QUALITY-AWARE AND METADATA-BASED INTEROPERABILITY FOR ENVIRONMENTAL HEALTH INFORMATION

A. Guemeida [a], R. Jeansoulin [b, *], G. Salzano [a]

[a] Laboratoire Sciences et Ingénierie de l'Information et de l'Intelligence Stratégique (S3IS). Université de Marne la Vallée, Bd. Descartes, Champs-sur-Marne, 77454 Marne-la-Vallée Cedex 2, France - {guemeida, salzano}@univ-mlv.fr
[b] Laboratoire d'Informatique de l'Institut Gaspard Monge, Université Marne la Vallée – robert.jeansoulin@univ-mlv.fr

Commission II, WG II/7

KEY WORDS: Metadata, Data Fusion, Data Quality, Impedance Mismatch, Early Requirements.

ABSTRACT:
This paper addresses impedance mismatch problems, which occur when two information systems are plugged for working together. Such problems are particularly crucial for environmental health, where business activities require data sets from several heterogeneous and distributed systems, within autonomous organizations. We first introduce an impedance mismatch metaphor, between Information Systems, in particular between geographic data sets, developing analogy on scale problems, semantic misfits, space incompleteness, vector-to-raster issue. Then we propose an approach for a quality-aware interoperability, which is based on a step-by-step use of metadata. It focuses on early requirements, in terms of three steps: (i) existence, of relevant data, (ii) quality, sufficient for fitting the intended uses, (iii) contents, for a consistent and effective use of the data sets. We present an integrated view of these three steps, which is driven by the application requirements. A global architecture is associated with this view, by coupling a reasoning system and an integration system (mediator). The top level of this architecture is detailed, in terms of: (i) application ontology, (ii) specification languages such as Description Logics and OWL-DL, (iii) implementation choices, with Protégé and Racer systems. Finally, a simplified example illustrates the approach, and outlines several questions whose understanding can be helped by the analogy with impedance mismatch problems.

RÉSUMÉ :
Nous présentons dans cet article les problèmes d'écart d'impédance, qui surviennent lorsque deux systèmes d'information sont connectés pour fonctionner ensemble. De tels problèmes sont cruciaux en santé environnementale, où l'organisation du travail implique l'usage de nombreux systèmes différents et distribués. Nous commençons par discuter la métaphore de l'impédance dans les systèmes d'information, en particulier pour l'information géographique, en développant plusieurs analogies, pour les problèmes d'échelle, de mésentente sémantique, d'incomplétude, ou de différences entre vecteur et rasteur. Puis nous présentons une approche d'interopérabilité qui fait un usage pas à pas des méta-données, en mettant l'accent sur les besoins premiers, selon trois étapes : (i) l'existence de données pertinentes, (ii) une qualité suffisante pour ces besoins, (iii) la cohérence des données elles-mêmes et la capacité à les traiter. Nous présentons une vue unifiée de ces trois étapes, dans une architecture globale qui associe un système de raisonnement et un médiateur. Ceci est détaillé en termes de : (i) ontologie d'application, (ii) langage de spécification, comme « OWL-Description-Logic », (iii) implémentation avec « Protégé » et « Racer ». Enfin, nous illustrons cette approche sur un exemple simplifié qui traduit certaines des questions posées dans les logiciels proposés, avec l'aide de la métaphore de l'adaptation d'impédance.

1. INTRODUCTION

1.1 Where Matching Matters

Due to the recent flooding of huge amounts of geographical data, available through the Internet, observation and data registration become completely independent on data querying or mining. Final users are no longer accustomed with the numerical data sets, as they may have been with paper maps; rather, they discover the data on the fly, out of large or very large information systems.
A consequence is the need to rely, more and more on automatic support for deciding what to choose or reject, among tons of candidate data. Also, as any application uses a mix of many data sets, it is more appropriate to talk about system interoperability rather than about simple data merging: this is to insist on the fact that each information system, which supports each different data, influences the interoperability process, and not only the chosen data.

1.2 Where Quality Matters

The necessity to rely on partly automated systems for geographic information, as we do already with text queries filtered by web robots, implies to take data quality into account. Thus, the quality of the service must compromise the declared quality of the data, with the quality level that the user is able to accept. The translation between these two visions of the quality is an issue since years, in geographic information, and is source of a certain mismatch between what is obtained and what was expected. This overall situation leads us to consider the phenomenon of "impedance mismatch" that occurs in general when two systems are plugged for working together.
In this paper, we will develop on the "impedance" metaphor (sec.2) as a possible hint to better structure and categorise, the various translators to include in the data flow. Then this structure is broken down into three levels (sec.3), and a mediation architecture is proposed for its implementation (sec.4). This is illustrated on a simple example, taken from the domain of public environmental Health (sec.5).

* Corresponding author.

2. IMPEDANCE MISMATCH IN GEOGRAPHIC INFORMATION FUSION

2.1 The Impedance Mismatch Metaphor

By analogy with electrical systems, acoustic, hydraