L'encodage de plusieurs sources de données en XML peut entraîner des conflits sémantiques, notamment si le même nommage est utilisé pour désigner différentes choses. Pour éviter ce type de problème, la solution consiste à utiliser la recommandation W3C des espaces de noms. *Les espaces de noms* (Bray, Hollander *et al.*, 2006) permettent de distinguer les différents vocabulaires utilisés par les applications et facilitent l'utilisation de vocabulaire commun. L'utilisation des espaces de noms consiste à associer des noms de vocabulaires à des IRIs/URIs (International/Uniform Resource Identifier) et de qualifier les éléments et les attributs par ces noms.

L'échange de données au format XML entre applications nécessite la connaissance préalable, de la part des applications cibles, de la structure des données échangées. Ainsi, elles peuvent réaliser des contrôles de document pour éviter d'éventuels problèmes de non-conformité à la structure attendue ou au vocabulaire utilisé. Plusieurs solutions sont disponibles pour définir de tels schémas : DTD (Document Type Definition), RelaxNG ou bien XML Schema. Ce dernier est d'ailleurs recommandé par le W3C (Thompson *et al.*, 2004). Un document XML est dit *valide*, s'il est *bien formé* et conforme à sa syntaxe et son vocabulaire. XML Schema offre, entre autres, la possibilité de spécifier des types prédéfinis ou d'en définir de nouveaux, simples ou complexes, avec ou sans restriction sur : les valeurs possibles, cardinalités, intervalles, ...etc.

XML Schema permet également d'exprimer les types de contraintes d'intégrité comme : les clés, le domaine des valeurs, les types de données, non nul, etc. (Van Der Vlist, 2002). Or XML Schema ne spécifie que la structure et la syntaxe des documents. Il n'est pas désigné

pour fournir une description sémantique du domaine dans lequel l'application opère (Volz, 2004). XML Schema est loin d'être un langage de modélisation.

### 3.2 Ressource Description Framework (RDF)

RDF (Klyne et Carroll, 2004) est un modèle de représentation de connaissances pour le web sémantique permettant de décrire des ressources et les liens qui les relient. Il consiste à formuler des assertions, appelée aussi *statement*, qui expriment qu'une ressource est associée à une autre ressource ou à une valeur à travers d'une propriété. Elle se présente donc sous la forme d'un triplet (*sujet, prédicat, objet*) tel que le *prédicat* est la propriété reliant le *sujet* à un *objet*. Le sujet, et l'objet dans le cas où c'est une ressource, peuvent être identifiés par des IRIs/URIs ou être des nœuds anonymes (*blank node*). Le prédicat est nécessairement identifié par une IRI/URI. Dans le cas où l'objet est une valeur, il est possible de lui associer un type de données comme défini par XML Schema. Un document RDF peut être écrit avec différentes syntaxes, y compris en XML.

En RDF, la représentation des relations ternaires et n-aires est possible à l'aide de la notion de réification et en exploitant des nœuds anonymes (*blank node*).

Les documents RDF bénéficient maintenant d'un langage de requête nommé SPARQL. Ce langage, recommandé par le W3C depuis janvier 2008, est adapté à la structure des graphes RDF, notamment en supportant la notion de nœud vide, en définition des motifs de triplets (*patterns*). La syntaxe de SPARQL est proche de celle de SQL et permet un apprentissage rapide pour la plupart des utilisateurs de bases de données relationnelles. Ce langage est maintenant supporté dans de nombreux outils, (Protégé, Pellet, le triple store Sesame, l'API Jena, etc.) ce qui en fait une technologie indispensable dans le domaine du Web Sémantique.

### 3.3 RDF Schema (RDFS)

RDF Schema (Brickley et Guha, 2004) est une extension de RDF qui fournit un vocabulaire additionnel par rapport à RDF, comme par exemple, les notions de : classe, propriété, sous classe, sous propriété, domaine et co-domaine d'une propriété, etc. Avec ces notions, RDFS est qualifié comme un langage de base pour la représentation d'ontologie. Dans une ontologie on définit les termes utilisés pour décrire et représenter un domaine de connaissances. Les personnes, les bases de données et les applications utilisent les ontologies pour partager l'information du domaine. Un domaine peut être spécifique à un sujet ou à un espace de la

connaissance, comme la santé, l'environnement, etc. Les ontologies incluent des définitions, utilisables par l'ordinateur, des concepts du domaine et des relations entre ces concepts. Le langage d'ontologies RDFS n'est pas très expressif, puisqu'il ne permet pas directement de : définir de nouvelles classes à partir des classes primitives, d'énumérer les valeurs possibles, de préciser les cardinalités sur les propriétés, d'indiquer des conditions nécessaires et/ou suffisantes pour les classes ou d'utiliser des notions utiles comme la négation, l'union et l'intersection de classes et les propriétés transitives. Une autre limite de RDFS, est le fait qu'une classe peut être définie en tant qu'instance d'une autre classe : ceci rend impossible la différenciation entre classes et les instances.

RDFS ne distingue pas entre les éléments du langage et les autres objets du domaine car les constructeurs sont utilisés à la fois pour définir des modèles et pour définir le langage lui-même. Cet aspect impose de lui attribuer une sémantique non standard, c'est-à-dire autre qu'une sémantique du type Tarski que l'on retrouve pour la logique du premier ordre (Volz, 2004).

### **3.4 Web Ontology Language (OWL)**

Le Langage OWL (Smith *et al.*, 2004), est développé comme une extension de RDFS et tire ses fondements des formalismes des logiques de description. Les logiques de description (LD) (Baader *et al.*, 2004) sont une famille de formalismes de représentation de connaissances, basées sur la logique du premier ordre et sur la notion de concept. Les LD jouent un rôle important dans ce domaine et sont reconnues d'une grande utilité dans plusieurs applications telles que l'intégration d'information, l'ingénierie de logiciel et la modélisation conceptuelle (Calvanese *et al.*, 2004). Elles sont considérées aussi comme les formalismes les plus utilisés actuellement dans la construction des ontologies (Horrocks, 2005). Elles permettent d'utiliser différents constructeurs pour bâtir de nouveaux concepts complexes (Table 1).

L'expressivité des LD est dénotée à l'aide de symboles indiquant le ou les constructeurs autorisés (Table 2).

<i>Syntaxe DL</i>	<i>Signification</i>	<i>Syntaxe DL</i>	<i>Signification</i>
$\forall P.C$	Quantifieur universel	$\exists P$	Quantificateur existentiel
$\exists P.C$	Quantificateur existentiel qualifié	$\exists P.\{i\}$	Restriction sur la valeur
$\geq n P$	Restriction numérique « au moins »	$\neg C$	Négation
$\leq n P$	Restriction numérique « au plus »	$C_1 \sqcup C_2$	Union
$= n P$	Restriction numérique « exacte »	$C_1 \sqcap C_2$	Intersection
$\{i_1, \dots, i_n\}$	Composition de classe		

**Table 1 :** Quelques constructeurs des LD

<b>Symbole</b>	<b>Signification</b>
$\mathcal{AL}$	Langage Attributif qui permet : la négation atomique, union, intersection,
$C$	négation de concept Complexe
$\mathcal{S}$	abréviation pour $\mathcal{ALC}$ étendue avec les propriétés transitives
$\mathcal{H}$	Hierarchie de rôles
$\mathcal{O}$	nominaux. (Classes dont l'extension est un seul individu)
$\mathcal{I}$	propriétés Inverses
$\mathcal{N}$	restrictions (non qualifiées) sur les Nombres (cardinalités)
$\mathcal{Q}$	restrictions Qualifiées sur les nombres.
$\mathcal{F}$	propriétés Fonctionnelles
$(\mathcal{D})$	utilisation des types de données (Datatypes)

**Table 2 :** Dénotation de l'expressivité des LD

Les LD ont l'avantage, dans leur grande majorité, d'être décidables et de donner lieu à des services de raisonnement complets et cohérents. Un système de LD comporte un système de raisonnement et une base de connaissance (KB). Un système de raisonnement, d'inférence ou encore un raisonneur, est l'outil logiciel qui réalise le processus d'inférence. La base de connaissances est composée de connaissances intentionnelles (TBox) et de connaissances assertionnelles (ABox). L'ontologie en constitue la partie intentionnelle (i.e. TBox). L'ABox comporte les individus et leurs descriptions. Le code OWL s'obtient à partir des formules de LD, en appliquant des correspondances (Table 3).

Syntaxe DL	Constructeur OWL	Syntaxe DL	Constructeur OWL
$\forall P.C$	owl:allValuesFrom	$C_1 \sqsubseteq C_2$	subClassOf
$\exists P.C$	someValuesFrom	$C_1 \equiv C_2$	equivalentClass
$\exists P.\{i\}$	hasValue	$P_1 \sqsubseteq P_2$	subPropertyOf
$\geq n P$	minCardinality	$P_1 \equiv P_2$	equivalentProperty
$\leq n P$	maxCardinality	$C_1 \sqsubseteq \neg C_2$	disjointWith
$= n P$	Cardinality	$\{i_1\} \equiv \{i_2\}$	sameAs
$C_1 \sqcap C_2$	intersectionOf	$\{i_1\} \sqsubseteq \neg \{i_2\}$	differentFrom
$C_1 \sqcup C_2$	unionOf	$P_1 \equiv P_2^{-}$	inverseOf
$\neg C$	complementOf	$P^{-} \equiv P$	symmetricProperty
$\{i_1, \dots, i_n\}$	one of	$P^+ \sqsubseteq P$	transitiveProperty

**Table 3 :** Correspondances entre les LD et OWL

### 3.4.1 Variantes du langage OWL

Le langage OWL est décliné en trois variantes avec une expressivité décroissante : Full, DL et Lite.

- OWL Full correspond au langage OWL « *complet* », c'est-à-dire qu'il autorise une libre combinaison des constructeurs du langage OWL et de RDF Schema sans imposer une séparation stricte entre les classes, les propriétés, les individus et les valeurs. Cet aspect viole les contraintes des raisonneurs de logiques de description (Smith *et al.*, 2004). OWL Full est destiné aux utilisateurs qui cherchent une expressivité maximale sans besoin de décidabilité.
- OWL DL, un sous langage d'OWL Full, impose des restrictions sur la combinaison des LD avec RDFS et nécessite de disjointre les classes, les propriétés, les individus et les valeurs. L'objectif premier d'OWL DL est de garantir l'expressivité maximale en préservant la décidabilité du langage. OWL DL correspond à la LD  $\mathcal{SHOIN}(\mathcal{D})$  tel que :  $\mathcal{S}$  désigne la logique de base  $\mathcal{ALC}$  étendue avec les rôles transitives,  $\mathcal{H}$  pour l'hierarchie de rôles,  $\mathcal{O}$  pour les nominaux,  $\mathcal{N}$  pour les restrictions (non qualifiées) sur les nombres et  $(\mathcal{D})$  pour les types de données (Table 2).

Par rapport à OWL Full, OWL DL impose les restrictions suivantes :

- Les quatre propriétés :  $\{owl:inverseOf, owl:InverseFunctionalProperty, owl:SymmetricProperty, owl:TransitiveProperty\}$  ne doivent jamais être spécifiées pour des  $owl:DatatypeProperty$ , car l'ensemble des  $\{owl:ObjectProperty, owl:DatatypeProperty\}$  sont disjoint.
- Pas de cardinalités sur les propriétés transitives, leurs inverses ou leurs super-propriétés
- Toutes les classes ou propriétés référencées doivent être typées respectivement comme des classes et propriétés OWL.
- Les axiomes doivent être bien formés, i.e. sans composante manquante ou supplémentaire et doivent constituer une structure d'arbre. Les axiomes sur l'égalité et la différence entre individus doivent être sur des individus nommés.

Le non-respect, en partie ou en totalité, de ces restrictions mène à une indécidabilité du langage ce qui explique la raison d'être de ces restrictions.

- OWL Lite, un sous langage d'OWL DL, supporte un sous ensemble de constructeurs d'OWL-DL. OWL Lite correspond à la LD  $SHIF(\mathcal{D})$ , qui correspond à  $SHOIN(\mathcal{D})$  sans les nominaux et avec les restrictions de nombre fonctionnelles uniquement. OWL Lite est destiné aux développeurs d'outils pour leur permettre de supporter OWL, en commençant par une version simplifiée des caractéristiques du langage (Smith *et al.*, 2004). OWL Lite peut répondre aux besoins des utilisateurs cherchant à exprimer des classifications et des contraintes simples. Par rapport à OWL DL, OWL Lite impose les restrictions suivantes :
  - L'utilisation de :  $\{owl:oneOf, owl:unionOf, owl:complementOf, owl:hasValue, owl:disjointWith, owl:DataRange\}$  est interdite.
  - L'utilisation de :  $owl:intersectionOf$  se fait uniquement dans le cas de listes contenant deux classes ou plus.
  - Les valeurs autorisées pour les cardinalités sont  $\{0, 1\}$
  - L'utilisation des identifiants de classes dans les restrictions est obligatoire.

### 3.4.2 Caractéristiques

Comme le langage OWL est basé sur les LD, il hérite de plusieurs caractéristiques des LD, comme les propriétés suivantes :

- *Logique monotone* : Le langage OWL, comme les LD, repose sur la logique monotone. Son principe est que chaque formule valide reste toujours valide même si on lui rajoute de nouveaux axiomes. La monotonie signifie que l'apparition de nouvelles connaissances ne réduit pas ce qu'on connaît déjà. En formule, si  $r \vdash c$  alors  $r, A \vdash c$
- *Négation logique* : OWL emploie la négation logique dont le principe est qu'une formule est vraie si et seulement si sa négation est fausse. Ce type de négation ne permet pas d'inférer si une formule est valide ou fausse si sa négation est inconnue.
- *Hypothèse du Monde Ouvert* (Open World Assumption- OWA) : dans cette hypothèse, on suppose que la connaissance stockée est incomplète et que l'absence d'information n'implique pas la fausseté. A titre d'exemple, soient  $C, D$  deux classes,  $i, t, v$  trois individus et  $P$  un prédicat. si  $C$  est définie comme :  $C \sqsubseteq \forall P.D$ , on ne peut pas conclure que  $i \in C$  même si on trouve que « pour tout  $t, P(i, t)$  il y a un  $t \in D$  ». La raison est qu'il peut y avoir un  $v$  non connu tel que  $P(i, v)$  mais  $v \notin D$ . Ce fait ne peut pas être déduit automatiquement vu l'OWA. Cette hypothèse est utile dans le cas où les connaissances stockées ne sont pas complètes. OWL se base sur cette hypothèse car on considère, dans le cadre du web sémantique, qu'il peut y avoir toujours une connaissance dont l'accès a échoué.
- *Hypothèse du Nom Unique* (Unique Name Assumption-UNA) : Dans cette hypothèse, on suppose que chaque individu nommé est différent des autres, i.e. deux noms différents correspondent à deux individus différents. A l'inverse des LD, OWL n'emploie pas cette hypothèse.

### 3.4.3 Limites

Par rapport à RDFS, le langage OWL donne la possibilité d'utiliser des axiomes pour bâtir de nouveaux concepts complexes et formuler des expressions relativement riches. Toutefois, OWL présente encore des limites dues principalement à ses caractéristiques. En voici quelques unes :

- *Classification des individus* : Le langage OWL permet d'exprimer des conditions nécessaires et/ou suffisantes sur les classes. Ce type de conditions permet, lors du processus d'inférence, de classer les individus dans les classes pour lesquelles ils vérifient leurs conditions. Néanmoins, ce n'est pas le cas pour les classes ayant des conditions nécessaires et suffisantes comportant l'un des constructeurs suivants : quantifieur universel ( $\forall$ ), négation ( $\neg$ ) ou la cardinalité maximale ( $\leq$ ). Ce fait est dû à l'emploi de l'hypothèse du monde ouvert. Pour remédier à ce problème, il est nécessaire de forcer l'utilisation de l'hypothèse du monde fermé (Closed World Assumption - CWA). A l'inverse de l'OWA, le CWA considère que la connaissance est complète et que tout ce qui n'est pas connu comme vrai est faux. Pour l'emploi du CWA nous proposons une solution qui s'appuie sur les deux points suivants :
  - Employer l'hypothèse du Nom Unique (UNA) : pour cela, il faut soit utiliser la propriété *owl:AllDifferentFrom*, pour lister tous les individus différents, soit de configurer les propriétés du raisonneur pour employer le UNA. Etant donné que le maintien d'une telle liste des individus différents est couteux en ressource, la deuxième alternative est sans doute la meilleure.
  - Clôturer le domaine : pour chaque individu et pour chaque propriété, il faut ajouter une contrainte de cardinalité maximale. Il est possible d'automatiser ce type de solution pour ajouter ces contraintes juste avant l'appel du raisonneur.
- *Modularité* : Si l'un des objectifs premiers des ontologies est de les partager et les réutiliser, on trouve que le langage OWL ne permet pas d'importer des parties d'ontologies mais l'intégralité. C'est aspect fait l'objet de recherches actuelles.
- *Expressivité* : Le langage OWL présente encore quelques limites identifiées, suite à son utilisation. C'est par exemple le cas pour le manque support des types de données XML Schema. Dans ce sens, une proposition d'une nouvelle version OWL 2, a été soumise au W3C. La proposition, OWL 2, vise à étendre la LD utilisée dans OWL-DL, de *SHOIN(D)* vers une LD plus expressive, *SROIQ(D)*. L'objectif est d'accroître l'expressivité du langage en maintenant sa décidabilité. Le W3C a mis récemment en place un groupe de travail pour une version recommandée. Cette extension fournit de nouvelles possibilités telles que les restrictions sur les *Datatype*, les restrictions de cardinalités qualifiées, les relations irreflexive, la combinaison de rôles de type  $R^\circ S \sqsubseteq R$  ou  $S^\circ R \sqsubseteq R$ , ... etc.

D'autre part, compte tenu de la forme restreinte des axiomes, qui doivent satisfaire une forme arborescente, il n'est pas permis d'introduire des chemins de propriétés, entre les individus, qui ne respectent pas cette contrainte. Cela montre la nécessité des formalismes de règles.

### 3.5 Les règles

Les règles, une famille de formalismes de représentation de connaissances issues de la programmation logique (PL), sont reconnues comme nécessaires pour des applications du monde réel. C'est notamment le cas dans le domaine des services web, l'informatique médicale et biomédicale et de façon générale pour relier les entrées et les sorties des processus.

Un programme logique ou un programme de règles est un ensemble fini de règles. Une règle est composée de deux parties : la condition et la conclusion, appelées aussi respectivement le corps et l'entête. Comme il n'existe pas, *pour l'instant*, de consensus sur le format des règles, la définition d'un format standard de règle est un sujet de recherche ouvert actuellement. Pour atteindre cet objectif, le W3C a mis récemment en place un groupe de travail intitulé « *Rules Interchange Format Group* » (RIF). En dehors des travaux du W3C, plusieurs formats de règles existent, avec des différences plus ou moins importantes. Un format général, cité dans (Antoniou *et al.*, 2005) est donné comme suivant :

$$H \leftarrow B_1 \wedge \dots \wedge B_m \wedge \sim B_{m+1} \wedge \dots \wedge \sim B_n \quad \text{tel que :}$$

H est appelé l'entête ou la conséquence.

$(B_1 \wedge \dots \wedge \sim B_n)$  est appelé le corps ou la condition.

H,  $B_i$ , avec  $1 \leq i \leq m$ , sont des littéraux classiques avec  $n \geq m \geq 0$  et les  $\sim B_j$ , avec  $m+1 \leq j \leq n$ , sont des littéraux de négation par l'échec. Les littéraux, ou littéraux classiques, peuvent être des atomes ou des atomes niés. Chaque atome correspond à un prédicat n-aire de la forme  $p(t_1, \dots, t_n)$ .

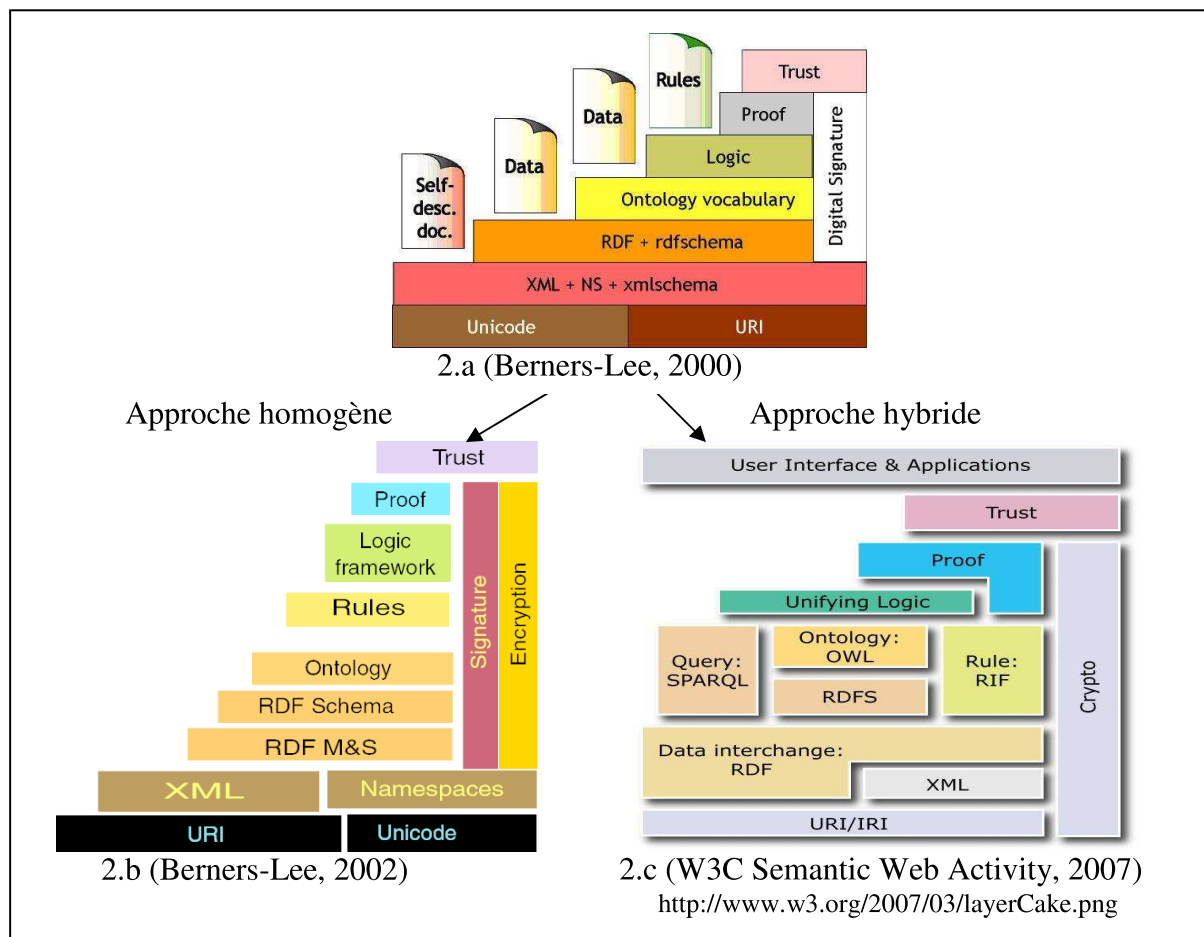
Une règle est dite sûre (*Safe Rule*) si les variables de la conclusion apparaissent dans quelques littéraux classiques de la condition.

Dans le projet du web sémantique, la problématique posée est de trouver une façon adéquate pour combiner les règles et OWL. En terme d'architecture elle se traduit par la question : Faut-il considérer les règles comme une couche au dessus du langage OWL ou bien les considérer comme deux piles, côte à côte (Figure 2) (Horrocks *et al.*, 2005) ?

Contrairement à OWL, les règles ne permettent pas de déclarer l'existence d'individus dont l'identité peut être non connue. Elles ne permettent pas d'exprimer les cardinalités, le quantifieur universel. De plus, les règles affectent seulement les individus. Certaines variantes de Datalog, un langage de règles bien connu dans le domaine des bases de données, n'autorisent pas l'utilisation de la négation.

### 3.6 Combinaison de règles et d'ontologies

Les ontologies OWL et les règles, sont deux formalismes différents pour la représentation de connaissances, basés respectivement sur les logiques de description (LD) et la programmation logique (PL). Deux articles récents (Antoniou *et al.*, 2005 ; Rosati, 2005) fournissent une classification des propositions existantes pour la combinaison des ontologies et des règles. Il existe principalement deux approches : homogène et hybride (figure 2).



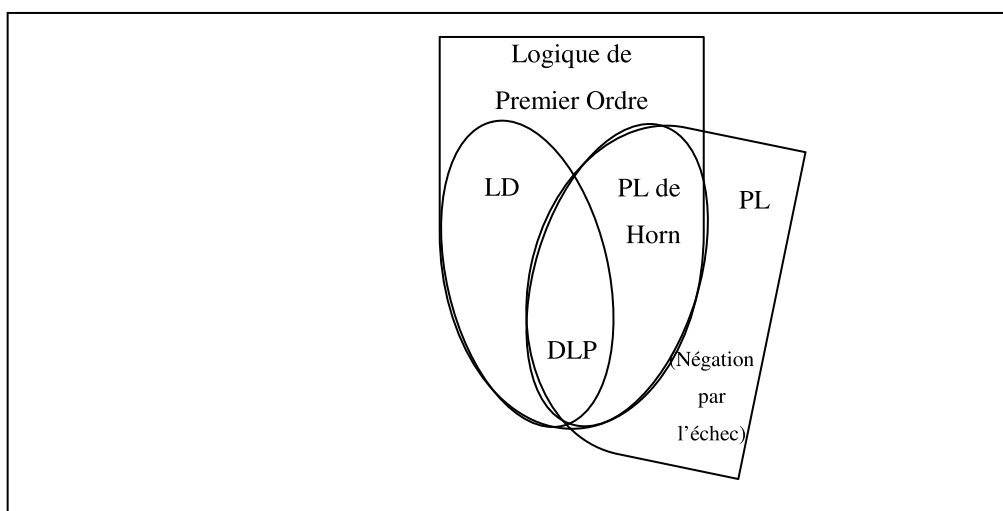
**Figure 2 :** Les deux approches combinant les ontologies et les règles

### 3.6.1 Approche homogène

Dans cette approche, les ontologies et les règles sont décrites dans un langage logique commun sans distinction entre les prédicats de l'ontologie et les prédicats de règles. Pour faire du raisonnement, un raisonneur général du langage commun peut être réutilisé ou un nouveau raisonneur spécialisé doit être développé. L'inconvénient de ce type d'approche est que le nouveau langage est indécidable ou bien son expressivité est limitée. Dans ce type d'approche, on trouve les propositions des langages SWRL et DLP

#### DLP

DLP, *Description Logic Programming*, (Grosf *et al.*, 2003) est défini comme une intersection entre la LD *SHIQ* et les clauses de Horn. La motivation principale du DLP est de faciliter l'interopérabilité entre les langages basés sur les LD et ceux basés sur les règles. Cela offre une traduction bidirectionnelle entre les règles et les ontologies et permet de construire les ontologies en dessus des règles et vice versa. De cette manière il est possible d'utiliser les outils de PL pour faire du raisonnement sur des ontologies LD de grande taille. Le langage DLP est décidable mais il fournit un support partiel pour le raisonnement OWL car la restriction universelle (lorsqu'elle est à gauche dans une formule), la restriction existentielle (lorsqu'elle est à droite), la négation et les cardinalités ne peuvent pas être utilisés car elles n'ont pas d'équivalences en règles.



**Figure 3** : Définition du langage DLP comme intersection de LD et PL (Grosf *et al.*, 2003)

## SWRL

SWRL, *Semantic Web Rule Language*, (Horroks *et al.*, 2004) est une extension au langage OWL pour formuler des règles de type Horn en limitant les prédicats aux classes et propriétés OWL. SWRL étend OWL en combinant les sous langages OWL (OWL DL et OWL Lite) avec le langage de règle RuleML (*Rule Markup Language*). Les spécifications du langage SWRL sont soumises comme proposition au W3C en 2004. La sémantique du langage SWRL, comme celle du langage OWL, est basée sur l'hypothèse du monde ouvert (OWA) et donc pas de possibilité pour employer la négation par l'échec. D'un autre côté, étant donné que SWRL n'impose aucune restriction sur les règles, le langage devient indécidable. Pour maintenir la décidabilité, la solution est de restreindre la partie règle aux règles dites *DL-Safe* comme proposé dans les travaux de KAON2.

### L'approche KAON2

Dans l'article (Motik *et al.*, 2004), l'auteur montre que la combinaison de la LD  $\mathcal{SHOIN}(\mathcal{D})$  avec les règles *DL-Safe* est décidable. Une règle  $r$  est appelée *DL-Safe* si chaque variable dans  $r$  apparaît dans un atome non LD dans le corps de la règle. Un atome LD est de la forme  $A(s)$  ou  $R(s, t)$ . La restriction aux règles *DL-Safe* permet de conserver la décidabilité. L'auteur propose ainsi un algorithme de raisonnement limité à la LD  $\mathcal{SHIQ}(\mathcal{D})$  correspondant à OWL DL sans le support des nominaux. Cet algorithme a été implémenté dans KAON2<sup>63</sup> et se base sur la réduction de la base de connaissance en un programme Datalog disjonctif. KAON2 est une infrastructure pour gérer les ontologies OWL DL et SWRL. Il supporte OWL Lite et la LD  $\mathcal{SHIQ}(\mathcal{D})$  un sous ensemble d'OWL DL étendues avec les règles *DL-Safe*, un sous ensemble de SWRL.

### 3.6.2 Approche hybride

Une séparation stricte est nécessaire dans ce type d'approche entre les prédicats de l'ontologie et les prédicats de règles. Cela se fait en combinant un langage d'ontologie et un langage de règles déjà existants. Le raisonnement se fait en interfaçant les systèmes de raisonnement d'ontologie existants avec les systèmes de règles existants. Ce type d'approche est bien adapté pour la construction d'outils pour l'accès aux systèmes d'information hétérogènes. C'est notamment le cas des systèmes basés sur la LD, disposant de programmes de règles

---

<sup>63</sup> KAON, <http://kaon2.semanticweb.org/>

spécifiques à des applications et ayant besoin pour intégrer les deux. Les propositions employant cette approche sont AL-Log, CARIN, dl-programs et récemment les bases de connaissances r-hybride.

### **AL-Log**

Le langage  $\mathcal{AL}\text{-Log}$  (Donini *et al.*, 1998) est une intégration hybride de la logique de description  $\mathcal{ALC}$  et des règles Datalog.  $\mathcal{ALC}$  est une simple variante de LD admettant les constructeurs suivants : union, intersection, négation, quantifieurs universel et existentiel. Les contraintes LD admises, dans le corps des règles AL-Log, sont limitées aux atomes unaires d'appartenance des individus aux classes.  $\mathcal{AL}\text{-Log}$  a l'avantage d'être décidable mais il restreint considérablement l'expressivité offerte par les deux formalismes.

### **CARIN**

CARIN (Levy *et al.*, 1998) est défini comme une famille de langages pour fournir une intégration hybride de Datalog avec différentes logiques de description. La classe de LD considérée inclut tout sous ensemble de la logique  $\mathcal{ALCN}\mathcal{R}$  admettant les constructeurs suivants : union, intersection, négation, quantifieurs universel et existentiel, la restriction de cardinalité non qualifiée et le constructeur d'intersection de propriétés. CARIN étend AL-Log en intégrant Datalog avec une LD plus expressive et en admettant des contraintes LD plus générales dans le corps des règles hybrides. De cette façon, CARIN devient indécidable. Les sous ensembles décidables de CARIN peuvent être obtenus en imposant des restrictions aux règles non récursives ou sur la façon dont les variables apparaissent dans les règles.

### **dl-programs**

L'article (Eiter *et al.*, 2004) propose les programmes de logique de description (*dl-programs*) comme une solution d'intégration hybride des LD  $\mathcal{SHOIN}(\mathcal{D})$  et  $\mathcal{SHIF}(\mathcal{D})$ , qui correspond respectivement aux sous langages OWL DL et OWL Lite, avec l'*Answer Set Programming*. L'*Answer Set Programming (ASP)* est un langage de programmation logique déclaratif basé sur la sémantique du modèle stable (*answer set*) de la programmation logique. Un programme de logique de description (*dl-program*) est constitué d'une base de connaissance LD et d'un ensemble fini de règles de logique de description (*dl-rules*). Les règles sont définies comme dans la programmation logique avec la négation par l'échec et peuvent aussi contenir des requêtes sur la base de connaissance avec la négation classique dans le corps.

L'implémentation de ce type d'approche se base sur l'utilisation d'un raisonneur LD et d'un moteur *Answer Set Programming* avec une séparation de raisonnement sur les deux parties.

### **Les bases de connaissances r-hybrid**

Les bases de connaissances r-hybrid (Rosati, 2005) sont proposées dans le cadre d'intégration hybride des ontologies et des règles. Une base de connaissances r-hybrid est constituée d'une composante structurelle (ontologie) et d'une composante règle. La composante ontologie peut être quelconque langage de LD et la composante règle consiste en un programme Datalog<sup>¬∨</sup>, i.e. un programme Datalog dans lequel la négation par l'échec dans le corps de règle et la disjonction dans l'entête de règle sont autorisées. L'auteur montre que si le raisonnement dans la logique L, utilisée pour spécifier la partie structurelle T, est décidable alors le raisonnement dans l'extension de T avec les règles Safe-Datalog<sup>¬∨</sup> (Datalog augmenté de la négation et disjonction) reste décidable. Dans ce dernier type de règles, chaque variable dans la règle doit apparaître dans un atome positif non LD. En effet, la libre interaction entre les deux composantes, structurelle et règle, mène à une indécidabilité importante, comme dans le cas de CARIN. Ainsi la décidabilité ne peut être maintenue que si on limite considérablement l'expressivité de l'une des deux composantes.

### **3.6.3 Discussion sur les approches combinant règles et ontologies**

Comme il est montré par l'approche DLP, l'intersection entre les formalismes des ontologies et de règles est trop limitée. Cela traduit l'incapacité de chacun des deux formalismes à inclure l'autre, ce qui implique la nécessité de combinaison des deux formalismes pour avoir le maximum d'expressivité. D'autre part, comme les deux formalismes ont des sémantiques et des caractéristiques différentes, on trouve que l'ensemble des approches proposées basculent entre maintien de la décidabilité et l'expressivité de la combinaison.

Devant la nécessité d'intégrer les ontologies et les règles (Rosati, 2005 ; Horrocks, 2005) et de considérer les deux types de négation (Wagner, 2003), nous partageons avec plusieurs auteurs (Kifer *et al.*, 2005) que les règles, étudiées depuis une trentaine d'années, ne peuvent pas être conçues comme une couche au-dessus du langage OWL. Ainsi, peut-on conclure que l'approche hybride est la mieux adaptée pour combiner indépendamment les ontologies et les règles avec un besoin d'interopérabilité. La version actuelle du *Semantic Layer Cake W3C* se focalise dans le sens de cette conclusion (figure 2.c) et considère les règles et les ontologies comme deux piles indépendantes côte à côte.

## 3.7 Les outils du web sémantique

### 3.7.1 Editeurs d'ontologie

Différents outils d'édition d'ontologies ont été proposés pour assister les utilisateurs dans la création et la manipulation des ontologies. Ils génèrent le code correspondant et permettent généralement d'interagir avec les systèmes de raisonnement. Parmi ces outils, on peut citer à titre d'exemples :

- OilEd (Bechhofer *et al.*, 2001) est développé initialement comme un éditeur d'ontologies OIL à l'université de Manchester. Il a été amélioré pour prendre en compte les formats d'ontologies DAML+OIL et OWL et tester leur cohérence à l'aide du raisonneur intégré FaCT. OilEd est un outil open source mais son développement n'est plus maintenu.
- OntoEdit (Sure *et al.*, 2002) est un environnement extensible basé sur une architecture de plug-in pour la construction d'ontologie. Développé par l'université Karlsruhe (Allemagne), OntoEdit est commercialisé actuellement, sous le nom OntoStudio, par la société Ontoprise avec deux versions dites de « base » et « complète ». OntoStudio version « basic » gère de nombreux formats comme RDF(S), DAML+OIL, OWL et F-Logic. La version complète contient un serveur d'édition d'ontologie multi utilisateur et un plug-in pour le test de cohérence d'ontologie.
- Protégé-OWL (Knublauch *et al.* 2004) est un éditeur d'ontologie OWL développé au Stanford Medical Informatics de l'université de Stanford. Il permet d'éditer des ontologies en RDFS et OWL ; de définir des axiomes comme expressions OWL ; de définir des individus OWL ; et d'utiliser les raisonneurs. Protégé est une application autonome et open source avec une architecture extensible. De nombreux plug-ins sont disponibles comme le plug-in SWRLTab, une extension de protégé pour l'édition et l'exécution des règles SWRL, et le plug-in JessTab, permettant d'utiliser le système de règle Jess. La version 4.x, sous développement, vise à améliorer la modularité et l'extensibilité du système et prendre en considération les caractéristiques prévues dans OWL 2.
- SWOOP (Kalyanpur *et al.*, 2005) est un éditeur d'ontologie en ligne développé par l'équipe de recherche Mindswap de l'Université de Maryland. Il utilise une interface de navigateur web et intègre le seul raisonneur Pellet (Parsia *et al.*, 2004). L'équipe envisage son développement pour s'interfacer avec d'autres raisonneurs.

- WebODE (Arpirez *et al.*, 2001), successeur du ODE (Ontology Design Environment), est une suite d'ingénierie d'ontologies avec une architecture extensible, développé à l'Université Polytechnique de Madrid. WebODE s'utilise uniquement comme application web avec trois interfaces : un éditeur basé sur le format HTML pour éditer tous les termes d'ontologies sauf les axiomes et les règles ; l'interface OntoDesigner, pour éditer les taxonomies de concepts et les relations ; et l'interface WebODE Axiom Builder pour créer des axiomes et les règles dans la logique du premier ordre. WebODE permet de faire le développement des ontologies qu'il sauvegarde dans une base de données oracle et ne manipule pas de langage d'ontologie. Cependant il permet d'importer et d'exporter des ontologies depuis et vers des formats comme RDFS, DAML+OIL et OWL.

### 3.7.2 Systèmes d'inférence

Comme le langage OWL est basé sur les LD, il n'est pas surprenant que les raisonneurs de LD soient les plus utilisés, vu leur décidabilité. Dans ce cadre, différents systèmes d'inférence ont été proposés. Ils se différencient principalement dans l'algorithme de raisonnement implémenté, la LD supportée et le mode d'utilisation. Les raisonneurs<sup>64</sup> de LD, les plus connus, sont : FaCT++, RacerPro et Pellet.

- FaCT++ (Tsarkov *et al.*, 2005) est le successeur du raisonneur FaCT. Développé en C++ et open source, FaCT++ implémente la LD  $\mathcal{SHOIQ}(\mathcal{D})$  qui correspond à OWL DL avec en plus, la restriction qualifiée de cardinalité. Il supporte le raisonnement sur la TBox seulement. Dans sa version la plus récente (septembre 2008), FaCT++ supporte désormais une LD plus expressive  $\mathcal{SROIQ}(\mathcal{D})$ .
- RacerPro (Haarslev *et al.*, 2001) est un raisonneur commercial (version d'évaluation et licence gratuites pour la recherche) développé en Lisp et implémente la logique  $\mathcal{SHIQ}(\mathcal{D})$ , i.e. OWL DL sans les nominaux et avec en plus, la restriction qualifiée de cardinalité. RacerPro supporte le raisonnement simultané sur la TBox et l'ABox. Pour interroger la base de connaissances, il faut utiliser impérativement le langage nRQL (new Racer Query Language), un langage propre à Racer.
- Pellet (Parsia *et al.*, 2004) est un raisonneur LD open source, développé en java. Il implémente la LD  $\mathcal{SROIQ}(\mathcal{D})$  qui correspond à la proposition OWL 2 et supporte des

---

<sup>64</sup> La liste complète des raisonneurs, maintenue par Uli Sattler, est disponible à cette adresse web : <http://www.cs.man.ac.uk/~sattler/reasoners.html>

inférences sur la TBox et l'ABox. Pellet a récemment inclut le support d'un fragment des règles DL-Safe de SWRL. Il permet d'interroger la base de connaissance avec le langage SPARQL (Prud'hommeaux *et al.*, 2008) et son prédécesseur RDQL.

Il existe également d'autres systèmes non LD basés sur d'autres formalismes, comme F-OWL (Zou *et al.*, 2005), Hoolet, Surnia mais qui ne sont pas décidables.

L'utilisation d'un raisonneur peut se faire de plusieurs manières : comme un API ou à l'aide d'une interface DIG. Cette dernière est une interface XML standardisée, développée par le *DL Implementation Group*<sup>65</sup>. Elle consiste en une interface commune aux raisonneurs de logique de description et permet aux outils tels que les éditeurs d'ontologie de faire usage des raisonneurs. La version 2 de l'interface DIG, sous développement, vise à améliorer la prise en charge de formats de types de données.

Les tâches qui peuvent être confiées au raisonneur sont :

- La vérification de la consistance de l'ontologie i.e. vérifier la syntaxe et l'usage des termes OWL et s'assurer que les instances respectent les restrictions.
- La vérification des relations inattendues entre les classes.
- La réduction des redondances dans les définitions de l'ontologie, découvrir les descriptions équivalentes, réutiliser les descriptions des concepts.
- Le classement automatique des instances dans les classes.

Le choix d'un raisonneur doit se faire en prenant en considération plusieurs paramètres comme : le sous langage OWL, les caractéristiques de l'ontologie (taille, complexité,...) et de la tâche à effectuer. L'article (Gardiner *et al.*, 2006) propose un cadre pour une comparaison automatique entre les différents raisonneurs, sur un ensemble de critères tel que la performance et la vitesse. Cette étude était limitée au raisonnement sur la TBox et elle nécessite la prise en compte de l'ABox. Parmi les défis soulignés dans ce domaine, nous retrouvons la problématique du passage à l'échelle du web et de la manipulation d'ontologies de grandes tailles.

---

<sup>65</sup> DL Implementation Group, <http://dl.kr.org/dig/index.html>

## 4. Evaluation des technologies par rapport aux objectifs

### 4.1 Formalismes de représentation de connaissances

La diversité des formalismes de représentation de connaissances nous amène à poser la question suivante : quel formalisme doit-on choisir ? RDFS, OWL (avec ses 3 dialectes), règles ou une combinaison LD/règle ? La réponse à cette question n'est pas standard et ne peut être décidée sans l'étude des besoins et la compréhension du contexte dans lequel nous travaillons. En effet, selon la complexité des connaissances à représenter, l'un ou l'autre de ces formalismes suffira ou non ! Allant de la simple capacité de représenter des taxonomies par RDFS, à l'utilisation conjointe des logiques de description dans OWL, ces formalismes peuvent néanmoins rester insuffisants dans des contextes nécessitant d'utiliser des règles. L'idée est de commencer premièrement par modéliser les connaissances, sous forme d'un graphe de concepts et de relations entre concepts, et voir par la suite, si nous avons besoin ou non d'utiliser les LD et les règles.

Pour représenter le contexte applicatif et décrire les besoins de qualité et d'aide à la décision, nous avons commencé par quelques classes primitives pour en dériver d'autres plus complexes. Dans notre cas et vu les limites du langage OWL, l'utilisation d'un langage de règles s'est fortement imposée. Ainsi, nous avons opté pour l'approche hybride pour la combinaison des ontologies et règles. C'est un choix tout à fait naturel pour pouvoir utiliser deux formalismes de représentation de connaissances ayant des caractéristiques totalement différentes.

### 4.2 Approche de médiation

Pour construire une représentation unifiée d'un ensemble de sources de données, deux approches principales existent : (i) une approche ascendante, dans la quelle on part des sources et on construit un schéma intégré de l'ensemble des sources, indépendamment de toute spécificité applicative, (ii) et une approche descendante, tel que on part d'un problème spécifique vers les sources locales. Dans cette approche, on construit un schéma global, représentant une description spécifique à une application ou un domaine particulier, pour mettre en correspondance par la suite les sources qui en relèvent.

La première approche guidée par les sources n'est pas extensible en termes du nombre de sources, ne fournit pas une intégration sémantique des sources dans un contexte applicatif, ne

justifie pas l'intégration de toutes les sources qui peuvent être en désaccord total avec le problème étudié, ne fournit pas une stabilité au système vis-à-vis l'observation de nouvelles sources qui peut amener à reconsidérer le schéma global.

La deuxième approche guidée par les besoins préliminaires assume la création d'un schéma global qui supporte l'interaction entre l'application et les sources en se basant sur le sujet et la description du domaine de l'application. Cette approche évite les problèmes de l'approche précédente.

En termes de techniques de médiation, il existe deux approches principales pour établir des correspondances entre le schéma global et les schémas des sources de données locales : le *Global As View* et le *Local As View*. Une généralisation de l'approche LAV a été également proposée, *Generalized Local As View*. (Lenzerini, 2002).

- **Global As View (GAV)** : dans GAV, qui correspond à l'approche ascendante, on décrit chaque élément  $g$  du schéma global  $G$  comme une vue sur les sources de données locales. GAV est une approche procédurale puisque le mapping indique explicitement au système comment extraire les données lorsqu'on évalue les éléments du schéma global. De cette façon, l'approche GAV facilite l'interrogation des sources locales, car l'évaluation d'une requête globale passe par un simple dépliement de la requête en remplaçant les vues par leur définition.
- **Local As View (LAV)** : dans LAV, qui correspond à l'approche descendante, on décrit chaque élément  $s$  d'une source  $S$  comme une vue sur le schéma global. L'approche LAV est déclarative puisque le contenu de chaque source est caractérisé en termes de vue sur le schéma global. Ainsi l'insertion d'une nouvelle source ne nécessite pas la modification du schéma global, mais uniquement la définition, au médiateur, d'une nouvelle vue sur le schéma global. De cette façon, l'approche LAV favorise l'extensibilité du système à large échelle.
- **Generalized Local As View (GLAV)** : GLAV est une généralisation de l'approche LAV autorisant des correspondances entre des vues sur les sources locales avec des vues sur le schéma global. Le traitement de requête passe par une réécriture puis par un dépliement. L'objectif de cette approche étant de combiner les avantages des approches LAV et GAV, cependant la réponse aux requêtes n'est pas toujours possible et la modification du schéma global nécessite la modification des expressions GLAV.

Même si l'approche GAV apparaît plus naturelle et sûrement plus simple à implémenter, elle 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 de nouveaux critères de qualité, ...). Ainsi, nous avons préféré l'approche LAV par rapport aux approches GAV et GLAV.

### 4.3 Architecture globale

Afin de supporter l'approche d'intégration de données *quality-aware*, nous proposons une architecture couplant un système de raisonnement avec un système d'intégration (figure 4). Schématiquement, on formule des requêtes à partir du schéma global. En nous appuyant sur une ontologie d'application, nous traitons les deux premières étapes de NARI, d'« identification / préparation des données » (cfr. §5.2 du Chapitre 1), tandis que nous utilisons le système de médiation pour traiter essentiellement le niveau « intégration de données » de NARI.

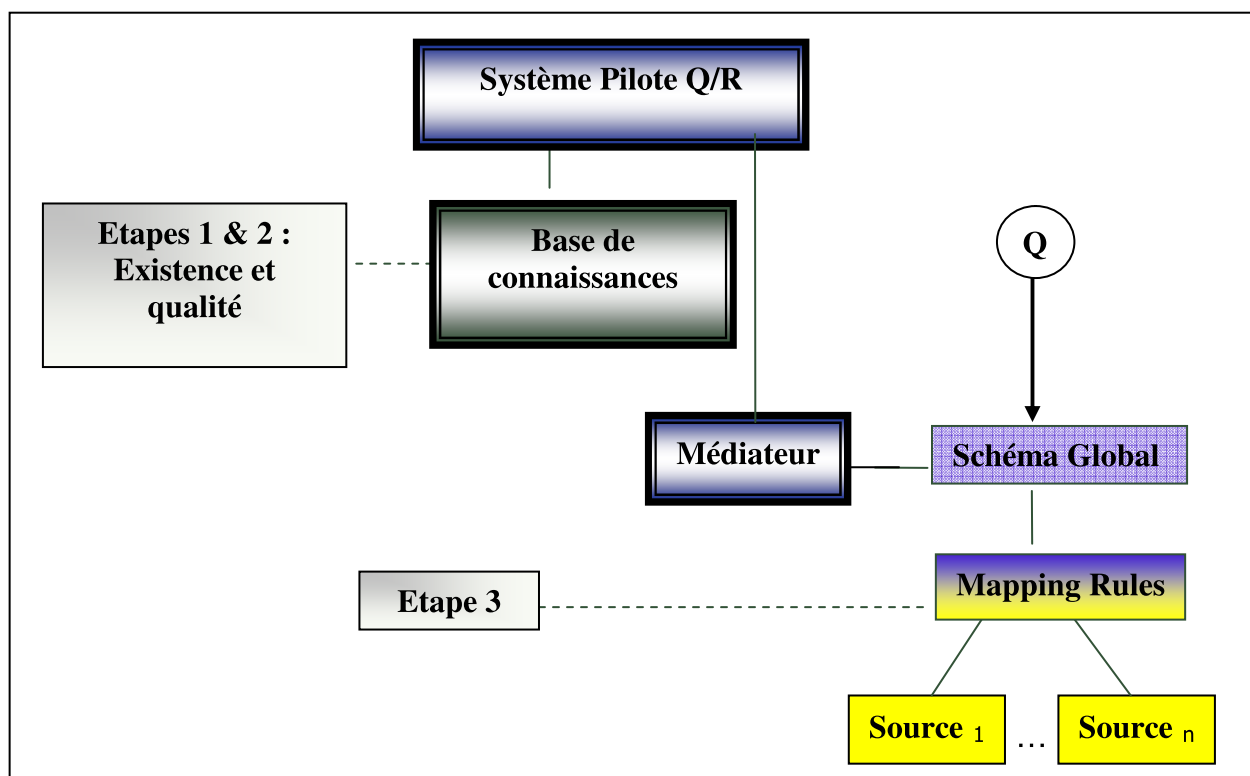


Figure 4 : Infrastructure technologique à trois niveaux

L'architecture globale comprend trois niveaux : application, médiation et sources locales.

### 4.3.1 Niveau application

Le niveau applicatif est celui dans lequel sont formalisés les besoins pour la prise de décision.

Il contient :

- Une base de connaissances basée sur l'ontologie d'application et qui comprend : (i) Une conceptualisation des différentes classes et rôles décrite avec un langage de LD, (ii) un ensemble des instances et (iii) un ensemble de règles.
- Un système de raisonnement LD pour vérifier la consistance de l'ontologie d'application, la satisfiabilité des concepts et inférer de nouvelles connaissances.
- Un système de règles pour inférer de nouvelles connaissances à partir de l'ensemble des règles.
- Un système pilote qui gère la base de connaissance et communique avec le système de médiation

**Ontologie d'application** : Nous démarrons par un nombre limité de classes et de relations et nous en dérivons ensuite de nouvelles classes, pour supporter les deux premières étapes de NARI, relatives respectivement à l'existence et à la qualité des sources de données.

### 4.3.2 Niveau de médiation

Le système de médiation comprend : (i) le médiateur (ii) un schéma global, et (iii) un ensemble de relations de mappage entre le schéma global et les sources de données locales.

**Schéma Global** : est donné par l'ontologie de domaine (Visser, 2004). Celle-ci, développée indépendamment des sources de données, fournit une vue unifiée pour formuler des requêtes au niveau global. Dans notre cas, il s'agit 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 analyses syntaxiques et lexicales sur les requêtes globales. Un concept clé, comme le Territoire, est commun aux deux ontologies, d'application et de domaine.

**Schéma de Médiation** : La médiation entre le schéma global et les sources locales se fait en suivant l'approche LAV et les principes d'une technique d'intégration décrite par (Amann *et al.*, 2002). Un ensemble de règles de mappage (*Mapping Rules*) relie 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 d'OQL (Berler *et al.*, 2000). 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.

### 4.3.3 Niveau sources de données locales

Elles sont référencées dans des catalogues et décrites par leurs schémas. Les schémas sont complétés par un ensemble opportun de métadonnées, correspondant aux critères de qualité. 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.

## 4.4 Choix d'implémentation technique

### 4.4.1 Choix des langages de représentation des connaissances

L'ontologie d'application est implémentée conjointement avec deux langages : (i) un langage de logiques de description OWL DL pour structurer les classes et les rôles et (ii) un langage de règle pour permettre la déduction de connaissances. OWL DL (Smith *et al.*, 2004), est un sous langage d'OWL, basé sur la LD  $\mathcal{SHOIN}(\mathcal{D})$ , auquel on peut associer des services de raisonnement, car il est décidable. En effet, l'expressivité offerte par les LD ne permet pas d'exprimer des règles et dans l'autre sens on a trouvé que les règles ne permettent pas non plus de capter les formules LD. Comme la logique  $\mathcal{SHIF}(\mathcal{D})$ , à qui correspond OWL Lite, n'inclut pas l'utilisation des cardinalités ( $\mathcal{N}$ ) nécessaire pour formuler des expressions CWA, il est donc indispensable d'utiliser OWL DL correspondant à la LD  $\mathcal{SHOIN}(\mathcal{D})$  (Baader *et al.*, 2005).

Comme OWL n'adopte pas l'hypothèse du monde fermé (CWA), la solution, citée dans le § 3.4.3, est d'employer l'hypothèse du nom unique, *Unique Name Assumption (UNA)*, qui assure l'unicité des individus nommés, et de clôturer le domaine. Pour forcer l'adoption de

l'hypothèse du nom unique, le mieux est de configurer le système de raisonnement pour que celui-ci utilise l'UNA.

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.

#### 4.4.2 Infrastructure technique

Nous utilisons l'éditeur d'ontologies Protégé (Knublauch *et al.*, 2004) pour définir la base de connaissances, Pellet (Parsia *et al.*, 2004) comme un système de raisonnement et Jess comme un système de règle.

Protégé est un environnement de développement open source pour construire des ontologies décrites en OWL-DL et des systèmes à base de connaissances, supportant  $\mathcal{SHOIQ}(\mathcal{D})$ . Il possède une architecture extensible et peut être utilisé en conjonction avec des systèmes de raisonnement, au travers de l'interface standardisée, DIG (Description logics Implementation Group)

Pellet (Parsia *et al.*, 2004) est un système d'inférence open source qui implémente une LD décidable et la plus expressive actuellement, i.e. la LD  $\mathcal{SR}OIQ(\mathcal{D})$ , et permet de faire des interrogations sur la base de connaissances à l'aide du langage de requête standardisé SPARQL.

JESS, *Java Expert System Shell*, est un système de règle commercial avec la possibilité d'obtenir une licence gratuite pour une utilisation académique. Le système Jess possède son propre format de règles. L'utilisation du système de règle Jess, dans notre cas, est motivée par le fait que l'éditeur d'ontologie Protégé implémente des plugins servant d'interfaces entre l'ontologie OWL et ses instances d'un côté et le système de règle Jess d'autre côté.

Pour maintenir un lien entre le système de raisonnement et le système de règles, nous utilisons les plug-ins JessTab et SWRLJessTab (Golbreich, 2004) sous Protégé. Le plugin JessTab permet d'importer l'ontologie OWL DL et ses instances dans la base de faits du système de règle Jess et dans l'autre sens de transférer les résultats vers la base de connaissances. Le plug-in SWRLJessTab permet de formuler des règles en SWRL et de les traduire ensuite vers le format Jess. Le plug-in SWRLJessTab nécessite encore du développement pour prendre en considération toutes les correspondances entre les formats SWRL et Jess. Il est également

possible de saisir directement les règles dans le propre format de Jess. A noter que le système de règle Jess accepte la négation par l'échec même si cette dernière n'est pas autorisée en SWRL. Une mise à jour interactive se fait en continu entre la base de faits Jess et l'ontologie avec ses instances.

Ce type de solution correspond à une combinaison hybride des ontologies et des règles avec une indépendance entre les deux systèmes. La solution répond aux besoins et objectifs de l'approche proposée surtout que les travaux de recherche sur la standardisation des règles et leur intégration avec les ontologies n'ont pas encore abouties.

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

## 5. Conclusions

Dans ce chapitre nous avons décrit les objectifs technologiques de NARI, en partant des objectifs généraux présentés dans le Chapitre 1. Ceux-ci, nous ont permis de mettre en évidence l'intérêt d'une approche basée sur les ontologies. Ensuite, nous avons analysé et évalué les technologies du web sémantique, pertinentes pour NARI, avec leurs rôles, avantages et limites, présenté les principales approches pour l'intégration de données et l'architecture globale retenue pour NARI.

Ainsi, le formalisme de représentation des connaissances correspondant aux logiques de descriptions nous semble particulièrement bien adapté à un environnement basé sur les ontologies où l'intégration de données et le raisonnement sont omniprésents. En particulier, les propriétés de décidabilité et cohérence/complétude des algorithmes d'inférence procurent à cette démarche des avantages sur les autres solutions gérant des bases de connaissances. Un autre aspect prépondérant consiste en l'existence de nombreux outils exploitant des standards reconnus, en particulier ceux du W3C.

Evidemment, les logiques de descriptions présentent leurs limites, avec en particulier l'absence d'un standard pour un couplage efficace et expressif avec un système à bases de règles.

L'architecture proposée pour NARI est en adéquation à l'architecture basée sur les services, préconisée par la directive européenne INSPIRE, concernant les données et les services géoréférencés. Elle comporte des modules répondant à des besoins fonctionnels diversifiés (identification des sources, transformations classifications, intégration de données). Des évolutions futures de cette architecture pourraient concerner l'ajout de composantes de visualisation, facilitant les interactions entre le système et les utilisateurs, ainsi que des composantes de géolocalisation, pour compléter, si nécessaire, la description des objets traités par NARI par des coordonnées géographiques.

## 6. Références

### Références du W3C

- Berners-Lee, T. Semantic Web talk. Invited Talk at XML 2000 Conference, Slides: <http://www.w3.org/2000/Talks/1206-xml2k-tbl/slide10-0.html>.
- Berners-Lee, T. *Semantic Web talk*. 2002 Slides: <http://www.w3.org/2002/Talks/04-sweb/slide12-0.html>
- Bray, T., Hollander, D., Layman, A., Tobin, R. *Namespaces in XML 1.0*. Recommendation, W3C, 2<sup>ème</sup> Edition, 2006. <http://www.w3.org/TR/REC-xml-names/>
- Bray, T., Paoli, J., Sperberg-McQueen C., Maler, E., Yergeau, F. *Extensible Markup Language (XML) 1.0*. Recommendation, W3C, 4<sup>ème</sup> Edition, 2006. <http://www.w3.org/TR/REC-xml/>
- Brickley, D., Guha, R.V. *RDF Vocabulary Description Language 1.0: RDF Schema*. Recommendation, W3C, 2004. <http://www.w3.org/TR/rdf-schema/>
- Horrocks, I., Patel-Schneider, P.F., Boley, H., Tabet, S., Grosz B., Dean, M. *SWRL : A Semantic Web Rule Language. Combining OWL and RuleML*. W3C Member Submission, 2004. <http://www.w3.org/Submission/SWRL/>
- Klyne, G., Carroll, J. *Resource Description Framework (RDF) : Concepts and Abstract Data Model*. Recommendation, W3C, 2004. <http://www.w3.org/TR/rdf-concepts/>
- Prud'hommeaux, E., Seaborne, A. *SPARQL Query Language for RDF*. Recommendation, W3C, 2008. <http://www.w3.org/TR/rdf-sparql-query/>

Smith, M.K., Welty, C., McGuinness, D. *OWL Web Ontology Language Guide*. Recommendation, W3C, 2004. <http://www.w3.org/TR/owl-guide/>

Thompson, H., Beech, D., Maloney, M., Mendelsohn, N. *XML Schema part 1: Structures*. Recommendation, W3C, 2004. <http://www.w3.org/TR/xmlschema-1/>

## Articles

Amann, B., Beeri, C., Fundulaki, I., Scholl, M. *Ontology-based integration of xml web resources*. In Proceedings of the 1<sup>st</sup> International Semantic Web Conference ISWC, Sardinia, Italy, Juin 2002. *Lecture Notes in Computer Science*, volume 2342, Springer, Berlin, 2002, pp. 117-131.

Antoniou, G., Damásio, C.V., Grosz, B., Horrocks, I., Kifer, M., Maluszyński, J., Patel-Schneider P.F. *Combining Rules and Ontologies. A survey*. REVERSE, 2005. <http://reverse.net/deliverables/m12/i3-d3.pdf>

Arpirez, J.C., Corcho, O., Fernandez-Lopez, M., Gomez-Perez, A. *WebODE: a scalable ontological engineering workbench*. In: 1<sup>st</sup> International Conference on Knowledge Capture (KCAP'01), ACM Press, Victoria, 2001, pp. 6-13.

Baader, F., Horrocks, I. Sattler, U. *Description Logics as Ontology Languages for the Semantic Web*. *Lecture Notes in Artificial Intelligence*, volume 2605, Springer, Berlin, 2005, pp. 228-248.

Bechhofer, S., Horrocks, I., Goble, C., Stevens, R. *OilEd: a reason-able ontology editor for the Semantic Web*. In Joint German/Austrian conference on Artificial Intelligence, *Lecture Notes in Artificial Intelligence*, volume. 2174, Springer, Berlin, 2001, pp. 396-408.

Berners-Lee, T., Hendler, J., Lassila, O. *The semantic Web*. *Scientific American*, May, 2001. <http://www.sciam.com/article.cfm?articleID=00048144-10D2-1C70-84A9809EC588EF21>

Donini, F.M., Lenzerini, M., Nardi, D., Schaerf, A. *AL-log: Integrating datalog and description logics*. *Journal of Intelligent Information Systems*, 10(3), 1998, pp.227-252.

Eiter, T., Lukasiewicz, T., Schindlauer, R., Tompits, H. *Combining Answer Set Programming with Description Logics for the Semantic Web*. In Dubois, D., Welty, C., Williams, M-A. (Eds), Proceedings of Ninth International Conference on Principles of Knowledge Representation and Reasoning (KR 2004), June 2-5, Whistler, British Columbia, Canada, pp. 141–151. Morgan Kaufmann, 2004.

- Gardiner, T., Tsarkov, D., Horrocks, I. *Framework For An Automated Comparison of Description Logic Reasoners*. In proceedings of the 5th ISWC, *Lecture Notes in Computer Science*, volume 4273, Springer, Berlin, 2006, pp. 654-667.
- Golbreich, C. *Combining Rule and Ontology Reasoners for the Semantic Web*. *Lecture Notes in Computer Science*, volume 3323, Springer, Berlin, 2004, pp. 6-22.
- Grosz, B.N., Horrocks, I., Volz, R., Decker, S. *Description logic programs: Combining logic programs with description logic*. In Proc. of the Twelfth International World Wide Web Conference (WWW 2003), ACM, 2003, pp. 48-57.
- Haarslev, V., Möller, R. *RACER System Description*. In Proc. IJCAR 2001, Siena, Italy, June 18–23 2001, *Lecture Notes in Artificial Intelligence*, volume 2083, Springer, Berlin, 2001, pp. 701-706.
- Horrocks, I. *Application of Description Logics: State of the Art and Research Challenges*. *Lecture Notes in Artificial Intelligence*, volume 3596, Springer, Berlin, 2005, pp. 78-90.
- Horrocks, I., Parsia, B., Patel-Schneider, P., Hendler, J. *Semantic Web Architecture: Stack or Two Towers ?*. *Lecture Notes in Computer Science*, volume 3703, Springer, Berlin, 2005, pp. 37-41.
- Kalyanpur, A., Parsia, B., Sirin, E., Cuenca-Grau, B., Hendler, J. *SWOOP: a web ontology editing browser*. *Journal of Web Semantics*, 4(2), 2005.
- Kifer, M., Bruijn, D., Boley, H., Fensel, D. *A Realistic Architecture for the Semantic Web*. *Lecture Notes in Computer Science*, volume, 3791, Springer, Berlin, 2005, pp. 17-29.
- Knublauch, H., Fergerson, R.W., Noy, N.F., Musen, M.A. *The Protégé OWL Plugin: An Open Development Environment for Semantic Web Applications*. *Lecture Notes in Computer Science*, volume, 3298, Springer, Berlin, 2004, pp. 229-243.
- Lenzerini, M. *Data integration: a theoretical perspective*. In Proceedings of the twenty-first ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems. ACM 2002.
- Levy, A.Y., Rousset, M-C. *Combining Horn rules and description logics in CARIN*. *Artificial Intelligence*, 104(1–2):165–209, 1998.

Motik, B., Sattler, U., Studer, R. *Query Answering for OWL-DL with Rules*. In S.A McIlraith *et al.* (Eds) : ISWC 2004. *Lecture Notes in Computer Science*, volume 3298, Springer, Berlin, 2004, pp. 549-563.

Noy, N., McGuinness, D. *Protege Ontology Development 101: A Guide to Creating Your First Ontology*. Stanford University, Stanford, 2005.

<http://www.ksl.stanford.edu/people/dlm/papers/ontology-tutorial-noy-mcguinness.pdf>

Parsia, B., Sirin, E. *Pellet: An OWL-DL Reasoner*. Poster, In Proc. ISWC'2004, Hiroshima, Japan, November 7–11, 2004.

Rosati, R. *Semantic and Computational Advantages of the Safe Integration of Ontologies and Rules*. *Lecture Notes in Computer Science*, volume 3703, Springer, Berlin, 2005, pp. 50-64.

Sure, Y., Erdmann, M., Angele, J., Staab, S., Studer, R., Wenke, D. *OntoEdit: collaborative ontology engineering for the semantic web*. In: First International Semantic Web Conference (ISWC'02), *Lecture Notes in Computer Science*, volume 2342, Springer, Berlin, 2002, pp. 221–235.

Tsarkov D., Horrocks, I. *Ordering Heuristics for Description Logic Reasoning*. In Proc. IJCAI'2005, 30 Juillet - 5 Août 2005, Edinburgh, UK, Morgan Kaufmann Publishers. pp. 609-614.

Visser, U. *Intelligent Information Integration for the Semantic Web*. *Lecture Notes in Computer Science*, volume 3159, Springer, Berlin, 2004, pp. 13-34.

Wagner, G. *Web Rules Need Two Kinds of Negation*. *Lecture Notes in Computer Science*, volume 2901, Springer, Berlin, 2003, pp. 33–50.

Zou, Y., Finin, T., Chen, H. *F-OWL: An Inference Engine for the Semantic Web*. *Lecture Notes in Artificial Intelligence*, volume 3228, Springer, Berlin, 2005, pp. 238-248.

## Ouvrages

Abiteboul, S., Buneman, P., Suciú, D. *Data on the Web: From Relations to Semistructured Data and XML*. San Francisco, Etas-Unis Morgan Kaufmann Publishers, 21 Octobre 1999, 257 p.

Baader, F., McGuinness, D., Nardi, D., Patel-Schneider, P-F. *The Description Logic Handbook: Theory, Implementation and Applications*. Cambridge, New York: Cambridge University Press, 2004, c2003, 555 p.

Berler, M., Eastman, J., Jordan, D., Russell, C., Schadow, O., Stanienda, T., Velez, F. *Object Query Language*. In : Cattell, R.G.G., Barry, D.K. (éd.), *Object Data Standard: ODMG 3.0*. San Francisco, CA, USA: Morgan Kaufmann Publishers, 2000, 280 p.

Calvanese, D., McGuinness, D., Nardi, D., Patel-Schneider, P., *The Description Logic Handbook: Theory, Implementation and Applications*, UK, Cambridge Univ. Press, 2004.

Shelley, P. *Practical RDF*. Cambridge, Royaume-Uni : O'Reilly Média, Juillet 2003, 331 p.

Van Der Vlist, E. *XML Schema*. Paris : Edition O'Reilly, 30 Juin 2002, 399 p.

### **Travaux universitaires**

Volz, R. *Web Ontology Reasoning with Logic Databases, Chapter 3: The Semantic Web*.  
Thèse de Doctorat, Université Karlsruhe, Allemagne, Février 2004, 287 p.

tel-00581322, version 1 - 30 Mar 2011

## **Chapitre 3 : Interopérabilité entre systèmes d'information géographiques et de santé : une approche basée sur les métadonnées**

G. SALZANO, A. GUEMEIDA

In Proceedings of Géomatique'2006, Montréal, Canada, 25-26 octobre 2006



Cet article constitue une première contribution à la conception de NARI. Nous focalisons d'abord sur l'Analyse des scénarios de RI en santé (cfr. l'objectif OS1, Chapitre 1, §5.1.1), et nous considérons des aspects de :

- l'Information géographique, structurante pour la Recherche d'Information
- l'intégration de données, guidée par des exigences de qualité, comme forme d'interopérabilité entre systèmes hétérogènes

Ensuite, nous nous situons au niveau de l'« intégration de données » de NARI, (Chapitre 1, §5.1.2), et nous proposons une ontologie d'application et une approche d'intégration de données de type LAV, basée sur les métadonnées (cfr. l'objectif OS2, §5.1.1 du Chapitre 1). Au niveau médiateur, nous représentons les connaissances par des logiques de descriptions et les outils associés (cfr. l'objectif OS3, §5.1.1 du Chapitre 1). Nous illustrons NARI avec des requêtes définies sur un schéma global associé à un simple exemple du domaine de la santé.

## 1. Introduction

La nécessité de relier des systèmes d'information géographiques et de santé pour des finalités de gestion et de diffusion d'information n'est plus à prouver. De nombreux projets et réalisations sont lancés au niveau international par l'Organisation Mondiale de la Santé, ou au niveau national et régional dans plusieurs pays.

Les différents niveaux du couplage de ces systèmes peuvent être appréhendés par des démarches d'interopérabilité des systèmes d'informations. Celles-ci, incluant des aspects organisationnels et techniques, nécessitent aussi bien des multiples expertises sectorielles que des expertises transversales, pour établir des corrélations entre les domaines. L'approche proposée à pour objectif de prendre en compte ces expertises diverses.

Le plan de cet article est le suivant. On explicite des liens entre la démarche systémique, le développement durable et l'interopérabilité des systèmes d'information (§2). Dans des systèmes à large échelle, l'interopérabilité vise à : (a) communiquer et échanger des données distribuées entre une multitude d'applications hétérogènes et autonomes ; (b) utiliser les données échangées en comprenant leurs relations et contextes. Dans le §3, on présente trois difficultés majeures pour l'élaboration de systèmes coopératifs : l'hétérogénéité sémantique,

l'indétermination et la complexité du cadre normatif. Dans le § 4, on présente une approche d'interopérabilité basée sur les métadonnées, pour donner un accès transparent à des sources d'information issues des deux domaines sectoriels, santé et géographie. A partir d'une vue globale du domaine et de ses exigences, l'approche permet de (i) évaluer les sources sur des critères de qualité externe spécifiques au contexte et (ii) décomposer les requêtes globales en requêtes partielles, extraire les contributions à partir des sources locales et les fusionner.

L'infrastructure technologique de cette approche, présentée dans le §5, couple un système d'inférence et un système de médiation et utilise des langages et outils de spécification et interrogation du web sémantique : Racer et OWL au niveau transversal, OQL pour le médiateur, XML-Schéma et XMLQuery au niveau des sources locales. L'application de l'approche à un jeu de données simplifié, représentatif d'un exemple de gestion de risques, ainsi que les résultats obtenus, sont commentés (§6), avant de présenter des perspectives de recherche.

## 2. Démarche systémique, développement durable et interopérabilité

La santé et les territoires sont liés structurellement dans leurs dimensions géographiques, sociales et économiques. Ainsi, par exemple, les programmes de développement durable (comme Agenda 21 des Nations Unies) et la constitution de l'Organisation Mondiale de la Santé (OMS) partagent un grand nombre de principes : vision à long terme, promotion de la qualité plutôt que de la quantité, équité, partenariats, formation. Ces principes sont applicables à différents domaines (environnement, social, ...) et secteurs d'activité (alimentaire, technologique, énergétique, ...).

Une approche systémique des systèmes d'information permet d'intégrer ces principes dans une vision transversale. Selon cette approche, le système d'information (S.I.), en tant que "système", est un "ensemble organisé de composantes intercorrélées qui accomplissent ensemble des tâches pour atteindre un objectif commun" (Sommerville, 2004).

En incluant les utilisateurs et les développeurs, les règles propres au métier et les contraintes organisationnelles, par exemple légales, sociales, économiques, pour atteindre l'objectif commun de l'organisation qu'il doit supporter, un S.I. inclue la connaissance de comment il devrait être utilisé. Ainsi, on peut considérer un S.I. comme appartenant à la classe de

systèmes définis comme "sociotechniques", à quatre niveaux (processus métier, logiciels applicatifs, logiciels de support, matériel). En général, sous l'impulsion des avancées technologiques, scientifiques et sociétales, les changements apportés à un niveau du système se répercutent largement sur les niveaux adjacents. Cette dynamique rend la conception de ces systèmes très complexe.

Plusieurs S.I. coopératifs visant le développement durable et la santé utilisent des plateformes de SIG (Systèmes d'Information Géographiques) pour des fonctions d'acquisition, d'analyse (statistique et spatiale) et d'affichage de données de santé (épidémiologiques, démographiques et sociales). Des corrélations thématiques et géographiques (changements climatiques, mobilité des personnes et des agents pathogènes) facilitent la prise de décision. Parmi les S.I. de cette catégorie, on citera le Réseau Global d'Intelligence en Santé Publique développé par Health Canada, pour identifier au niveau international des informations pouvant montrer le déclenchement d'événements en relation avec la santé publique, ainsi que deux produits développés par l'OMS : le Système d'Information Statistique qui collecte les valeurs d'indicateurs de santé (population, mortalité, espérance de vie, ...) sur 192 pays, et l'Atlas Global des Maladies Infectieuses.

Dans les objectifs du développement durable s'inscrivent aussi les réseaux de santé. En regroupant plusieurs établissements et professionnels de santé, ceux-ci visent à recomposer l'offre de soins pour répondre à multiples changements : allongement de la durée de vie, émergence de multipathologies, inégalité des accès aux soins due à des facteurs géographiques, sociaux, économiques. Les réseaux de santé se développent avec différentes spécificités au Canada (Réseau Canadien de la Santé), et particulièrement au Québec, et dans plusieurs pays d'Europe, comme l'Angleterre (National Health Services), l'Espagne (Catalan Health Service CatSalut), la Finlande (Finnish Society of Telemedicine ) et la France (Nationale des Réseaux de santé, sanitaires et sociaux ). Pour améliorer la coordination, la continuité, l'interdisciplinarité des prises en charge sanitaires et maîtriser les dépenses, ces organisations doivent s'appuyer sur des S.I. coopératifs supportant les différents systèmes partenaires.

Des situations de gestion de risques naturels, comme par exemple lors de la "canicule" de 2003 en France, du tsunami de 2004 en Indonésie ou des inondations à la Nouvelle Orléans en 2005, ont tragiquement mis en évidence les attentes des gestionnaires et des populations pour des S.I. coopératifs, reliant la santé et la géographie. Nos recherches visent à apporter une

contribution à ces attentes, au niveau de l'analyse et de la conception de systèmes d'information. Dans le paragraphe suivant on analyse des facteurs de complexité pour l'élaboration de ces systèmes, avant de présenter une approche informatique d'interopérabilité pour des services d'interrogation. Cette démarche, s'appuie sur des recherches théoriques (Sheth, 1999), (CNRS AS 97, 2003) et peut être complémentaire à divers systèmes opérationnels pour la gestion des risques naturels (RF-PNC, 2005), (RF-SI-Eau, 2003).

### 3. Facteurs de complexité de l'interopérabilité de S.I. coopératifs à large échelle

L'interopérabilité est définie par l'IEEE (Institute of Electrical and Electronics Engineers, Inc.) comme la "capacité de deux ou plusieurs systèmes ou composants à échanger de l'information et à utiliser l'information qu'ils ont échangées".

L'analyse suivante, concernant la complexité de l'interopérabilité dans la gestion des risques naturels, peut être généralisée à d'autres domaines. Dans (Salzano, 2005), on analyse le partage de données en utilisant les dimensions de la distribution : *Pourquoi, Quoi, Qui, Où, Quand, Comment*. Schématiquement :

- *Pourquoi* représente la motivation du partage de ressources d'information. La motivation est liée à plusieurs activités : par exemple, évaluation de risques ou diffusion d'information.
- *Quoi* représente une source d'information, qui peut être liée à un ou plusieurs domaines d'activités. Par exemple, les sources d'information décrivant les établissements de santé relèvent en priorité de la santé mais aussi de l'économie.
- *Où* représente un territoire, dont l'interprétation est liée au domaine d'activité. Ainsi les territoires ont plusieurs spécialisations : par exemple, territoire de santé, administratif, géographique.
- *Qui* représente les acteurs. Ceux-ci 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 de données, incluant les moyens matériels et logiciels.

Cette analyse met en évidence trois difficultés pour l'interopérabilité : l'hétérogénéité sémantique, l'indétermination des sources de données et la complexité du cadre normatif.

- **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, différentes de celles, figées, des territoires administratifs. Elle concerne aussi le niveau de la granularité, spatiale et temporelle, des données qui doivent être prises en compte lors des opérations 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 les maisons de retraite ou les services de transport médicalisé.
- **L'indétermination** est liée à la multiplicité des relations de type n-m entre les classes des dimensions citées ci-dessus et à la multiplicité des valeurs pour certains attributs de ces classes. 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.
- La complexité du **cadre normatif** est liée à la richesse, à la fragmentation et au développement très récent du processus de standardisation. Il s'agit d'identifier les standards les plus adéquats pour le système cible et dans cette perspective analyser les standards adoptés dans les sources. Pour l'approche présentée dans le § 4, on a analysé des standards relevant en priorité de l'informatique et de domaines sectoriels, santé et géographie.

De très larges consortiums et groupements opèrent pour *l'interopérabilité informatique* : par exemple, le W3C, en ce qui concerne une architecture pour le Web sémantique et les Web services, et l'Object Management Group (OMG), en ce qui concerne les modèles et développements orientés-objet. Des métadonnées applicables à tous les domaines d'applications sont proposées par le Dublin Core Metadata Initiative (DC, 2006) et par la très récente norme 'Information Technology - Metadata Registries' (MDR, ISO/IEC 11179, <http://metadata-standards.org/11179/>) dont l'objectif est d'enrichir sémantiquement les

métadonnées. Dans le § 5 on indiquera les éléments de cette catégorie retenus dans l'approche proposée.

*Les normes sectorielles en santé et géographie* émanent d'organismes officiels, mandatés ou reconnus au niveau mondial, comme l'ISO (International Standardization Organization), ou au niveau européen, comme le CEN (European Committee for Normalization), avec des comités techniques organisés par discipline. Ils sont respectivement le TC 211 et le TC215, au sein de l'ISO, et le TC 287 et TC 251, au sein du CEN. De nombreux organismes internationaux ou nationaux contribuent à alimenter les processus de standardisation, sur des domaines d'étendue variable, comme par exemple les consortiums OpenGIS en cartographie. Dans chaque domaine, des normes ou standards très récents tendent à se diffuser, comme ISO 19115 (UK Gemini, 2003) en géographie et HL7 (HL7, 2005) en informatique médicale. Promu par l'ANSI et initialement orienté vers l'échange de messages médicaux, HL7 se généralise progressivement.

Des problèmes d'hétérogénéité et d'indétermination, analogues à ceux décrits au niveau de sources d'information, se posent au niveau des standards, dans chaque domaine sectoriel (Van Bommel et Musen, 1997 ; Devillers et Jeansoulin, 2005) et a fortiori pour les systèmes d'information coopératifs. Ils concernent par exemple la signification et la représentation de la localisation des ressources, les nombreuses valeurs optionnelles, souvent non renseignées, et l'absence de corrélations thématiques entre domaines.

Dans la suite, on présente une architecture d'interopérabilité, basée sur les métadonnées, visant à maîtriser ces trois facteurs de complexité : hétérogénéité sémantique, indétermination et adéquation aux normes et standards.

#### **4. Une approche d'interopérabilité basée sur les métadonnées**

L'approche d'interopérabilité proposée couple un système de médiation et un système d'inférence. Le premier vise à décomposer les requêtes globales en requêtes partielles, extraire les contributions à partir des sources locales et les fusionner ; le second, décrit dans le §5, évalue les sources sur des critères de qualité externe spécifiques au contexte.

La médiation est une architecture d'interopérabilité qui facilite l'accès transparent des utilisateurs à des ressources hétérogènes et réparties. Le catalogage, la localisation, la

navigation et l'interrogation sont des exemples de services de médiation. La médiation 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. On a bâti une architecture de médiation pour supporter les besoins potentiels d'un large éventail d'applications basées sur un ensemble commun et très vaste de sources, comme dans les applications d'e-gouvernement. Elle généralise les approches de médiation basées sur les schémas (Lenzerini, 2005), en s'appuyant fortement sur les métadonnées, pour réduire l'indétermination des interrogations de sources réparties et augmenter la cohérence et performance des interrogations, en imposant des contraintes au niveau applicatif, en fonction des contextes. Dans ce domaine, d'une part de nombreux verrous scientifiques et techniques sont encore à relever (CNRS AS 97, 2003), en ce qui concerne la "personnalisation" de l'information. D'autre part, le processus de standardisation dans le domaine des métadonnées est très actif, en particulier sous l'impulsion de directives européennes comme INSPIRE "Infrastructure for SPatial Information in Europe" (<http://www.ec-gis.org/inspire/>), visant à améliorer la disponibilité, la qualité, l'organisation et l'accessibilité des informations spatiales.

Dans le système de médiation, on considère trois niveaux : global, local et de médiation.

- Au niveau global, on explicite des classes et des exigences relatives au domaine applicatif. Celles-ci sont liées aux usages des sources pour différentes activités, par exemple surveillance systématique du territoire, gestion d'une situation de crise particulière, .... On exprime via des métadonnées des critères de qualité portant sur les sources en fonction du contexte : certains critères sont souvent utilisés en bases de données, comme par exemple la fraîcheur ou l'actualité (Berti-Equille, 2004), d'autres sont en relation avec la dimension géographique, comme la couverture spatiale des sources. Le diagramme UML du niveau global, représenté dans la figure 1, indique les principales classes et relations, à partir desquelles on définit des contraintes de gestion. Celles-ci, détaillées dans le §6, concernent par exemple l'identification de l'ensemble des sources nécessaires à la gestion d'un risque et des critères d'évaluation et de sélectivité sur ces sources.
- Au niveau local, des sources, on explicite les métadonnées induites du niveau global. L'analyse des valeurs des métadonnées sur les sources renseigne sur la pertinence et la qualité des sources par rapport aux objectifs et contextes des interrogations.
- Un médiateur relie les niveaux global et local. L'approche choisie, de type LAV (Local As View), permet de spécifier le domaine applicatif indépendamment des sources

disponibles, facilite le passage à des échelles très larges et l'intégration progressive de sources de données.

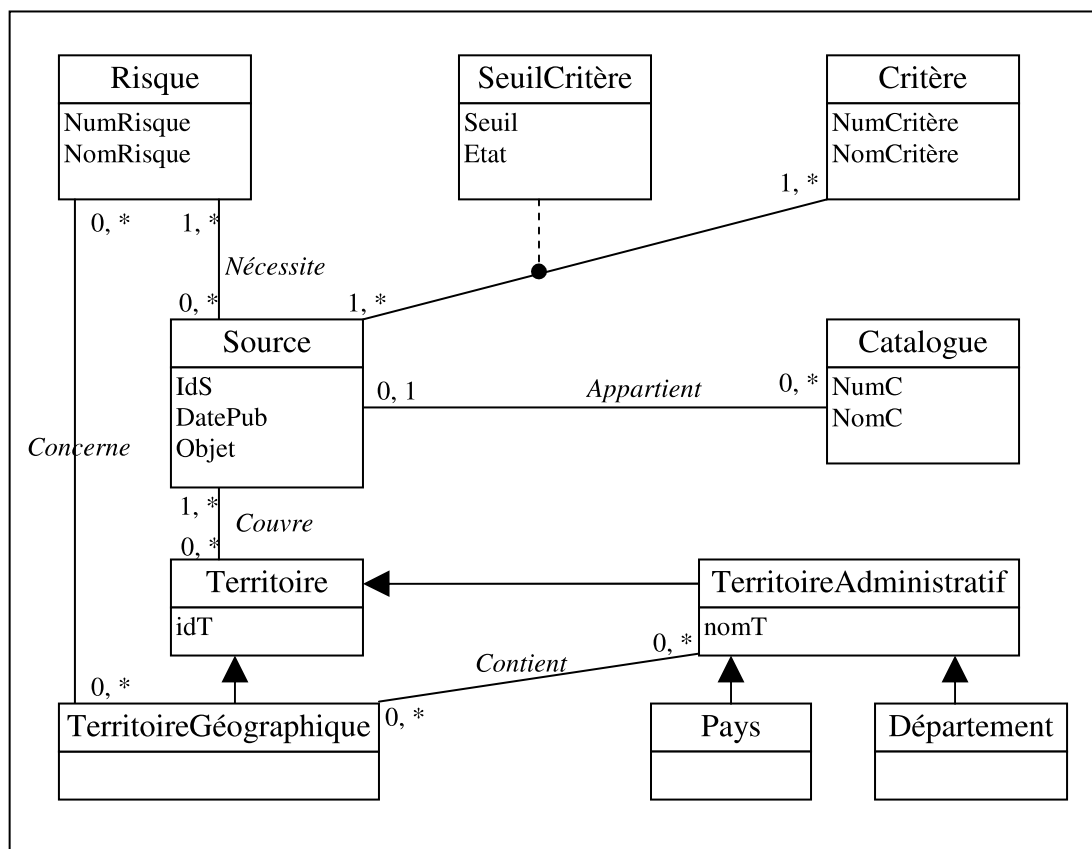


Figure 1 : Diagramme UML du domaine applicatif.

**Classification des interrogations.** Les interrogations visées sont de deux types et concernent respectivement la qualité des sources par rapport au contexte applicatif et le contenu des sources. On liste ci-dessous des exemples d'interrogations, par rapport à un risque et à deux activités de gestion : surveillance du territoire et gestion d'une situation de crise sur un territoire géographique. Ce deuxième cas nécessite des correspondances entre territoires géographiques et administratifs.

A) *interrogations sur la qualité :*

a1) *surveillance systématique :*

1. Le risque X est-il décrit sur tous les territoires administratifs ? Si non, quelles sont les sources de données manquantes ?
2. Le risque X est-il bien décrit ? Si non quelles sont les sources de données qui ne respectent pas les critères de qualité définis au niveau de l'application ?

a2) gestion d'une situation de crise pour un risque X dans un territoire géographique TG :

3. Le risque X est-il décrit sur le territoire géographique TG ? Si non, quelles sont les sources de données manquantes ?
4. Le risque X est-il bien décrit sur le territoire géographique TG ? Si non quelles sont les sources de données qui ne respectent pas les critères de qualité ?

B) interrogations sur le contenu des sources :

b1) surveillance systématique :

5. Pour chaque département à risque X, quel est le nombre d'établissements de santé ?
6. Pour chaque département à risque X, quel est le nombre de personnes à risque ?

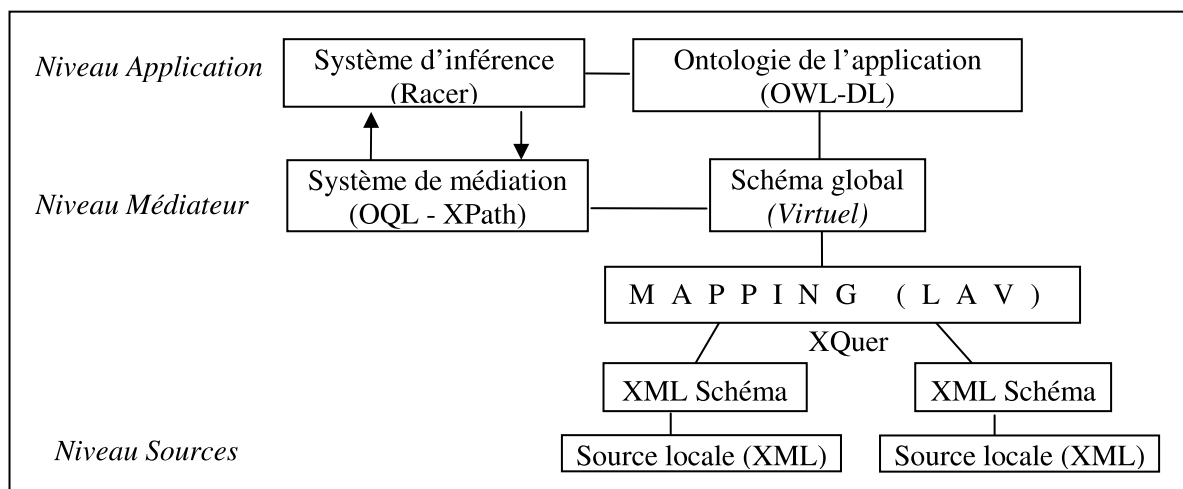
b2) gestion d'une situation de crise pour un risque X dans un territoire géographique TG :

7. Pour chaque département de TG quel est le nombre d'établissements de santé ?
8. Pour chaque département de TG quel est le nombre de personnes à risque ?

On présente ci-dessous l'infrastructure technique de l'architecture d'interopérabilité.

## 5. Infrastructure technologique

L'infrastructure utilisée se décompose en trois niveaux : Application, Médiateur et Sources, présentés dans (Salzano, Guemeida, 2006). Le niveau Application a été ultérieurement développé, par l'ajout d'un système d'inférence (figure 2), pour spécifier des concepts de gestion, à partir des classes du modèle au niveau global, et pour interroger ces concepts via des inférences.

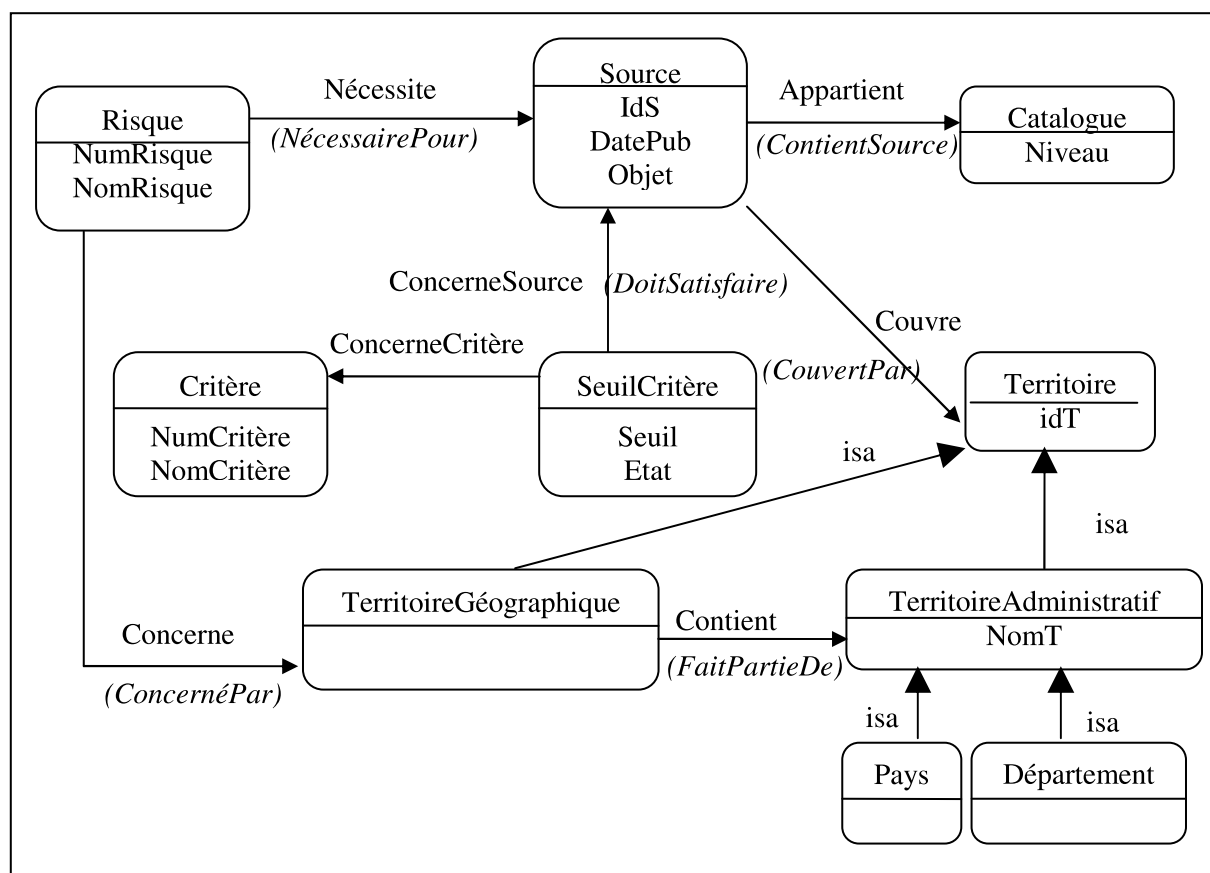


**Figure 2 :** Infrastructure technologique à trois niveaux

**5.1 Niveau Application.** On utilise une ontologie décrite à l'aide du langage du web sémantique OWL-DL. Par rapport à OWL-Light (OWL, 2004) et RDFS (RDFS, 2004), qui donnent une sémantique au modèle, OWL-DL est plus riche, grâce à la possibilité de l'utilisation maximale de la logique de description (DL) tout en maintenant sa décidabilité non garantie en OWL-Full. Un raisonneur est utilisé pour maintenir la cohérence et la consistance de l'ontologie et en déduire de nouvelles connaissances. Parmi les raisonneurs permettant de raisonner au même temps sur la TBox (Concepts) et la ABox (Instances), on a choisi (Racer, 2006). Pour éditer l'ontologie, on a utilisé Protégé (Noy, McGuinness, 2005) pour son interface avec Racer ou tout autre raisonneur utilisant l'interface standard DIG. La figure 3 donne une vue sur l'ontologie utilisée avec l'ensemble des concepts et des rôles employés. Les rôles inverses sont indiqués en italique entre parenthèses. Les généralisations entre les concepts sont représentées à l'aide des rôles « *Isa* ».

Les concepts illustrés dans la figure 3 sont des concepts primitifs. Les concepts définis à partir des concepts primitifs en utilisant la DL n'apparaissent pas sur le schéma. La définition de ces concepts en DL permet de préciser les conditions nécessaires et/ou nécessaires et suffisantes pour reclasser les instances.

Au niveau des instances de l'ontologie, des mises à jour sont réalisées régulièrement par l'extraction des métadonnées des sources.



**Figure 3** : Ontologie au niveau application

### Formulation des interrogations

#### A) interrogations sur la qualité :

Pour ces interrogations on définit un ensemble de concepts dans l'ontologie de l'application en utilisant le formalisme du langage OWL-DL. Chaque définition de concept est une formule basée sur : les concepts primitifs ou les concepts déjà définis, l'ensemble des propriétés et l'ensemble des symboles suivants { $\sqcap$  : intersection,  $\sqcup$  : union,  $\forall$  : toutes les valeurs dans,  $\exists$  : au moins une valeur dans,  $\neg$  : complément de,  $\ni$  : possède la valeur,  $=$  : cardinalité,  $\geq$  : cardinalité minimale,  $\leq$  : cardinalité maximale,  $\{ \}$  : énumération}. Les propriétés peuvent être des DatatypeProperty, tel que *NomRisque* avec comme type de données : *string*, ou des ObjectProperty comme *Appartient* reliant deux concepts.

a1) surveillance systématique :

*Interrogation 1 : Le risque X est-il décrit ? Si non, quelles sont les sources de données manquantes ?*

On considère qu'un risque est décrit si toutes les sources de données, nécessaires à la gestion de ce risque, sont disponibles. La disponibilité des sources permet d'obtenir des réponses complètes aux interrogations sur le contenu. La disponibilité d'une source est marquée par son appartenance à l'un des catalogues. Les sources non disponibles seront considérées comme des sources manquantes. Ainsi, le concept « SourceManquante » est le complément du concept « SourceDisponible » par rapport au concept « Source ». Les concepts « RisqueDécrit » et « RisqueNonDécrit » sont complémentaires par rapport au concept « Risque ». La complémentarité est introduite dans les formules par l'utilisation du symbole ( $\neg$ ) signifiant la négation.

- RisqueDécrit  $\equiv$  Risque  $\square \quad \forall$  Necessite SourceDisponible
- SourceDisponible  $\equiv$  Source  $\square \quad \exists$  Appartient Catalogue
- SourceManquante  $\equiv$  Source  $\square \quad \neg$  SourceDisponible
- RisqueNonDécrit  $\equiv$  Risque  $\square \quad \neg$  RisqueDécrit

Si, après interrogation, le risque X est considéré comme non décrit, cela veut dire qu'il manque une ou plusieurs sources, nécessaires à la gestion de X. Pour connaître les sources manquantes, on a défini le concept « SourceManquanteX » ainsi :

- SourceManquanteX  $\equiv$  SourceManquante  $\square \quad \exists$  NecessairePour (Risque  $\square$  NomRisque  $\ni$  "X")

*Interrogation 2 : Le risque X est-il bien décrit ? Si non, quelles sont les sources de données qui ne respectent pas les critères de qualité définis au niveau de l'application ?*

Un risque est bien décrit si toutes les sources, nécessaires pour sa gestion, vérifient les critères de qualité spécifiques au contexte de l'application. Une source disponible et vérifiant tous les critères de qualité est dite « SourceQualifiée ». Une source défaillante est une source non qualifiée. Une source est qualifiée, si les seuils définis pour les critères ne sont pas dépassés. Dans le cas du critère de fraîcheur, si d1 est la date de publication de la source, d2 est la date

actuelle et  $S$  est le seuil défini au niveau application pour cette source, alors le seuil est respecté si et seulement si :  $d2 - d1 < S$ .

- $SourceQualifiée \equiv SourceDisponible \sqcap \forall DoitSatisfaire SeuilCritereRespecté$
- $SourceDéfaillante \equiv SourceDisponible \sqcap \neg SourceQualifiée$
- $RisqueBienDécrit \equiv RisqueDécrit \sqcap \forall Necessite SourceQualifiée$
- $RisqueMalDécrit \equiv RisqueDécrit \sqcap \neg RisqueBienDécrit$

Si, à l'issue de l'interrogation, le risque  $X$  est parmi les risques mal décrits, alors il existe une ou plusieurs sources qui ne respectent pas les critères de qualité. Afin d'identifier ces sources défaillantes, on a défini le concept « Source Défaillante  $X$  » tel que :

- $SourceDéfaillanteX \equiv SourceDéfaillante \sqcap \exists NecessairePour (Risque \sqcap NomRisque \ni "X")$

*a2) gestion d'une situation de crise pour un risque  $X$  dans un territoire géographique  $TG$  :*

*Interrogation 3 : Le risque  $X$  est-il décrit sur le territoire géographique  $TG$  ? Si non, quelles sont les sources de données manquantes ?*

On suppose, comme indiqué dans l'ontologie de l'application, qu'un territoire géographique peut être composé d'un ou de plusieurs territoires administratifs. De ce fait, un risque est décrit sur un territoire géographique  $TG$  s'il est décrit sur l'ensemble des territoires administratifs composant  $TG$ . On considère qu'un risque est décrit sur un territoire administratif si les sources de données qui le couvrent sont disponibles (i.e. ne sont pas manquantes).

- $RisqueDécritTG \equiv Risque \sqcap \exists Concerne \{TG\} \sqcap \neg (\exists Necessite (SourceManquante \sqcap \exists Couvre (TerritoireAdministratif \sqcap \exists FaitPartieDe \{TG\})))$

Si le risque  $X$  n'est pas parmi les risques décrits sur  $TG$ , alors il manque une ou plusieurs sources en correspondance de territoires administratifs de  $TG$ . Pour identifier ces sources manquantes, on a défini les deux concepts suivants :

- $SourceTG \equiv Source \sqcap \exists Couvre (TerritoireAdministratif \sqcap (\exists FaitPartieDe \{TG\}))$
- $SourceManquanteX\_TG \equiv SourceTG \sqcap \neg SourceManquanteX$

*Interrogation 4 : Le risque X est-il bien décrit sur le territoire géographique TG ? Si non quelles sont les sources de données défaillantes ?*

Un risque est bien décrit sur un TG si toutes les sources de données qui couvrent les territoires administratifs de ce TG respectent les critères de qualité (i.e. ne sont pas défaillantes).

- $\text{RisqueBienDécritTG} \equiv \text{RisqueDécritTG} \wedge \neg (\exists \text{Necessite} (\text{SourceDéfaillante} \wedge (\exists \text{Couvre} (\text{TerritoireAdministratif} \wedge (\exists \text{FaitPartieDe} \{TG\}))))))$

Le concept « *SourceDéfaillanteX\_TG* » permet de connaître les sources défaillantes nécessaires à la gestion du risque X sur le territoire géographique TG.

- $\text{SourceDéfaillanteX\_TG} \equiv \text{SourceTG} \wedge \text{SourceDéfaillanteX}$

## 5.2 Niveau Sources

On considère des sources de données au format XML. Leurs structures sont décrites avec XML-Schéma. Le schéma de chaque source contient, en plus des éléments relatifs aux données, des attributs de métadonnées pour des besoins du niveau application, par exemple : identifiant de la source, objet, date de publication, couverture géographique. Dans des cas concrets, ces attributs peuvent faire référence à plusieurs espaces de noms, par exemple des éléments normatifs du Dublin Core ou de la norme ISO 19115.

## 5.3 Niveau Médiateur

Les correspondances (Mapping) entre le schéma global et les différents schémas des sources locales représentent les métadonnées de ce niveau. En suivant l'approche développée dans le prototype STyX (Fundulaki, Amann, & al, 2002), elles sont exprimées par des règles de transformation. Celles-ci associent aux chemins XPath de la source des chemins conceptuels dans un graphe de concepts associé à l'ontologie du domaine étudié. Ainsi, à gauche d'une règle, on trouve un chemin XPath dans la source et à droite le nom du concept du schéma global.

Les requêtes globales sont exprimées en fonction des termes du graphe de concepts à l'aide d'une variante simplifiée d'OQL. Pour être évaluées, elles sont décomposées 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 les règles de transformation. 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.

## 6. Illustration de l'approche

Pour illustrer l'approche on a considéré :

- Un ensemble de risques R1, R2 et R3 désignant respectivement les risques de Canicule, Grippe aviaire et Inondation. Chaque risque concerne un ou plusieurs territoires géographiques et sa gestion nécessite un ensemble de sources de données (Table 1). Pour le risque R1 (Canicule), les informations sur les personnes âgées dépendantes (PA) constituent la catégorie des personnes les plus vulnérables à ce risque. Pour le risque R2 (Grippe aviaire), les informations sur les élevages de volaille sont indispensables pour prendre les mesures empêchant la transmission du virus. Le risque R3 (Inondation) nécessite des informations sur les campings, qui constituent des infrastructures de tourisme vulnérables à ce risque.
- Un ensemble de territoires géographiques TG1, TG2 et TG3. Chaque territoire géographique est composé d'un ou de plusieurs territoires administratifs. (Table 2)
- Un ensemble de territoires administratifs : P1, D1, D2, D3 et D4. P1 est un pays et D1, D2, D3, D4 sont des départements qui appartiennent administrativement à P1.
- Un ensemble de catalogues C1 et C2. Un catalogue contient une ou plusieurs sources de données. (Table 3)
- Un ensemble de sources de données réparties S1,...S8. Chaque source sera complétée par des éléments de métadonnées apportant des informations complémentaires sur la source comme : l'objet, la date de publication et la couverture. Chaque source de données couvre un ou plusieurs territoires administratifs. On suppose que si une source de données couvre un pays alors elle couvre tous les départements de ce pays (Table 4). Ces sources sont issues des domaines de la santé et de la géographie.
- Un ensemble de critères de qualité que les sources de données doivent vérifier. Pour chaque critère et chaque source on indique le seuil maximum autorisé. La table 5 montre les seuils définis pour le critère de fraîcheur.

Risque	Nom risque	Sources nécessaires	Territoires géographiques concernés
R1	Canicule	S1, S2, S3, S4, S5	TG1
R2	Grippe aviaire	S1, S2, S5, S6, S7	TG2
R3	Inondation	S1, S2, S5, S8	TG1, TG2

**Table 1 :** Correspondances entre risques et sources de données

Territoire géographique	Composition
TG1	D1, D2
TG2	D3, D4

Catalogue	Niveau	Contenu
C1	National	S1, S2, S5
C2	Départemental	S3, S4, S6, S7, S8

**Table 2 :** Territoires géographiques et départements

**Table 3 :** Catalogues et sources

Source	Objet	Date de publication	Couverture
S1	Carte de risque	2005-10-15	P1
S2	Etablissements de santé	2005-12-16	P1
S3	PA (Personnes Agées dépendantes)	2006-07-02	D1
S4	PA	2006-01-02	D2
S5	Liste des risques	2005-06-12	P1
S6	Elevages de volailles	2003-01-12	D3
S7	Elevages de volailles	2006-02-24	D4
S8	Campings	2006-09-08	P1

**Table 4 :** Sources de données

Critère	Source	Seuil
Cr1 : Fraîcheur	S1	2 ans
	S2	2 ans
	S3	6 mois
	S4	6 mois
	S5	3 ans
	S6	1 an
	S7	1 an
	S8	1 mois

**Table 5 :** Critères et seuils

### *Schéma et contenu des sources de données*

Les sources de données sont au format XML, avec les éléments suivants :

- S1 : départements soumis à des risques, avec numéro de risque, numéro et nom de département.
- S2 : établissements de santé, avec numéro, nom et type d'établissement, numéro de département.

- S3, S4 : personnes âgées dépendantes (PA), avec numéro et nom, respectivement sur les deux départements D1 et D2.
- S5 : risques, avec numéro et nom du risque.
- S6, S7 : élevages de volailles, avec numéro, type et adresse de l'élevage, respectivement sur les deux départements D3 et D4.
- S8 : campings, avec numéro, capacité, adresse.

On suppose que les sources S1, S2 et S5 sont établies au niveau national, tandis que les sources S3, S4, S6, S7 et S8 sont établies au niveau départemental. La source S5 se réfère aux risques : R1 - Canicule, R2 -Grippe aviaire, R3 -Inondation. Pour les sources S1 à S4, on a considéré le jeu de données suivant :

S1		
NumR	NumD	NomD
1	1	D1
1	2	D2
2	3	D3
2	4	D4
3	1	D1
3	2	D2
3	3	D3
3	4	D4

S2			
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

S3	
NumP	NomP
1	PA1
2	PA2
3	PA3

S4	
NumP	NomP
4	PA4
5	PA5

### *Résultats des interrogations sur la qualité*

Les interrogations sont lancées par le système d'inférence Racer à partir de Protégé<sup>66</sup>. Les résultats obtenus sont :

#### *Interrogation 1 :*

- SourceDisponible = {S1, S2, S3, S4, S5, S7, S8}, car chaque source  $S_i$  appartient à l'un des catalogues.
- SourceManquante = {S6} : S6 est manquante car elle n'appartient à aucun catalogue.
- RisqueDécrit = {R1, R3}, car toutes les sources nécessaires à la gestion des risques R1 et R3 sont disponibles.

<sup>66</sup> Par défaut, Protégé s'interface avec Racer sur le port 8080. Les résultats rendus par Racer sont visibles dans Protégé en cliquant sur l'onglet « Inferred ». L'onglet « Asserted » contient les déclarations des concepts.

- RisqueNonDécrit = {R2}, car pour le risque R2 on a besoin des sources S1, S2, S5, S6 et S7 mais la source S6 est manquante.
- SourceManquanteCanicule = { $\emptyset$ }, car le risque X est décrit, donc aucune source n'est manquante (i.e. toutes les sources nécessaires à sa gestion sont disponibles).

*Interrogation 2 :*

- SourceQualifiée = {S1, S2, S3, S5, S7, S8}, toutes ces sources respectent les critères de qualité définis. La vérification est faite en utilisant la date du « 2006-09-08 ».
- SourceDéfaillante {S4}, S4 est défaillante car le seuil du critère fraîcheur n'est pas respecté : la source S4 est publiée le « 2006-01-02 », le seuil de fraîcheur est de 6 mois et la date courante est « 2006-09-08 ». Donc la durée écoulée entre les deux dates, 8 mois et 6 jours, dépasse le seuil maximum de fraîcheur autorisé (6 mois).
- RisqueBienDécrit = {R3}, car toutes les sources nécessaires à la gestion de R3 sont des sources qualifiées.
- RisqueMalDécrit = {R1}, car S4 est défaillante.
- SourceDéfaillanteCanicule = {S4}, car le seuil de fraîcheur n'est pas respecté

*Interrogation 3 :*

- RisqueDécrit\_TG1 = {R1}, car R1 est décrit sur tous les départements de TG1 et donc aucune source nécessaire au risque Canicule n'est manquante.

*Interrogation 4 :*

- RisqueBienDécrit\_TG1 = {R3}

Comme le risque Canicule (R1) n'est pas parmi les risques décrits sur TG1, alors il y a des sources défaillantes : SourceDéfaillanteCanicule\_TG1 = {S4}, car le seuil de fraîcheur n'est pas respecté.

## 7. Conclusions

L'objectif de ce papier a été d'illustrer comment les systèmes d'information coopératifs, notamment dans la gestion de risques naturels ayant un impact sur la santé, requièrent des expertises multiples, en systèmes d'informations géographiques, de santé, en gestion, ainsi

que la construction de "ponts" entre ces systèmes. Dans cet objectif, on a proposé une approche d'interopérabilité de systèmes d'information basée sur les métadonnées, avec une infrastructure couplant un système d'inférence et un système de médiation. On a illustrée cette approche sur des interrogations correspondantes à deux activités de gestion différentes.

En perspective, on envisage de compléter le système d'inférence, notamment par l'introduction de la dimension temporelle. Cette extension permet de détecter un risque sur un territoire géographique, à partir de données réparties dans l'espace et dans le temps, et aussi de situer dans le temps les processus d'évaluation de la gestion de l'information. On envisage aussi d'illustrer le système ainsi étendu sur des données réalistes issues du Plan Canicule français.

## 8. Bibliographie

- Berti-Equille, L. (2004) "Un état de l'art sur la qualité des données". In *Qualité des systèmes d'Information - Ingénierie des systèmes d'information - RSTI - Série ISI*, 9/2004, pages 117-143
- CNRS AS 97 du département STIC du CNRS (2003). "Médiation via les métadonnées". <http://www.lirmm.fr/~libourel/MM/MetaMedia.htm>
- DC (2006). Dublin Core Metadata Initiative, <http://dublincore.org/>
- Devillers, R.; Jeansoulin, R. (Eds.) (2005). "Qualité de l'information géographique". Hermès Science
- Fundulaki, I.; Amann, B.; Beerli, C.; Scholl, M. (2002). "STYX : Connecting the XML World to the World of Semantics (Demo)". In *Proceedings EDBT'2002*, 2002
- Lenzerini, M. (2005) "Data Integration: A theoretical Perspective. Course: logic-based information integration", ESSLLI, 2005
- Noy, N.; McGuinness, D.; (2005). "Protégé Ontology Development 101: A Guide to Creating Your First Ontology", Stanford University, Stanford, <http://www.ksl.stanford.edu/people/dlm/papers/ontology-tutorial-noy-mcguinness.pdf>
- OWL (2004). Web Ontology Language Guide. W3C Recommendation 10 February 2004, <http://www.w3.org/TR/owl-guide/>
- Racer (2006). <http://www.racer-systems.com/>

RDF (2004). RDF Vocabulary Description Language 1.0: RDF Schema. W3C Recommendation 10 February 2004. <http://www.w3.org/TR/rdf-schema/>

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

RF-SI-Eau (2005). Ministère de l'Ecologie et du Développement Durable, France (2005) "Architecture du Système d'Information sur l'Eau. Livre vert".

Salzano, G., Guemeida A. (2006). "Métadonnées pour l'interrogation de sources de données hétérogènes et distribuées - Illustration sur un scénario couplant la santé et l'environnement". In Actes du Séminaire VSST'2006, Lille, 16-17 janvier 2006, 12p et CD

Salzano, G. (2005). "Modeling Metadata for Multidomain Health and Geography Information". In Kremers, H. (Ed.) Proceedings of International Symposium on Generalization of Information ISGI 2005, Berlin, Germany, 14-17 September 2005, 12 p.

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

Sommerville, I. (2004). "Software Engineering". 7th Edition. Addison Wesley

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

HL7 (2006). Health Level 7: <http://www.hl7.org/>

Van Bommel, J.H.; Musen, M.A. (Eds.) (1997). "Handbook of Medical Informatics". Springer

**Chapitre 4 : Qualité de l'information  
géographique :  
approche basée sur les besoins préliminaires**

A. GUEMEIDA, R. JEANSOULIN, G. SALZANO

Colloque International de Géomatique et d'Analyse Spatiale, SAGEO'2007,  
Clermont-Ferrand, 18, 19 Juin 2007



Dans cet article nous introduisons la métaphore de l'impédance, qui contribue à l'originalité de NARI.

Nous focalisons ensuite sur le niveau « identification / préparation des sources de données » du processus de recherche d'information (cfr. Chapitre 1, § 5.1.2.2).

Nous utilisons la métaphore de l'impédance pour conduire la conception de NARI, selon les objectifs spécifiques (OS) détaillés dans Chapitre 1, §5.1.1.

Nous développons en particulier l'analyse des hétérogénéités ("*résistances*") relevant de l'Information Géographique et plus généralement des Systèmes d'Information. D'autre part nous développons l'analyse des exigences de qualité externe dans des systèmes géographiques et de santé (cfr. l'objectif OS1).

Nous affinons aussi l'ontologie d'application et nous indiquons les étapes destinées à réduire progressivement l'écart d'impédance, avec les aspects à traiter (cfr. l'objectif OS2).

La conception de ces étapes mérite d'être approfondie dans des recherches futures, notamment à l'aide d'estimations quantitatives de l'écart d'impédance, en liaison aux problèmes d'hétérogénéité, de qualité et de volumes des données.

Le chapitre suivant aborde la conception de modules logiciels qui participent à ces étapes.

## 1. Introduction

Les informations accessibles sont aujourd'hui si nombreuses et diverses, que les processus d'interrogation ou de fouille de données, sont de plus en plus indépendants des observations et de leur enregistrement initial. Les utilisateurs ont du mal à savoir qui a produit quoi, à quelle date, dans quel but, et il devient impossible de tracer une histoire détaillée des données. En conséquence, l'utilisateur a besoin d'une aide automatique pour choisir des données, parmi des quantités sinon ingérables. Mais les critères sont très difficiles à maîtriser, car cette aide complexe consiste souvent en "robots" logiciels, en interfaces d'accès, à des catalogues ou métadonnées, généralement peu transparents.

L'information géographique n'échappe pas à ce mouvement et il faut faire toujours plus confiance à des processus intermédiaires, qui ajoutent de l'incertitude à l'incertitude liée à la qualité des données. Ainsi, la qualité globale du service doit devenir un compromis entre la

qualité annoncée par le producteur, selon ce qui est accessible par le service, et la qualité qui est acceptable pour l'application.

Ce compromis est un problème majeur de l'information géographique depuis des années. Il est à l'origine du divorce ('mismatch') entre ce qui est obtenu et ce qui était attendu. Nous abordons ce problème en utilisant la métaphore du divorce d'impédance ("impedance mismatch"), qui est un phénomène classique lorsqu'on essaye de brancher ensemble deux systèmes physiques. Cette métaphore peut aider à structurer et classer les divers modules de traduction à introduire dans le flot de données (section 2). Cette structure est découpée en trois niveaux (s.3), et une architecture de médiation est proposée pour sa mise en œuvre (s.4). Nous illustrons cette démarche par un exemple simple, pris dans le domaine de la santé (s.5).

## 2. Interopérabilité de systèmes d'information et impédance

### 2.1 La métaphore de l'Impédance Mismatch

Par analogie avec les systèmes électriques, mécaniques ou hydrauliques, nous parlerons de "impedance mismatch" entre producteur et utilisateur de données. Sans détailler, rappelons que ce phénomène a besoin d'un aller-retour entre la source et la destination, comme en courant alternatif (AC), sinon, en courant continu (DC), il s'agit d'une simple résistance. Tentons une analogie avec l'information.

**Intensité** :  $I$ . Peut être reliée au volume du flot de données ;

**Voltage** :  $V$ . Différence de potentiel, ou degré de spécialisation des données ; la **puissance**  $P = V \times I$ , est le produit de ces deux grandeurs.

**Impédance** : nombre complexe  $Z = R + iX$ , entre résistance  $R$  et réactance  $X$ .

**Résistance** :  $R$ . Perte due à la traduction entre le référent et sa représentation.

**Réactance** : si  $X > 0$  : *réactance inductive*, interprétée comme une inertie à conserver des représentations antérieures, léguées. Elle crée une énergie externe, sous forme d'annotations, alertes, etc... à usage externe. Si  $X < 0$  : *réactance capacitive*, qui crée une énergie interne, utilisable sous forme de « cache » d'information, utilisable par un filtre (ex.: résumé statistique, coordonnées lissées), pour le transfert d'une information intensionnelle, ordinale, plutôt qu'extensionnelle.

La différence entre les impédances  $Z_{in}$  et  $Z_{out}$ , de deux systèmes connectés, provoque un problème de traduction et doit recevoir un traitement adéquat :

**Égalisation d'impédance** : si on veut maximiser le produit  $I \times V$  (cf. maximal power theorem), il faut ajouter les modules, capacités ou inducteurs, qui permettent d'égaliser les résistances  $R_{out} = R_{in}$ , et d'opposer les réactances  $X_{out} = -X_{in}$ .

**Pontage d'impédance** : si on souhaite un contrôle fin mais partiel, par exemple beaucoup de détails sur un volume restreint d'information, on prendra  $Z_{in} \gg Z_{out}$ .

**Transformateurs** : peuvent modifier  $V$ , seulement en AC

Ces traitements peuvent être flexibles dans les limites d'un usage possible de l'information. Nous allons montrer dans ce papier, comment anticiper de telles limites le plus tôt possible, en fonction de l'expression des besoins, afin de déterminer quel type d'accord d'impédance doit être mis en œuvre. Ceci entraîne un surcoût, mais qui est justifié par le possible gain d'éviter l'accès à de nombreuses données qui pourraient s'avérer finalement inutilisables.

## 2.2 Impedance Mismatch entre systèmes d'information

Les systèmes d'information (SI) font partie d'une organisation et se composent de plusieurs niveaux : logiciels d'application, de support, matériel. L'introduction d'une impédance est inévitable, entre ces niveaux et selon les tâches : la traduction des besoins préliminaires et globaux de l'organisation est reconnue comme un point critique pour le succès ou l'échec d'un système. Les aspects administratifs, culturels, légaux de l'interopérabilité, s'ajoutent aux aspects purement techniques, pour causer des résistances et des frictions.

On qualifie de 'impedance mismatch' (Castro *et al.*, 2002) l'écart entre le système 'nominal' et son environnement opérationnel : le système nominal est déterminé en fonction d'un objectif, dont les concepts diffèrent de ceux de l'environnement opérationnel. Ils proposent donc une approche basée sur l'objectif, démarrant dès la première expression des besoins par les divers acteurs.

## 2.3 Impedance Mismatch entre systèmes d'information géographique

En information géographique, plusieurs problèmes classiques peuvent provoquer des écarts d'impédance. Nous les examinons ici, avec une illustration issue du domaine de la santé en

France, qui met en évidence la complexité résultant de l'usage de nombreux systèmes, comme : le PMSI<sup>67</sup> (Plan de Médicalisation des Systèmes d'Information), les recensements de décès de l'INSEE, ou des registres communaux. Les objectifs sont très différents. Les données du PMSI ont un impact direct sur le financement des hôpitaux, mais mesurent mal l'efficacité des soins selon les maladies, alors que certaines maladies font l'objet d'un suivi particulier par ailleurs, mais pas de manière homogène sur le territoire. Ces systèmes doivent coopérer en fonction des spécialités des divers acteurs.

**Vecteur-Rasteur** (Champ-Objet) : les objets ne sont pas assimilables à des sous-ensembles de champs et des groupes de pixels ne forment pas toujours des objets. Ce type de désaccord rappelle ce qui a été identifié sous le terme de "object-relational impedance mismatch" (Ambler, 2001) entre bases de données relationnelles et orienté-objets. L'accès, à des groupes de pixels, est de type ensembliste et ne ressemble pas à une méthode par objet individuel. Pour rapprocher les points de vue, il faut créer une information annexe et la mémoriser dans l'attente éventuelle de s'en servir à nouveau. Sinon le surcoût est prohibitif. Il faut donc préparer plusieurs agents logiciels pour effectuer des transformations vecteur-rasteur, délimiter des contours, constituer une topologie, ou discrétiser les pixels à une taille donnée, etc. On peut utiliser l'analogie d'un inducteur, qui régularise un flot de données et crée une énergie annexe.

**Géométrie-Topologie** : il est certes possible de dériver la topologie à partir d'une géométrie correcte, mais cela nécessite un niveau de détail rarement atteint en réalité, à moins de surcharger inutilement la base de données, en accumulant les risques d'erreur et donc les risques d'incohérence. Ce cas survient avec les données géométriques issues d'une transformation rasteur-vecteur (cf. supra), mais aussi lorsque le recueil de la géométrie n'est pas systématiquement couplé à la topologie, par exemple après des calculs d'intersection entre jeux de données différents, etc. Cet aller-retour entre géométrie et topologie nécessite une 'capacité' qui retienne la part d'information utile.

**Echelle spatiale** : un changement apparemment continu peut ne pas se traduire par un facteur linéaire (zoom homothétique). En théorie du signal, le "rapport de Nyquist" nous enseigne qu'un signal utile doit avoir une bande passante, double de celle du canal de transmission. C'est la résistance qui fournit une analogie : il faut adapter nos besoins à un niveau

---

<sup>67</sup> PMSI : <http://www.atih.sante.fr/en/index.php?id=0002300005FF>

compatible avec les sources de données. Mais lorsque plusieurs sources mêlent plusieurs échelles, on peut tenter un traitement d'agrégation-désagrégation pour harmoniser sur une échelle unique, alors c'est plutôt une série de capacités et d'inducteurs, qui peut servir d'analogie.

**Echelle temporelle** : similaire à l'espace mais avec un impact humain direct en termes d'évolution, vieillissement, accumulation (ex : toxicité) ... avec des effets de seuil irréversibles, donc non symétriques, des effets de mémoire (ex : antécédents) qui peuvent être assimilés à une réactance inductive.

**Échelle de spécialisation** : le nombre de niveaux de détail ne s'accroît pas forcément avec le nombre de descripteurs, ou la taille du vocabulaire. Des écarts peuvent résulter de l'usage plus ou moins large d'un même mot dans des contextes différents, ou des mots différents décrivant des choses indistinctes dans un certain contexte. Par exemple, les nomenclatures médicales International Statistical Classification of Diseases & Health Problems (ICD) et Diagnostic & Statistical Manual of Mental Disorders (DSM), sont très différentes de la Chinese Classification of Mental Disorders (CCMD)<sup>68</sup>.

**Granularité** : cet aspect est complémentaire des problèmes d'échelle spatiale, temporelle ou de spécialisation, il concerne la taille des populations sur lesquelles est observée une certaine variable. Par exemple, des données médicales sont créées pour chaque patient, pour chaque médecin, pour chaque service hospitalier, pour une sous-population donnée (ex: grande entreprise), pour un groupe de maladies données, etc... sans considération des intersections possibles. Pourtant, lors de l'établissement de protocoles de soin collectif ou individuel, il est nécessaire de croiser ces données avec d'autres, relevant d'autres granularités (ex: travail, logement, éducation).

**Fitness for use** : ce terme anglais est très répandu pour nommer ce qui est appelé la 'qualité externe' définie par l'utilisateur. Elle s'oppose à la 'qualité interne' que le producteur est censé fournir avec les jeux de données qu'il diffuse. Dans un rapport signal-sur-bruit, si le bruit au sens de l'utilisateur, diffère du bruit au sens du producteur, le rapport perd beaucoup de sa signification. Il est illusoire de ramener le fitness-for-use à un simple rapport ; la notion d'accord d'impédance est plus riche et l'on peut remarquer que, comme dans l'analogie

---

<sup>68</sup> SFMG : Société Française de Médecine Générale :  
[http://www.sfm.org/Publication/documents\\_recherche/num\\_49.html](http://www.sfm.org/Publication/documents_recherche/num_49.html)

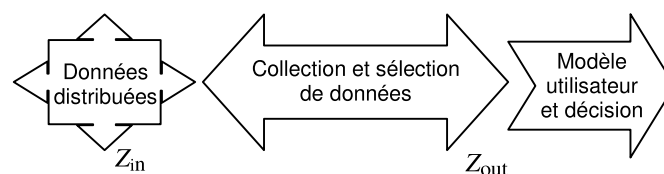
physique, il est plus difficile de trouver l'accord lorsque la puissance du signal est proche de celle du bruit (au sens utilisateur). Par exemple, pour constituer le 'dossier médical' d'un patient, la connaissance des épisodes antérieurs, doit être accessible et complète. Dans les tentatives de mise en place d'un tel système, la difficulté reconnue est de préserver une variété de vues qui permette le partage d'information durant les soins, sans préjudice des décisions.

### 3. Une approche en trois étapes

#### 3.1 Existence, Qualité et Contenu : trois aspects de l'information

La collecte et la sélection des données constituent une partie lourde et coûteuse du travail, suivie par la mise en oeuvre du modèle de l'application et qui se termine par la prise de décision. Dans les petites applications, il est raisonnable de grouper tout cela en un seul processus, pour en contrôler directement l'impédance. Dans les grandes applications, nous pensons qu'il est indispensable de considérer la sélection des données comme une tâche à part. Sa conception va nécessairement produire de l'impédance : le problème est d'anticiper avec quelles autres impédances celle-ci risque devoir être adaptée, afin de faciliter les 'médiations' ultérieures.

La question est alors : combien de modèles utilisateurs pourra-t-on gérer avec un système de sélection ainsi conçu ? (Plus d'un ?).



**Figure 1** : Les principaux lieux de confrontation d'impédance

Les organismes de Santé Publique collectent des quantités énormes de données, sur des systèmes hétérogènes, distribués à des niveaux différents (international, national, régional, départemental, communal) et avec des besoins différents (épidémiologie, contrôle des dépenses de santé, gestion hospitalière ...). Pour tout objectif particulier, comme l'étude d'une maladie, il faut s'interroger sur :

- **L'existence** de données pertinentes pour l'étude ;
- **La qualité**, de ces données, en adéquation (*fitness*) avec le but fixé ;

- **Le contenu**, qui doit assurer un usage cohérent et opérationnel.

### 3.2 Catalogues, Métadonnées, et Données : trois étapes d'interrogation

Les questions sur la pertinence, l'adéquation et la cohérence, font partie de l'accord à bâtir entre l'impédance  $Z_{in}$  des différentes sources de données et celle  $Z_{out}$  en entrée du ou des modèles utilisateurs. Doit-on tout connaître sur les données potentielles avant de bâtir cet accord ? Nous pensons avec d'autres auteurs (Barde *et al.*, 2005), qu'il est possible de construire graduellement cet accord d'impédance.

D'abord, nous accédons aux **Catalogues**, qui identifient les sources de données et fournissent quelques informations sur la couverture, le format, parfois la date comme dans l'exemple MDWEB (Desconnets *et al.*, 2003). 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 données, le découpage en niveaux de spécialisation, etc. Enfin nous pouvons accéder aux **Données**. Ceci nous suggère d'utiliser trois niveaux successifs dans la conception graduelle des composants de  $Z_{out}$ .

### 3.3 Construire une impédance en trois étapes

#### 3.3.1 Etape 1 : Existence

(a) **Espace**. Il semble aisé de trouver des données en fonction de coordonnées géographiques : le changement éventuel de référentiel cartographique s'apparente à une égalisation de résistance. La géolocalisation d'une adresse peut nécessiter un inducteur (interpolateur) et une capacité (raffinement ultérieur). L'interrogation des *getCapabilities* OGC du Catalogue peut inclure cette impédance.

(b) **Temps**. L'égalisation d'intervalle temporel est aisée, mais la comparaison de périodes référencées diversement peut être plus complexe et peut nécessiter des filtres (inducteurs) et des mémoires pour une révision éventuelle (capacités).

(c) **Thème**. L'interrogation à partir de mots-clés est souvent très approximative. Par exemple, pour le mot 'altitude', il faut savoir que SRTM est un sigle pertinent (*Shuttle Radar Topographic Mission*). La constitution de thesaurus spécialisés peut être une aide, mais souvent insuffisante. Or, la sélection thématique doit commencer dès le Catalogue et même dès le Catalogue des Catalogues (contexte WMS). L'association de termes relève de

l'égalisation de résistance, mais le choix même des termes nécessite filtrage (inducteur) et mémorisation (capacité) pour affiner ultérieurement ou remettre en cause les similarités. En particulier, la capacité de bâtir une hiérarchie de termes liés est très importante ; elle peut être guidée par une série de géolocalisation directe et inverse qui permet d'associer des thèmes en fonction de leur proximité locale. Le réglage d'impédance de telles opérations est difficile, entre une sélection trop permissive, beaucoup trop large, ou trop sévère. Mais il est nécessaire de faire des choix pour rendre le problème traitable.

### 3.3.2 Etape 2 : Qualité

L'adéquation de la qualité est un choix multi-dimensionnel, qui ne se limite pas aux seuls éléments de qualité interne : une aide logicielle efficace est nécessaire, ni trop sévère, afin de ne pas éliminer des données très significatives, ni trop laxiste afin de ne pas poursuivre la sélection sur un volume trop grande de données inutilisables. Les éléments de qualité normalisés (ISO19115), bien qu'insuffisants, sont utiles dans la description de l'impédance d'entrée  $Z_{in}$ .

**Précision de position.** La précision absolue mesure l'écart d'un échantillon de coordonnées avec des valeurs de référence considérées exactes par le producteur. Si la précision est insuffisante, il faut adapter la résistance dans  $Z_{in}$ , si elle est trop grande, il faut réduire le flot par induction. La précision relative est importante dans d'autres requêtes, avec d'autres contraintes (ex. : de topologie), il faut la mémoriser dans une capacité.

**Précision d'attribut.** Opération similaire : inclure une résistance dans l'analyse des métadonnées et un inducteur pour réduire la taille des données avant l'étape 3.

**Complétude.** Il s'agit du degré auquel les diverses caractéristiques sont représentées ou omises. La solution est d'établir une combinaison de résistances (si sous-représentation), d'inducteurs (sur-représentation), ou de capacités (si une inférence positive ou négative doit être ultérieurement dérivée).

**Précision, complétude, et validité temporelle.** Opérations similaires.

**Cohérence logique.** Pour anticiper les calculs de contraintes qui ne pourront être effectués complètement qu'avec les données, il est utile de réduire la cardinalité de certains ensembles dès l'étape 2. Ceci est un analogue de capacité.

### 3.3.3 Etape 3 : Contenu

Lorsqu'une liste réduite a été sélectionnée, on peut enfin confronter ses 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. Il est obligatoire de réduire a priori la taille d'exploration en utilisant un ordre de préférence approprié, comme un ordre partiel sur les croyances entre les solutions.

La construction de ces ordres partiels se fait par apprentissage statistique (ex : bayésien) ou qualitatif (ex : analyse formelle des concepts) ou toute combinaison. Le résultat nécessite une information d'accompagnement, mémorisée par des capacités, pour une aide à la reformulation des requêtes ultérieures, moins restrictives, mais avec le même objectif. On voit bien ici l'analogie avec l'énergie potentielle des capacités électriques.

Finalement, le processus peut boucler jusqu'à une décision acceptée pour un niveau de risque d'erreur remis à jour.

## 4. Une vue intégrée en trois étapes, guidée par les besoins

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 aisément à la structure INSPIRE et aux normes ISO 19113-115 etc... (Figure 2).

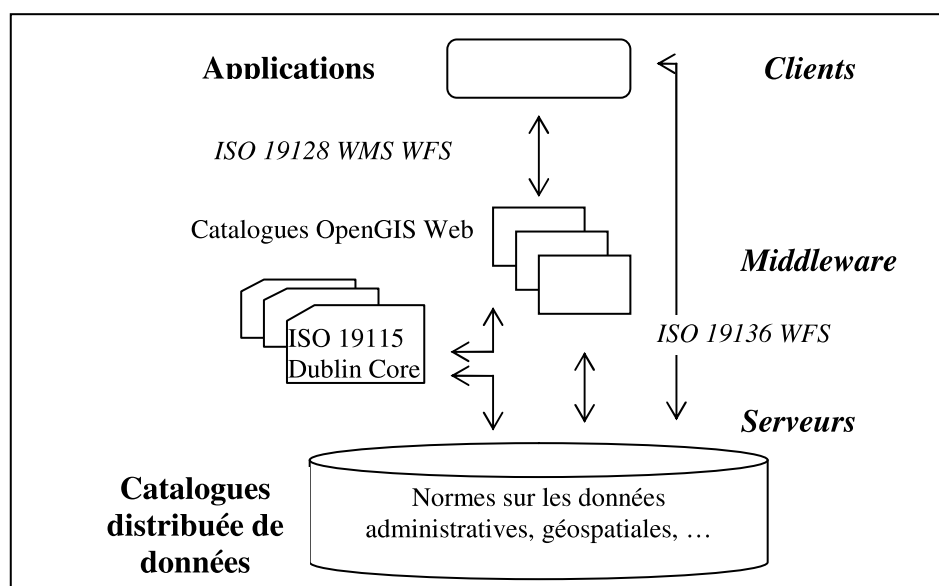
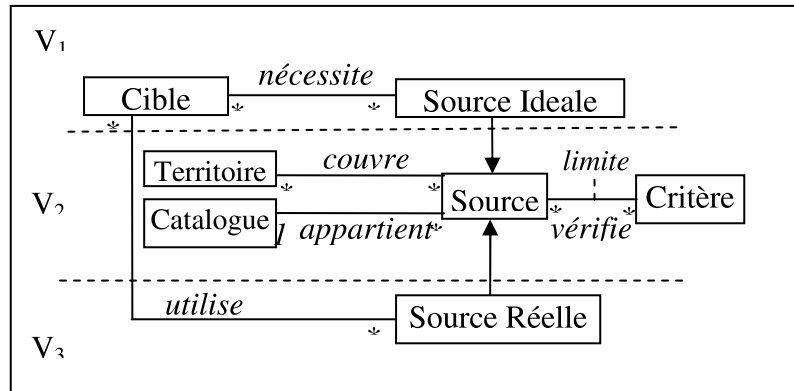


Figure 2 : Le modèle de référence d'INSPIRE (d'après INSPIRE, 2002)



**Figure 3 :** Vue intégrée des aspects d'existence, qualité et contenu des données

Une vue intégrée des trois étapes est représentée dans la figure 3. Les étapes 1 (existence) et 2 (qualité) sont représentées par les niveaux  $V_1$  et  $V_2$ , tandis que le niveau  $V_3$ , est nécessaire seulement pour les requêtes de l'étape 3, portant sur les contenus des sources de données réelles.

**Étape 1.** Pour une cible donnée  $T$  (*target*), on note  $\mathbf{rids}(T) = \{S_1, S_2, \dots, S_m\}$  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*)  $\mathbf{uds}(T) = \{S'_1, S'_2, \dots, S'_k\}$  où les sources  $S'$  vérifient les deux conditions :

- (1) il existe un catalogue  $C$  tel que  $S' \in C$  [1]
- (2)  $S'$  est en correspondance avec une source  $S_i, i = 1, \dots, m$

Selon la terminologie de (Parent, Spaccapietra, 2000) sur l'intégration de bases de données, les schémas de deux sources de données sont en correspondance si les parties du monde réel qu'ils représentent ont une intersection non vide, ou s'ils présentent des éléments qui peuvent être associés dans un but applicatif. Au niveau des instances, deux éléments (individu, ligne, relation, ...) sont dits en correspondance, s'ils décrivent le même élément du monde réel (objet, lien, propriété, ...).

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  $\mathbf{uds}(T)$ .

**Etape 2.** On note  $\Delta(T) = \mathbf{d}(\text{rids}(T), \text{uds}(T))$  la 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 :

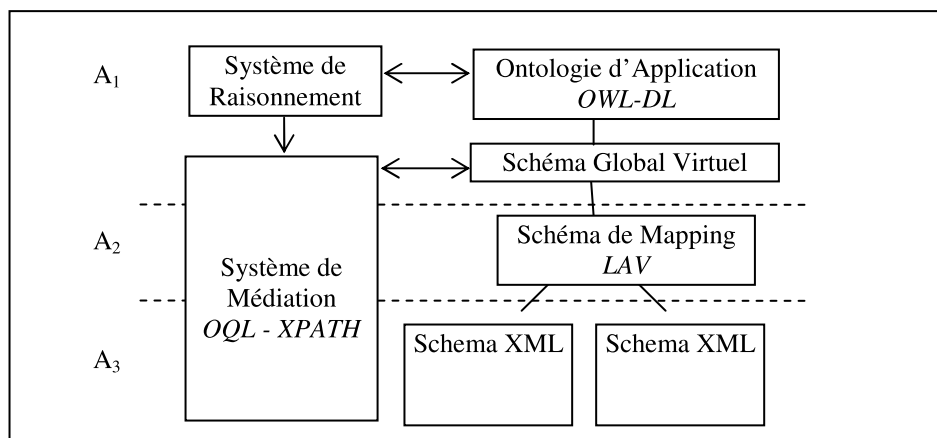
- évaluer la qualité des sources de  $\text{uds}(T)$ , par rapport aux critères et limites dérivés à partir de ceux définis sur les sources de données idéales,  $\text{rids}(T)$ ,
- choisir un ensemble optimal  $\text{uds}(T)$ , noté  $\text{ouds}(T)$ , qui minimise  $\Delta(T)$ . Des aspects organisationnels, 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  $\text{rids}(T) \equiv \text{ouds}(T)$ , i.e. : la formule [1] est vérifiée avec des correspondances qui sont des identités, tandis que la situation la plus défavorable a lieu si  $\text{ouds}(T) = \emptyset$ . En pratique, l'étape 2 consiste à trouver le meilleur compromis entre les deux.

**Etape 3.** On effectue les requêtes sur les sources de  $\text{ouds}(T)$ , si  $\text{ouds}(T) \neq \emptyset$ .

#### 4.1 Choix d'architecture

Afin de supporter ces trois étapes, on propose une architecture qui couple un système de raisonnement avec un système d'intégration. Le premier travaille sur une ontologie d'application, tandis que le deuxième est basé sur un système de médiation qui 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.



**Figure 4 :** Architecture à trois niveaux

### *Approche de médiation de type LAV*

Les principales approches, pour relier le schéma global aux sources locales, sont : *Global\_As\_View* "GAV" et *Local\_As\_Views* "LAV" (Lenzerini, 2002).

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.

## **4.2 Trois niveaux d'architecture : global, de médiation et local**

### **4.2.1 Niveau A1 : ontologie d'application et schéma global**

Le niveau global  $A_1$  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.

**Ontologie d'application** : il s'agit d'une conceptualisation formelle et explicite de concepts, propriétés et rôles (Gruber, 1993). On démarre par un nombre limité de ceux-ci, extraits des deux premiers niveaux,  $V_1$  et  $V_2$ , du modèle des classes (Figure 2). Pour représenter la qualité du processus décisionnel (étape 1) et la qualité requise pour les sources de données (étape 2), 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 :

- $\text{rads}(T) = \text{rids}(T) \cap \text{uds}(T)$
- $\text{qds}(T) = \{ds \mid ds \in \text{rads}(T) \text{ et } ds \text{ satisfait les critères}\}$
- $\text{DT} = \{T \mid \text{rids}(T) \subseteq \text{rads}(T)\}$  [2]
- $\text{WDT} = \{T \mid \text{rids}(T) \subseteq \text{qds}(T)\}$

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.

**Formalisme des Logiques de Description** : les concepts et les relations, du type [2], sont spécifiés formellement à l'aide d'un langage de Logique de Description (LD) (Calvanese *et al.*, 2004), sur lequel opère un système de raisonnement. Les LD constituent une famille de formalismes de représentation des connaissances basés sur une logique du premier ordre et sur la notion de classes (concept). Les LD fournissent plusieurs constructeurs permettant de combiner des concepts et rôles (atomiques ou composés) pour définir des concepts complexes. Dans leur grande majorité, les LD sont décidables et fournissent des services d'inférence. Une base de connaissance spécifiée en LD comprend un ensemble d'axiomes terminologiques (TBox), et un ensemble d'axiomes assertionnels (ABox), satisfaits par une interprétation I.

L'expressivité des LD est liée aux constructeurs supportés. La Table 1 présente des niveaux d'expressivité de LD, où :  $\perp$  est le concept plus spécifique, A un concept atomique, C, C<sub>1</sub>, C<sub>2</sub> des classes, P, P<sub>1</sub>, P<sub>2</sub> des propriétés et n un entier.

Symbole	Constructeurs supportés	Signification
$\mathcal{ALC}$	$\perp \mid A \mid \forall P.C \mid \exists P.C \mid C_1 \sqcap C_2$ $\mid C_1 \sqcup C_2 \mid \neg C$	$\mathcal{AL}$ est la logique de description de base et $\neg C$ représente la négation complexe.
$\mathcal{S}$	$\mathcal{ALC}$ avec $P^+ \sqsubseteq P$	abréviation pour $\mathcal{ALC}$ étendue avec les propriétés transitives
$\mathcal{H}$	$P_1 \sqsubseteq P_2$	hiérarchie de rôles
$\mathcal{O}$	$\{i\} \mid \{i_1, \dots, i_n\}$	nominaux ( <i>hasValue</i> , <i>oneOf</i> )
$\mathcal{I}$	$P_1 \sqsubseteq P_2^-$	rôles inverses
$\mathcal{N}$	$\geq n P \mid \leq n P$	restrictions non qualifiées sur les nombres
$\mathcal{Q}$	$n P.C \mid \leq n P.C$	restrictions qualifiées sur les nombres
$\mathcal{F}$	$\leq 1 P$	propriétés fonctionnelles
$(\mathcal{D})$		utilisation des types de données ou des propriétés avec types de données

**Table 1 :** Expressivité des Logiques de Description

Les LD respectent l'hypothèse du "monde ouvert" (OWA, "open world assumption"), selon laquelle l'absence d'information stockée dans une base de données est interprétée comme "ignorance" et pas comme "négation". L'hypothèse opposée, utilisée dans les bases de données relationnelles, est celle "du monde fermé" (CWA, "closed world assumption").

L'adoption de l'hypothèse CWA dans les systèmes de raisonnement peut être forcée de deux façons : (i) en utilisant des types anonymes pour spécifier les cardinalités des rôles pour chaque individu de l'ABox ; (ii) en limitant l'univers des individus connus à l'aide de l'axiome "oneOf" et en limitant le domaine des rôles aux seuls rôles participant réellement des assertions, à l'aide de restrictions de cardinalité. La première méthode permet l'insertion dans la base de nouveaux individus, sans modifier la définition des classes ou des rôles.

**Schéma Global** : est donné par l'ontologie de domaine (Visser, 2004). Celle-ci, développée indépendamment des sources de données, fournit une vue unifiée pour formuler des requêtes au niveau global. Dans notre cas, il s'agit 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.

#### **4.2.2 Niveau A<sub>2</sub> : Schéma de Médiation**

On décrit le niveau de médiation A<sub>2</sub> en suivant l'approche LAV et les principes d'une technique d'intégration décrite par (Amann *et al.*, 2002). Un ensemble de règles ('mapping rules') relie 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 d'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.

#### **4.2.3 Niveau A<sub>3</sub> : Sources Locales**

Le niveau local A<sub>3</sub> 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.

### **4.3 Choix d'implémentation technique**

#### **4.3.1 Formalismes pour la représentation des connaissances**

L'ontologie d'application est implémentée avec OWL DL (Smith *et al.*, 2004), sous langage d'OWL basé sur  $\mathcal{SHOIN}(\mathcal{D})$ , auquel on peut associer des services de raisonnement, car il est décidable.

En effets, pour formuler les concepts décrits en [1], l'expressivité de LD requise est  $\mathcal{ALCOIN}(\mathcal{D})$ .  $\mathcal{N}$  est nécessaire pour formuler des expressions CWA, avec des constructeurs qui limitent les cardinalités. Or,  $\mathcal{ALCOIN}(\mathcal{D})$  est un sous-ensemble de  $\mathcal{SHOIN}(\mathcal{D})$  (Baader *et al.*,

2005), et pas de  $SHIF(\mathcal{D})$ , auquel, dans la famille des langages OWL, est associé le sous langage OWL Light.

OWL n'adopte pas l'hypothèse Unique Name Assumption (UNA), qui assure l'unicité des individus nommés, et qui est nécessaire pour l'hypothèse CWA. Deux méthodes sont alors possibles pour forcer l'adoption de cette hypothèse : (1) configurer le système de raisonnement pour que celui-ci utilise la UNA ; (2) utiliser la description *owl:AllDifferent* et ajouter tous les individus nommés dans la liste *owl:distinctMembers*. Cette solution est plus coûteuse que la première.

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.

#### 4.3.2 Infrastructure technique

On utilise l'éditeur d'ontologies Protégé (Knublauch *et al.*, 2004) pour définir la base de connaissances et Racer (Haarslev *et al.*, 2001) 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  $SHOIN(\mathcal{D})$ . 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).

Racer est un système d'inférence 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  $SHIQ(\mathcal{D})$  (Horrocks *et al.*, 2005).

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

## 5. Application de l'approche à un exemple

Nous illustrons l'approche sur un exemple d'évaluation de systèmes de risque de santé, dans un contexte "gouvernement vers gouvernement" (G2G). Les applications cibles nécessitent des sources de données, qui couvrent des territoires (administratifs ou géographiques), à une

certaine période. Ces sources sont listées dans des catalogues. Les cibles concernent la gestion de risques de santé (canicule, vague de froid ...). Les données concernent des structures d'aide du domaine santé-sociale (les hôpitaux par exemple) et des personnes vulnérables. Ces données, produites par plusieurs organismes, peuvent être requises par plusieurs cibles. L'existence et la qualité des sources de données, renseignent sur la qualité des processus décisionnels.

### Requêtes

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

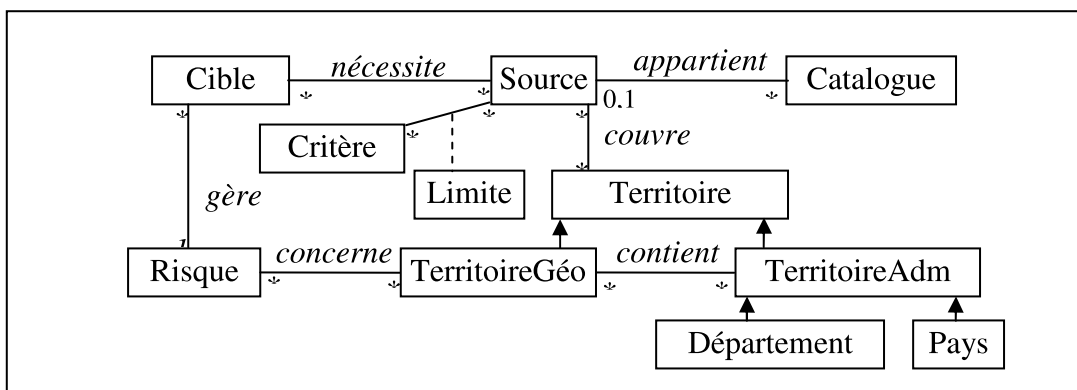
**Étape 1 - Q1 :** Quelles sont les cibles décrites sur un territoire géographique ?

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

Dans (Guemeida *et al.*, 2007) on détaille le traitement de requêtes sur le contenu, comme par exemple "Quel est le nombre de personnes âgées dépendantes, dans les départements inclus dans un territoire géographique donné ?".

## 5.1 Ontologie d'application

On restreint la recherche des correspondances dans la formule [1] aux seules identités. Dans ce cas,  $\text{rids}(T) \supseteq \text{ouids}(T)$  et l'étape 1 se limite à vérifier l'existence, dans les catalogues, des sources demandées, comme illustré dans la figure 5.



**Figure 5 :** Modèle UML de l'ontologie d'application

Ce modèle est approprié dans les contextes où les sources sont clairement identifiées, comme dans des plans nationaux d'urgences climatiques (RF-PNC, 2005). Pour souligner des

difficultés d'interopérabilité organisationnelle, on considère plusieurs spécialisations de la classe Territoire. Parmi les différentes relations (géométriques et topologiques) définies sur ces territoires, on a indiqué une relation d'inclusion.

### 5.1.1 Définition des concepts

#### Etape 1 - Q1 :

SourceDisponible  $\equiv$  Source  $\sqcap$   $\exists$  appartient.Catalogue

SourceManquante  $\equiv$  Source  $\sqcap$   $\neg$  SourceDisponible

TerritoireTG  $\equiv$  Territoire  $\sqcap$   $\exists$  contient.TG

SourceTG  $\equiv$  Source  $\sqcap$   $\exists$  couvre.TerritoireTG

SourceDisponibleTG  $\equiv$  SourceDisponible  $\sqcap$  SourceTG

CibleDécriteTG  $\equiv$  Cible  $\sqcap$   $\exists$  gère.(Risque  $\sqcap$   $\exists$  concerne.TG)  $\sqcap$

$\forall$  nécessite.( SourceDisponibleTG  $\sqcup$   $\neg$  SourceTG )

#### Etape 2 - Q2 :

SourceQualifiée  $\equiv$  SourceDisponible  $\sqcap$   $\forall$  vérifie.LimiteCritèreRespecté

SourceNonQualifiée  $\equiv$  SourceDisponible  $\sqcap$   $\neg$  SourceQualifiée

SourceQualifiéeTG  $\equiv$  SourceQualifiée  $\sqcap$  SourceTG

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

### 5.1.2 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:ObjectProperty rdf:ID="couvre">
<owl:Class rdf:ID="Catalogue"/>	<rdfs:domain rdf:resource="#Source"/>
<owl:Class rdf:ID="Territoire"/>	<rdfs:range rdf:resource="#Territoire"/>
<owl:Class rdf:ID="Source"/>	</owl:ObjectProperty>
<owl:ObjectProperty rdf:ID="contient">	<owl:DatatypeProperty rdf:ID="DatePub">
<rdfs:domain rdf:resource="#Catalogue"/>	<rdfs:domain rdf:resource="#Source"/>
<rdfs:range rdf:resource="#Source"/>	</owl:DatatypeProperty>
</owl:ObjectProperty>	...

## 5.2 Données utilisées et résultats

Les données utilisées vérifient les assertions de la table 2. Elles concernent les risques (canicule, vague de froid, ...), des hôpitaux, des personnes vulnérables. Ces dernières données sont renseignées sur des territoires administratifs.

$\text{rids}(T_1) \equiv \{S_1, S_2, S_3, S_4, S_5\}$	$\text{rids}(T_i)$ est défini par [1], $T_i$ application cible
$\text{rids}(T_2) \equiv \{S_1, S_2, S_5, S_6, S_7\}$	$S_i$ source de données
$\text{rids}(T_3) \equiv \{S_1, S_2, S_5, S_8\}$	
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 départemental	
Couvre $\{(S_1, P_1), (S_2, P_1), (S_5, P_1), (S_8, P_1), (S_3, TA_1), (S_4, TA_2),$	
$(S_6, TA_3), (S_7, TA_4)\}$	
$F(S_1, 2Y) \dots F(S_4, 6M) \dots$	
(F, Y et M indiquant respectivement : critère de Fraicheur, Année et Mois)	

**Table 2** : Jeu de données

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

$T_3$  ne fait pas partie de cette classe, car  $T_3$  appartient à une sous-classe et le système Racer classe les instances seulement au niveau des spécialisations.

**Etape 2 - Q2** : *CibleBienDécrète* $TG_1 = \{T_3\}$ , car toutes les sources de  $\text{rids}(T_3)$  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 un critère de *fraîcheur*.

D'autres concepts, basés sur *SourceManquante* et *SourceNonQualifiée*, peuvent aider l'identification de sources alternatives, moins bien qualifiées, sur ce territoire.

## 6. Conclusions

Nous avons introduit et analysé la métaphore de l'écart d'impédance dans les systèmes géographiques et de santé, dans une approche d'interopérabilité entre ces systèmes. Les écarts d'impédance sont traités à trois niveaux, qui concernent successivement l'existence, la qualité et le contenu des sources de données. Ces trois dimensions guident la conception de l'architecture. La connaissance de critères de qualité est un problème d'optimisation multidimensionnel. Ainsi, nous visons à construire des relations de préférence, pour obtenir les meilleurs compromis entre les besoins exprimés et les sources disponibles.

## 7. Références

- Amann, B., Beeri, C., Fundulaki, I., Scholl, M., «Ontology-based integration of xml web resources», *Lecture Notes in Computer Science*, vol. 2342, 2002, p. 117-131.
- Ambler, S.W., «Agile Modeling: A Brief Overview», *In Evans, France, Moreira & Rumpe (Eds.) Proc. of 'Practical UML-Based Rigorous Development Methods' Workshop, UML2001 Conference*, Toronto, Canada, 1er Octobre 2001, LNI series, vol. 7, p. 7-11.
- Baader, F., Horrocks, I., Sattler, U., «Description Logics as Ontology Languages for the Semantic Web», *Lecture Notes in Artificial Intelligence*, vol. 2605, 2005, p. 228–248.
- Barde, J., Libourel, T., Maurel, P., «A Metadata Service for Integrated Management of Knowledges Related to Coastal Areas », *Multimedia Tools and Applications*, Springer, 25(3) / March 2005, pp. 419-429.
- Calvanese, D., McGuinness, D., Nardi, D., Patel-Schneider, P., *The Description Logic Handbook: Theory, Implementation and Applications*, UK, Cambridge Univ. Press, 2004.
- Castro, J., Kolp, M., Mylopoulos, J., «Towards requirements-driven information systems engineering: the Tropos project», *Information Systems*, vol. 27 n°6, 2002, p. 365-389.
- Desconnets, J.C., Moyroud, N., Libourel, T. « Méthodologie de mise en place d'observatoires virtuels via les métadonnées » *InforSid*, Nancy, Juin 2003. Voir demo Mdweb: <http://www.mdweb-project.org>

- Guemeida, A., Jeansoulin, R., Salzano, G., «Quality-aware and Metadata-based Interoperability for Environmental Health Information», *In Proc. of the 5<sup>th</sup> International Symposium on Spatial Data Quality ISSDQ'07*, Enschede, Netherlands, 13-15 juin 2007.
- Gruber, T.R., *Knowledge Acquisition, chap.: A translation approach to portable ontology specifications*, London, Academic Press Ltd, 1993, p. 199-220.
- Haarslev, V., Möller, R., «RACER System Description», *Lecture Notes in Computer Science*, vol. 2083, 2001, p. 701-705.
- Horrocks, I., Sattler, U., «A tableaux decision procedure for SHOIQ», *In Proc. of 19<sup>th</sup> International Joint Conference on Artificial Intelligence IJCAI-05*, Edinburgh, Scotland, 30 juillet-05 août 2005, San Fransisco, Morgan-Kaufman, p. 448-453.
- Knublauch, H., Ferguson, R.W., Noy, N.F., 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.
- Lenzerini, M., «Data integration: A theoretical perspective», *In: Proc. of 21<sup>st</sup> ACM SIGACT-SIGMOD-SIGART Symposium on Principles of Database Systems PODS*, Madison, Wisconsin, 03-06 juin 2002, New York, ACM Press, p. 233–246.
- Parent, C., Spaccapietra, S., *Advances in Object-Oriented Data Modeling in Database Integration: The Key to Data Interoperability*, p. 221-254, Cambridge, The MIT Press, 2000.
- 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.
- Smith, M.K., Welty, C., McGuinness, D., OWL Web Ontology Language Guide, Recommendation W3C, 2004, <http://www.w3.org/TR/owl-guide> (accédée le 23-01-2007)
- Visser, U., «Intelligent Information Integration for the Semantic Web», *Lecture Notes in Artificial Intelligence*, vol. 3159, 2004, p. 13-34.



# **Chapitre 5: Quality-aware Agents for e-Government Information Systems Architecture**

A. GUEMEIDA, R. JEANSOULIN, G. SALZANO

In Proceedings of 3rd International Conference on e-Government, University of Quebec at Montreal, Canada on the 27-28 September 2007



Cet article approfondit des aspects de la conception de NARI, selon les objectifs spécifiques (OS) détaillés dans le Chapitre 1, §5.1.1. Nous nous appuyons sur la métaphore de l'écart d'impédance pour :

- Analyser des scénarios de Recherche d'Information en santé dans le contexte de l'e-gouvernement et bâtir un nouvel exemple d'illustration. Cet exemple permet de mettre en œuvre des relations d'inclusion et des modèles de similarité entre territoires, ainsi que des relations sémantiques entre thématiques (cfr. l'objectif OS1)
- Partir des exigences de qualité de la prise de décision pour affiner la conception de l'architecture globale de NARI et pour indiquer et formaliser des modules logiciels avec diverses fonctions (décomposition et transformation de requêtes, classifications, coordination, ...) (cfr l'objectif OS2). Ces modules peuvent être interprétés en termes de "*réactances*", qui font face aux "*résistances*" liées aux problèmes d'hétérogénéité et de qualité.
- Etendre le choix des langages de représentation des connaissances, par l'utilisation des règles et des outils associés, pour supporter ces divers modules logiciels (cfr. l'objectif OS3).

## 1. Introduction

Environmental health events, such as the French "heat wave" of August 2003, have illustrated the difficulties faced by governmental agencies to manage emergencies. Most of these difficulties, organizational, conceptual or technical, are related to the share of information, provided by huge sets of autonomous systems, related to environment, health, law, etc.

Different types of e-Government relationships are identified: Government-to-Citizen (G2C), Government-to-Business (G2B) and Government-to-Government (G2G). According with the "iceberg" metaphor (Scholl 2005), we consider G2G and inter agency collaboration (OECD 2003) as preliminary to other e-Government forms. We focus on back-office aspects of G2G problems, related to public services provision. We use the impedance mismatch metaphor in this context, and we propose a quality-aware approach for a Query-Answering System (QAS), to support a more effective decision-making.

We illustrate some critical aspects of G2G on an example in environmental health urgencies (§2). The impedance metaphor helps to structure architectural components (§ 3). The structure

is layered in three levels (§ 4), and is characterized by cooperative multi-agents patterns (§ 5). An implementation of this architecture is proposed (§ 6) and illustrated on an example (§ 7).

## 2. Critical requirements for G2G systems in health emergencies

The French heat wave, which generated a real shock, can illustrate critical requirements for G2G systems. Simultaneous difficulties to manage this crisis were:

- few days to predict crisis periods
- coordinated awareness of vulnerabilities and resources
- interactions between several phenomena, and resources shortage: hot temperature, pollution, medical services vacancies, transportation issues.

This crisis motivated a “heat national plan”, layered with actions of vigilance, alert, intervention and requisition. This plan, regularly evaluated and revised, is based on the coordination of many agents (prefectures, hospitals, urgency services), and on a set of regulations, in particular, about data collection.

### 2.1 Requirements

From this crisis, and from similar emergencies, some lessons learnt are:

- act very quickly
- associate multiple governmental agencies, with well defined roles in the organizations;
- manage and diffuse information about vulnerabilities and available resources;
- if necessary, induce adaptations to the established plans.

Then, global requirements for e-government systems are: multi-country extensibility of the system, multi-disciplinary contexts, and quality of the sources.

#### *Towards the design of an e-government Query-Answering System*

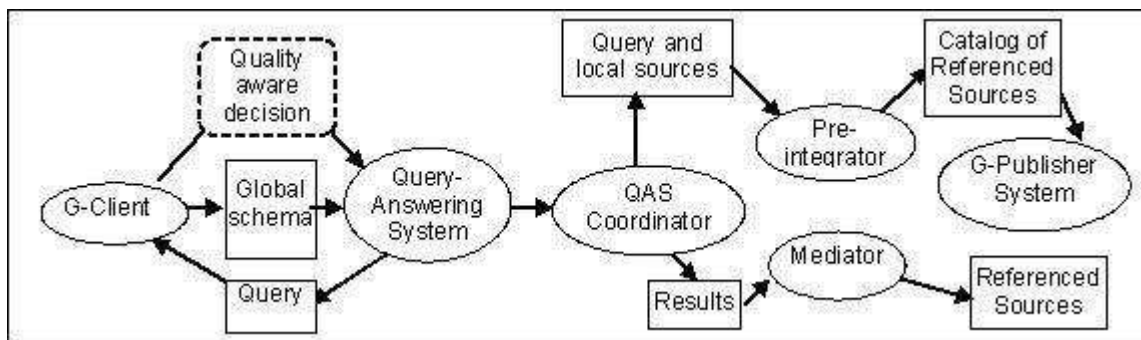
From these requirements, we propose to build a QAS in a G2G context, with three functions, related to the following problems:

- the existence of pertinent data sets corresponding to requirements;

- the quality of selected data;
- the capacity to query and merge them.

Figure 1 draws the organizational model of our QAS between two government services: a *G-Publisher* who provides referenced sources, and a *G-Client* who formulates queries on the global schema of the QAS. A Pre-integrator agent queries the existence and quality aspects of sources. A Mediator agent is in charge of translating and mapping contents. A QAS Coordinator drives the actions of the Pre-Integrator and Mediator (iteration, relaxation, exploration, etc.).

This model is based on Tropos (Castro 2002), a requirement driven development framework, to elicit early requirements about agents, their goals and dependencies. A "dependency" describes an agreement (*dependum*) between two agents, a *depender* and a *dependee*. Dependums can be goals, resources or tasks. We represent dependencies by depender → dependum → dependee, and agents, goals and resources respectively by ovals, dotted rounded rectangles and rectangles. Agents' dependencies allow identifying architectural components.



**Figure 1:** Organizational model for a Query-Answering System in an e-Government context

## 2.2 Characteristics of our approach

The main characteristics are quality awareness, traceability and adaptability.

Quality aware, in opposition to agnostic approach, constitutes a critical topic for e-Government (Bianchinia, 2006). Context and quality comparisons should be used together to perform several kinds of search, to enhance information discovery and to make it more precise (Devillers, 2005). In fact, context and quality requirements (i) reduce the set of

candidate sources; (ii) allow feedback with the client and agreement with the context; (iii) guarantee a given quality of decision and traceability.

Traceability contributes to an aware decision and to optimize the whole process. We take trace of actions of different agents: updated queries, selected catalogs and sources, transformed data sets etc. We take trace also of links between requirements and design, when a query fails and must be reformulated: memorizing failures, anticipating answers to similar queries on relaxed constraints.

Adaptability consists in exploring answers on approximate queries, validated by the G-Client, if initial queries cannot be answered.

Given a query Q with a set of referenced Catalogs, several coordinated treatments are necessary to satisfy these goals. Traceability implies that each action should classify Q at Catalog and Source levels, and store information necessary to coordination and reuse.

Example: Public agents plan actions for dependent older people (DOP), which are vulnerable in a certain territory (GT). Several cases may occur: data can be unavailable, or may not satisfy quality criteria. In these cases, one may act on the basis of approximate data, which can be calculated by filtering / aggregating data known on correlated territories; or by applying similarity models on data from territories with similar demographic and economic situations.

<i>Actions</i>	<i>Test</i>		<i>Next Actions</i>
Search for DOP/GT in catalogs	Catalogs are	Available	Search for usable sources in them
		not available	Proposed Adaptation: GT → GT', and/or DOP → P to be validated by G-Client
Search for DOP/GT in sources	Sources are	Available	Query on the contents
		not available	Proposed Adaptation to be validated by G-Client

**Table 1:** Query-Answering System steps on a simple example.

### 3. The Integration of Information and the Impedance Metaphor

The integration of information is a hard task because we must merge several constraints in one process: (i) to comply with same or similar definitions; (ii) to have same or similar data

structure; (iii) to have same or similar behavior with respect to queries and updates. This was “named” at different occasions, the impedance mismatch problem. When the object-oriented programming was introduced in the database domain, people told about the object-relational impedance mismatch (Ambler 2001). However, this idea was mostly limited to a simple word. Now, that system integration is more mature, that software agents are better identified, we propose to revisit the impedance metaphor (Guemeida 07).

The impedance was formalized in the context of electricity, hydrology, acoustics, and mechanics. The electricity domain is typical: with direct current (DC), impedance is a simple resistance, and the metaphor is poor; but it is more complex with alternating current (AC). Thus, in the information context, we need time dimension and a form of alternated interaction. Let's try to develop:

- **Intensity:**  $I$ , can be the amount of data (flow),
- **Voltage:**  $V$ , difference of potential can be the diversity range within data structure,
- **Transformers:** can modify the voltage,
- **Impedance:** complex number made of a resistance  $R$  and a reactance  $X$ :  $Z = R + iX$ ,
- **Resistance:**  $R$  can estimate a bias between the referent and its representation by the system,
- **Reactance:** if positive,  $X$  is an **inductive reactance**: preserving existing representations, and producing external energy: warning, annotation etc.
- If negative,  $X$  is a **capacitive reactance**: acts as energy storage, and can possibly help to compute aggregates for subsequent queries.

We must control the impedances  $Z_{in}$  and  $Z_{out}$ , of two connected systems, regarding the objective:

- **Impedance matching:** for a maximal power transmission. Obtained with  $Z_{in} = Z_{out}^*$ , the impedance of the user system is the conjugate to the impedance of the source.
- **Impedance bridging:** for a best control on the use of a limited amount of data, to get a rich structure, with:  $Z_{in} \gg Z_{out}$ .

When geographic information is concerned, like in Health information systems, we need specialized operators to tackle various mismatch aspects:

- **Vector versus Raster:** sets of pixels do not behave as objects, e.g.: with respect to boundaries; then we need vector-to-raster and raster-to-vector transformers that act as capacitors (keeping structure) or inductors (deriving information), and which regularize the data flow (influence on access time).
- **Geometry versus topology:** in case of uncertainty, topology constraints and coordinate information do not lead to the same issues: special capacitors are required.
- **Space scale:** a continuous change in space resolution does not behave as a linear zoom. By analogy with resistance we can avoid deriving more than what the input allows. We can also combine inductors and capacitors in an aggregation-disaggregation process, to harmonize the data to a unique scale: similar to a tuning effect.
- **Time scale:** akin space scale issue.
- **Fitness for use:** can be linked to the signal-to-noise ratio, or producer-to-user ratio. The matching becomes more difficult when the signal power goes down to the noise level.
- **Granularity:** the number of detail levels is not proportional to the number of words in a vocabulary, and discrepancies can be provoked by various uses of similar words intended to designate similar things.

## 4. Ontology and sequence structure for information integration

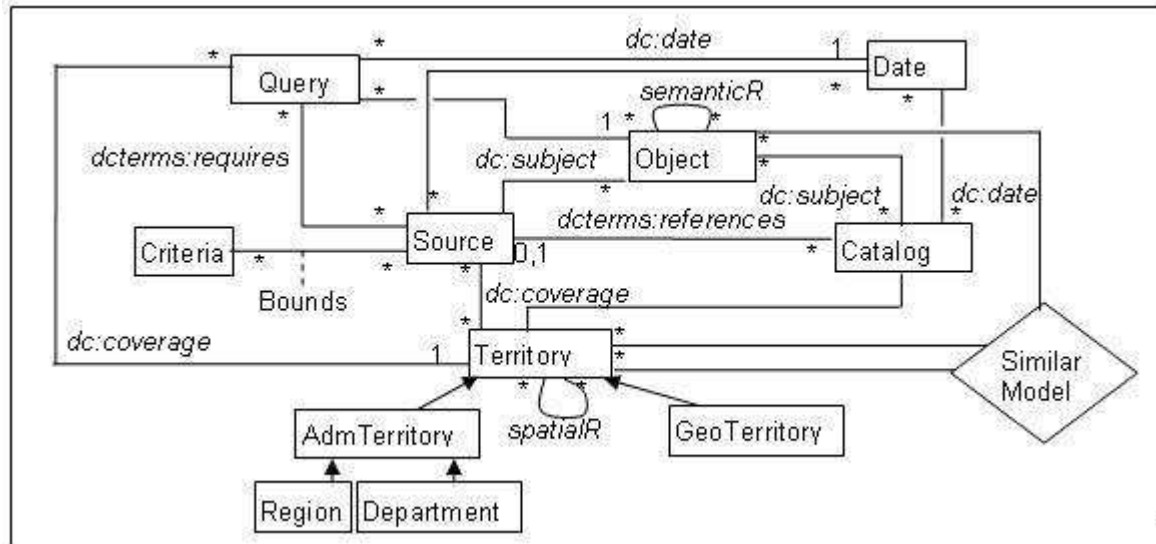
First, we need to build onto a set of concepts that allows a generic representation of the geographic information. We combine metadata and application ontology, to describe target requirements, while handling terminological and semantic discrepancies.

Then the integrated view should be broken down into three parts: Existence, Quality, and Contents.

### 4.1 Application ontology

The application ontology (Figure 2) can be interpreted as a specialization of the approach ontology and the domain ontology. We consider the Object, Territory and Date classes, corresponding to the Theme, Space and Time dimensions. Semantic and spatial relations allow exploring solutions to reduce impedance mismatches that concern Object and Territory

(SimilarModel). The "dc" prefix refers to Dublin Core Metadata (DCMI 2007). The multiplicity of links outlines the indetermination and heterogeneity of the sources for a given query, cf. § 2. The approach allows an early reduction, before full data access.



**Figure 2:** Application Ontology for a Query-Answering system (UML diagram)

## 4.2 Three steps approach

We use three levels to gradually design the components of the system (INSPIRE 2007), (Guemeida, 2007). First, we query the Catalogs that identify, locate and describe the data sets (coverage, format, etc.). Then, the Metadata give a richer description, about quality elements, the vocabulary, its granularity. Finally, the contents of the selected datasets can be queried.

The first two steps can be achieved at an early stage, and with a rapid response time, hence, efforts deserve to be done there, for improving the overall process. Let's sketch these steps.

### 4.2.1 Existence

To be checked along three dimensions:

- *Theme*: choosing relevant data starts at the very Catalog level. We need operators to match textual data description, to establish similarities, and to structure the answers. We can take advantage of direct or reverse geo-location, with smart text-processors.

- *Space*: it may look easy to simply query the *getCapabilities* OGC feature, with a zone of interest. But in case of failure, larger zones or adjacent zones can be necessary, to re-compute or approximate features.
- *Time*, to meet time requirements uses interval comparison, and some re-computing.

#### 4.2.2 Quality

Fitness for use is a multi-dimension issue. Though standard quality elements (ISO19115) do not represent the user quality needs, it makes sense to use them.

- *Positional accuracy*. if under user requirements, we must relax them, if above, we must downsize the data: combine inductors and capacitors;
- *Attribute accuracy*. similar operations to analyze metadata or to downsize the data;
- *Completeness*. degree to which the features, are included or omitted in a dataset. We can relax target requirements (undershoot), downsize data (overshoot), and/or derive new data;
- *Time accuracy, time completeness, time validity*. similar to space;
- *Lineage, logical consistency*: must be collected for further combination with additional constraints.

#### 4.2.3 Contents

It is necessary to confront the selected data from the reduced list, with the integrity constraints of the global schema. But it is cost effective because the volume is very large, and the probable detection of conflicts, makes the process intractable. We have to reduce the size of the search space, and to rank related confidence, by using appropriate preferences, i.e. partial orders.

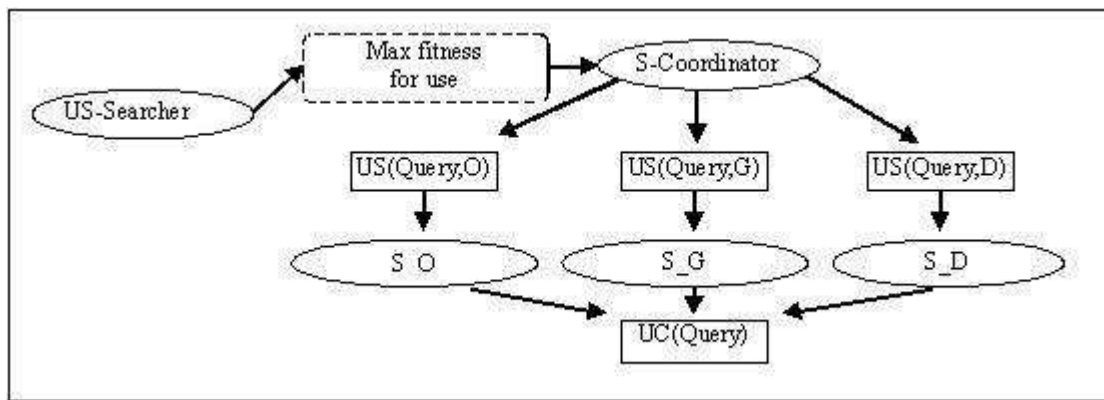
## 5. Cooperative multi-agent patterns

A suitable architecture requires advanced multi-agent patterns. The agents' coordination must reduce the impedance mismatch between the input circuit -global access to resources and catalogs- and the output circuit -expression of user's needs and queries. It must comply with early requirements, along the dimensions of space, theme, and time.

## 5.1 Catalogs and Sources Searchers

We need an agent (UC-Searcher) to provide a set of Usable Catalogs (UC) for the initial, or any subsequent query. It requires services from agents searching catalogs along specific dimensions: respectively C\_O, C\_G and C\_D for Object, Geographic Territory and Date. The agents have per input a Query and Published Catalogs; their goal is to find relevant catalogs, under the supervision of a C-Coordinator agent.

Similarly, the US-Searcher provides a set of Usable Sources (US) for the Query, using agents S\_O, S\_G and S\_D, acting on some previously selected catalogs, to find (or build) sources answering the query along its dimension, under the supervision of a S-Coordinator (Figure 3).

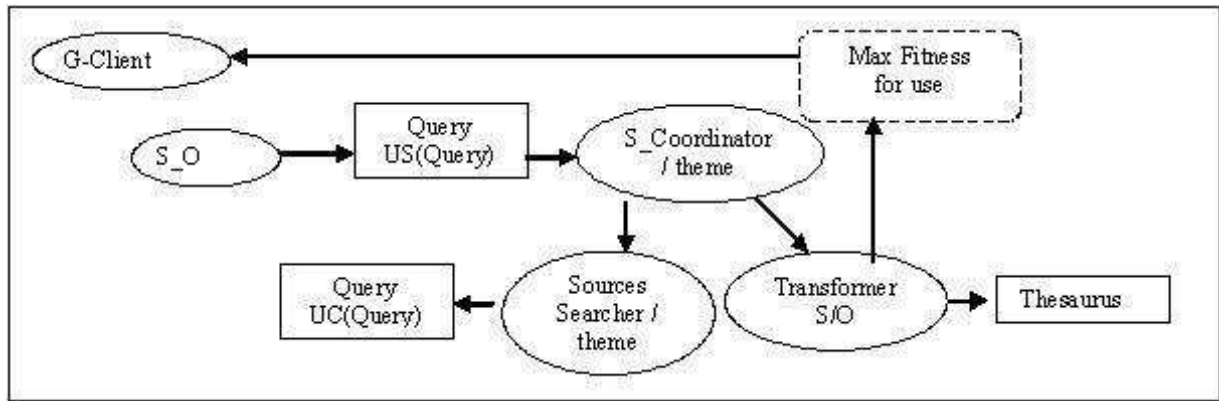


**Figure 3:** Sources Searcher composition.

## 5.2 Catalogs and Sources Searchers along specific dimensions

Catalogs and sources searchers are sets of agents. For instance, S\_O checks for sources aligned with theme O. If any source exists, TransformerS/O checks for similar themes from thesauri or other semantic structures. For similar themes validated by G-Client, it filters, aggregates, or transforms these sources, to obtain approximated and usable data.

This approach (Figure 4), adaptable to space and time, shows the necessary validation by the G-Client during relaxation choices.



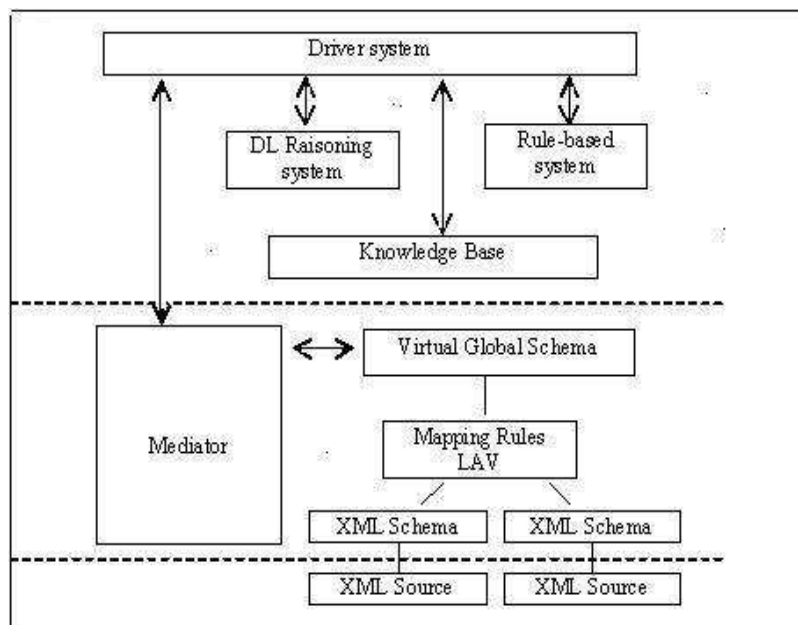
**Figure 4:** Sources Searcher along the theme dimension.

### *Organizational styles*

QAS architecture should be adapted to the organizational style of the e-government context. Organizational styles model, at a macroscopic level, the coordination of different agents towards common goals (Fuxman, 2001). Figures 3 and 4 show distributed and circular organizational styles, necessary to reduce impedance mismatches.

## 6. Global architecture

The global architecture for QAS presents three levels: Application, Mediation and Local sources (Figure 5).



**Figure 5:** QAS global architecture

## 6.1 Application level

It includes:

- A knowledge base, built on the application ontology, containing: (i) OWL DL structured classes and roles, transformation models, semantic and spatial relations; (ii) instances; (iii) rules, to generate deductive knowledge (Baader *et al.*, 2005), (Calvanese *et al.*, 2004)
- A Description Logic reasoning system, to verify consistency of the application ontology, concepts satisfiability, and to find instances of specialized classes
- A rule-based system, to infer new things, such as catalogs and sources associated to a given query, and classify these queries
- A driver system, which manages the knowledge base and communicates with Pre-integrator and Mediator agents

### 6.1.1 Agents specifications

A set of rules specify some agents' treatments (Table 2).

<i>Identifier</i>	<i>Input</i>	<i>Output</i>	<i>Treatments</i>	
C_O	Catalogs, Q(O,G,D)	UC_O	$UC\_O(C) \leftarrow Catalog(C) \wedge dc:subject(C,O)$	
C_G		UC_G	$UC\_G(C) \leftarrow Catalog(C) \wedge dc:coverage(C,G)$	
C_D		UC_D	$UC\_D(C) \leftarrow Catalog(C) \wedge dc:date(C,D)$	
TransformerC_O		Q'(O',G,D), Exists_O'	$UC\_O'(C) \leftarrow Catalog(C) \wedge dc:subject(Q,O) \wedge dc:subject(C,O') \wedge semanticR(O,O')$ IF $UC\_O'(C) \neq \emptyset$ , then $Exists\_O'=T$ , and Q is updated.	
S_Coordinator		Referenced_Sources	$Referenced\_Sources(S) \leftarrow Catalog(C) \wedge Source(S) \wedge dc:references(C,S)$	
S_O		Sources, Q(O,G,D)	US_O	$US\_O(S) \leftarrow Source(S) \wedge dc:subject(S,O)$
S_G			US_G	$US\_G(S) \leftarrow Source(S) \wedge dc:coverage(S,G)$
S_D	US_D		$US\_D(S) \leftarrow Source(S) \wedge dc:date(S,D)$	
TransformerS_O	Q'(O',G,D), Exists_O'		$US\_O'(S) \leftarrow Source(S) \wedge dc:subject(Q,O) \wedge dc:subject(S,O') \wedge semanticR(O,O')$ IF $US\_O'(C) \neq \emptyset$ , then $Exists\_O'=T$ , and Q is updated.	

**Table 2:** Examples of treatments specifications

### 6.1.2 Classifications and Coordinators

*Classifications* are used to notify the G-Clients and to drive coordination agents. Queries are classified at Catalogs and Sources levels, by using metadata.

All these classifications can be specified with a first order rules language. Description Logics (DL) are also allowed, if the involved portion of the application ontology constitutes a tree and its expression doesn't contain universal quantifiers or values comparisons. Table 3 shows classifications, expressed with rules language, while Table 4 shows classifications built in DL. Required expressivity for DL is SHI(D), corresponding to OWL-Lite.

<i>Classification</i>	<i>Necessary conditions</i>
Q_FD_C_O(Q)	Query(Q) $\wedge$ dc:subject(Q,O) $\wedge$ Catalog(C) $\wedge$ dc:subject(C,O)
Q_FD_C_G(Q)	Query(Q) $\wedge$ dc:coverage(Q,T) $\wedge$ Catalog(C) $\wedge$ dc:coverage(C,T)
Q_FD_C_D(Q)	Query(Q) $\wedge$ dc:date(Q,DQ) $\wedge$ Catalog(C) $\wedge$ dc:date(C,DC) $\wedge$ lessThen(DQ,DC)
Q_PD_C_O(Q)	Query(Q) $\wedge$ $\neg$ Q_FD_CO(Q) $\wedge$ dc:subject(Q,O) $\wedge$ Catalog(C) $\wedge$ dc:subject(C,O') $\wedge$ semanticR(O,O')
Q_PD_C_G(Q)	Query(Q) $\wedge$ $\neg$ Q_FD_CG(Q) $\wedge$ dc:coverage(Q,G) $\wedge$ Catalog(C) $\wedge$ dc:coverage(C,G') $\wedge$ spatialR(G,G')
	Query(Q) $\wedge$ dc:subject(Q,O) $\wedge$ dc:coverage(Q,G) $\wedge$ similarModel(M) $\wedge$ territory1(M,G) $\wedge$ territory2(M,G') $\wedge$ object(M,O') $\wedge$ generalization(O,O')
Q_PD_C_D(Q)	Query(Q) $\wedge$ dc:date(Q,DQ) $\wedge$ Catalog(C) $\wedge$ dc:date(C,DC) $\wedge$ graterThenOrEqual(DQ,DC)
Q_FD_C_OG(Q)	Query(Q) $\wedge$ dc:subject(Q,O) $\wedge$ dc:coverage(Q,G) $\wedge$ Catalog(C) $\wedge$ dc:subject(C,O) $\wedge$ dc:coverage(C,G)
Q_FD_C_OGD(Q)	Query(Q) $\wedge$ dc:subject(Q,O) $\wedge$ dc:coverage(Q,G) $\wedge$ dc:date(Q,DQ) $\wedge$ Catalog(C) $\wedge$ dc:subject(C,O) $\wedge$ dc:coverage(C,G) $\wedge$ dc:date(C,DC) $\wedge$ lessThen(DQ,DC)

**Table 3:** Query classifications, expressed with rules

<i>Classification</i>	<i>Necessary conditions</i>
Q_UD_C_O	Query $\sqcap$ $\neg$ (Q_FD_CO $\sqcup$ Q_PD_CO)
Q_UD_C_G	Query $\sqcap$ $\neg$ (Q_FD_CG $\sqcup$ Q_PD_CG)
Q_UD_C_D	Query $\sqcap$ $\neg$ (Q_FD_CD $\sqcup$ Q_PD_CD)
Q_UD_C_OGD	Q_UD_CO $\sqcup$ Q_UD_CG $\sqcup$ Q_UD_CD

**Table 4:** Query classifications, expressed with DL

*Used notations:* FD, PD and UD correspond to fully, partially and not described queries; C\_O, C\_G, C\_D mean "with respect to Catalogs and Object", "Catalogs and Space", "Catalogs and Time", and so on. Similar classifications arise at Sources level.

*Coordinators* active agents' arms, eventually with several loop, and classify the queries.

Coordinators could also optimize the sequences. For instance, if C\_O is the first agent called by C\_Catalog\_Coordinator, the next agent depends on the Query classification:

<i>Classification</i>	<i>Next agent</i>	<i>Input</i>
Q_FD_CO	C_G	UC_O
Q_PD_CO	TransformerC_O	Catalogs
Q_UD_CO	Exit	

## 6.2 Mediation level

### Domain ontology

The domain ontology, developed independently from data sources, describes domain semantics (Visser, 2004). It generates a global schema, from which user formulates global queries. In our case, this virtual, object-oriented schema describes concepts, with typed attributes, connected by binary relations.

In an e\_Government context, Territory is a structuring concept for the QAS, common to application and domain ontology.

### *A mediation technique*

We use an academic LAV mediation technique (Amann *et al.*, 2002). In this approach a set of mapping rules, express the correspondences between the concepts defined by the global classes and the data sources. Queries on the global concepts are formulated in an OQL variant. Global queries requiring many data sources are broken into a set of local queries, executed by the local systems, and the results are merged.

We adapt this approach by enriching the descriptions of the catalogued data sources, with quality criteria. Only data sources corresponding directly or indirectly, to global requirements are considered as Usable Sources. The other data sources, definitely out of scope, are ignored.

### 6.3 Technical choices

The technical infrastructure is based on W3C standards tools.

#### *Application level*

Two main approaches combine DL and rules (Antoniou *et al.*, 2005):

- The homogenous approaches like the Semantic Web Rule Language SWRL, which extends OWL, but the language is un-decidable and no reasoner supports it.
- The Hybrid approach, consisting in the use of existing tools, i.e. a DL reasoner and a rule engine with a common interface for knowledge base updating.

We implement the latter, using the Jena rule engine and the Pellet reasoner, in conjunction with Protégé.

Protégé (Knublauch *et al.*, 2004) is an open-source development environment for OWL ontologies building, and knowledge-based systems development. Its extensible architecture allows use of several plug-ins. Its Editor Interface can be used to generate OWL code and to communicate with a reasoning system through a standardized XML common interface.

Pellet (Sirin *et al.*, 2006), an open-source Java-based reasoner, covers full expressivity of OWL-DL. It implements a tableau-based decision procedure for general TBoxes and ABoxes. It supports the OWL-API and Jena interface.

Jena (Jena, 2007), a Java framework for building Semantic Web applications, provides a programmatic environment for RDF, RDFS, OWL and SPARQL. It includes a rule-based inference engine.

SPARQL (SPARQL, 2007), a W3C language specification to query ontology built with RDF, RDFS or OWL, does not support inference by default. However a reasoner like Pellet, allows querying the Knowledge Base ABox using SPARQL with some use limitation (Sirin *et al.* 2007).

#### *Towards Local levels*

XMLSchema and XQuery are respectively used to represent data source schemas and constraints and to query it.

## 7. Illustration

### 7.1 Presentation of a running example

We illustrate the approach on the example of section 2. For a given territory (GT) and a given health risk (hot wave, cold wave ...), the QAS would correlate the demands of services, for dependent older people (DOP), with the offer of services (hospital, beds ...). The difficulty in collecting social data, as data related to DOP, on administrative territories, is well known.

#### *Running data*

Data associated with the applications ontology and used notations are:

<i>Territories and their relations</i>		
<i>Territory</i>	<i>spatialR</i>	<i>Territory</i>
R1	contains	D11, D12, D13
R2	contains	D21, D22, D23, D24

<i>Objects and their relations</i>			
<i>Object</i>	<i>Description</i>	<i>semanticR</i>	<i>Linked Objects</i>
O1	Population	generalization	O2
O2	DOP	specialization	O1
O3	Health	generalization	O4

<i>Similar Territories with respect to Object</i>		
<i>Territory 1</i>	<i>Territory 2</i>	<i>Object</i>
D21	D24	Population

<i>Available catalogs and sources</i>
Population in all regions
Population in all departments of R1, all years
ODP in all departments of R1, all years
Population in all departments of R2, all years
ODP in departments 1, 2, and 3 of R2, only for year 2005

<i>Notations</i>	
R:	Region
D:	Administrative Department
Cik:	Catalog concerning region i and object k
Cijk:	Catalog concerning region i, department j and object k
Sikm:	Source concerning region i, object k and year m
Sijkm:	Source concerning region i, department j, object k and year m
m=1, 2	correspond respectively to 2005 and 2006

### 7.2 Queries and Results

#### 7.2.1 Q1: Number of DOP in D13 for 2006

For Q1=Q (DOP, D13, 2006), catalogs and sources exist.

Sequences of the agents called by the Catalogs and Sources Coordinators, are listed below.

Query	Agent	Input	Output	Next
Q1	C_O	Catalogs Q(DOP,D13,2006)	UC_O={C112, C122, C132, C212, C222, C232}	C_G
	C_G	UC_O	UC_G={C132}	C_D
	C_D	UC_G	UC_D={C132}	C_Coordinator
	C_Coordinator		Classification: Q_FD_C_OGD	S_Coordinator
	S_Coordinator	UC={C132}	Referenced_Sources={S1321, S1322}	S_O
	S_O	Referenced_Sources	US_O={S1321, S1322}	S_G
	S_G	US_O	US_G={S1321, S1322}	S_D
	S_D	US_G	US_D={S1322}	S_Coordinator
	S_Coordinator	Catalogs Q(DOP, D13, 2006)	Classification: Q_FD_S_OGD US={S1322}	Mediator

### 7.2.2 Q2: Number of DOP in D24 for 2005

For Q2=Q (DOP, D24, 2005), data aren't available. Approximated data are obtained by using similarity relation between D21 and D24 on the super-object *population*, and DOP data in D21. We list an extract of the sequences of the agents called by the Catalogs and Sources Coordinators, with corresponding classifications.

	Agent	Input	Output	Next
Q2	C_O	C=Catalogs Q(DOP, D24, 2005)	UC_O={C112, C122, C132, C212, C222, C232}	C_G
	C_G	UC_O Q(DOP, D24, 2005)	UC_G={ } UC_O Q(DOP, D24, 2005) Exists_G'=T; G'=D21 O'=Population	TransformerC_G
	TransformerC_G	UC_O Q(DOP, D24, 2005) Exists_G'=T; G'=D21 O'=Population	Q2.1(Population, D24, 2005)	C_Coordinator
			Q2.2(DOP, D21, 2005)	C_Coordinator
		Q2.3(Population, D21, 2005)	C_Coordinator	
Q2.1	Q(Population, D24, 2005)			
	C_Coordinator	Catalogs=C	UC={C241} Classification: Q_FD_C_O	S_Coordinator
	S_Coordinator	Catalogs={C241}	US={S2411} Classification: Q_FD_S_O	Mediator
Q2.2	Q(DOP, D21, 2005)			
	C_Coordinator	Catalogs=UC_O(Q2)	UC={C212} Classification: Q_FD_C_G	S_Coordinator
	S_Coordinator	Catalogs={C212}	US={S2121} Classification: Q_FD_S_G	Mediator
Q2.3	Q(Population, D21, 2005)			
	C_Coordinator	Catalogs=C	UC={C211} Classification: Q_FD_C_D	S_Coordinator
	S_Coordinator	UC	US={S2111} Classification: Q_FD_S_D	Mediator
Q2	C_Coordinator	Classification: Q_PD_C_OGD		
	Mediator	US={S2411} * {S2121} / {S2111} (Approximated Result Q2=Result Q2.1 * Result Q2.2 / Result Q2.3)		

## 8. Conclusions

This paper presents a quality aware approach to design global interoperability architectures for G2G systems. The approach is driven by two critical early requirements for the integration of information: traceability and contextual adaptability.

Investigation and analysis of physical systems and corresponding technical solutions could enhance the design of complex architectural components. This requires creativity and the ability to generalize solutions to organizational problems.

Thanks to the impedance metaphor, we have established some parallels between information systems and physical "circuits". Coordinated agents' patterns represent architectural components, each with different functions: to identify and select geo-localized information resources, to filter and/or aggregate data, to adapt and qualify requests from government users' and the available local sources.

Such design is based on a Local As View integration approach: we have combined metadata and application ontology to describe target requirements, at same time handling terminological and semantic discrepancies.

Knowledge representation is based on description logics and rules, and is implemented with tools (Protégé and Pellet) compliant with the appropriate standards.

Classification indicators permit the support of two information systems' functions: (i) a cooperative information management and (ii) a continuous evaluation of the organization itself. Both functions are strategic in the context of e-government.

## 9. References

- Amann, B., Beeri, C., Fundulaki, I. and Scholl, M. (2002) "Ontology-based integration of xml web resources", *LNCS*, Vol. 2342, pp 117-131.
- Ambler, S.W. (2001) "Agile Modeling: A Brief Overview", In Evans, France, Moreira & Rumpe (Eds.) Proc. of 'Practical UML-Based Rigorous Development Methods' Workshop, UML2001 Conference, Toronto, Canada, 1er Octobre 2001, LNI series, Vol. 7, pp 7-11.
- Antoniou, G., Damásio, C.V., Grosz, B., Horrocks, I., Kifer, M., Maluszynski, J. and Patel-Schneider P.F. (2005) "Combining Rules and Ontologies. A survey". REVERSE, 2005. <http://reverse.net/deliverables/m12/i3-d3.pdf>
- Baader, F., Horrocks, I. and Sattler, U. (2005) "Description Logics as Ontology Languages for the Semantic Web", *LNAI*, Vol. 2605, pp 228–248.
- Bianchinia, D., De Antonellis, V., Pernici, B. and Plebani, P. (2006) "Ontology-based methodology for e-service discovery", *Information Systems*, 31, pp 361–380
- Calvanese, D., McGuinness, D., Nardi, D. and Patel-Schneider, P. (2004) *The Description Logic Handbook: Theory, Implementation and Applications*, UK, Cambridge Univ. Press.
- Castro, J., Kolp, M. and Mylopoulos, J. (2002) "Towards requirements-driven information systems engineering: the Tropos project", *Information Systems*, Vol. 27 n°6, pp 365-389.
- Devilleers, R. and Jeansoulin, R. (Eds.) (2005) *Qualité de l'information géographique*. Hermès Science
- DCMI, Dublin Core Metadata Initiative (2007): <http://dublincore.org/>
- Fuxman, A., Giorgini, P., Kolp M. and Mylopoulos, J. (2001) "Information Systems as Social Structures", FOIS'01, October 17-19, Ogunquit, Maine, USA.
- Guemeida, A., Jeansoulin, R. and Salzano, G. (2007) "Quality-aware and metadata-based Interoperability for environmental Health information", ISSDQ, Enschede, The Netherlands.
- INSPIRE (2007), <http://inspire.jrc.it/home.html>

Jena (2007). <http://jena.sourceforge.net/>

Knublauch, H., Fergerson, R.W., Noy, N.F. and Musen, M.A. (2004) "The Protégé OWL Plugin: An Open Development Environment for Semantic Web Applications", *LNCS*, Vol. 3298, pp 229-243.

Lenzerini, M. (2002) "Data integration: A theoretical perspective", Proc. of 21st ACM SIGACT-SIGMOD-SIGART Symposium on Principles of Database Systems PODS, Madison, Wisconsin, 03-06 juin, New York, ACM Press, pp 233–246.

OECD (2003) "The e-government imperative: main findings", <http://www.oecd.org/dataoecd/60/60/2502539.pdf>

Scholl, H. J. (2005) "Organizational Transformation Through E-Government: Myth or Reality?", M.A. Wimmer *et al.* (Eds.): *EGOV 2005*, *LNCS*, Vol. 3591, pp 1-11

Sheth, A. (1998) "Changing focus on interoperability in Information System", Goodchild, M. F., Egenhofer M. J., Fegeas R. and Kottman C. A. (Eds.) *Interoperating Geographic Information Systems*. Kluwer, pp 1-25

Sirin, E., Parsia, B., Cuenca Grau, B., Kalyanpur, A. and Katz, Y. (2006) "Pellet: A practical OWL-DL reasoner", Submitted for publication to *Journal of Web Semantics*.

Sirin, E. and Parsia, B. (2007) "SPARQL-DL: SPARQL Query for OWL-DL", 3rd OWL Experiences and Directions Workshop (OWLED-2007)

SPARQL (2007), <http://www.w3.org/TR/rdf-sparql-query/>

Visser, U. (2004) "Intelligent Information Integration for the Semantic Web", *LNAI*, Vol. 3159, pp 13-34.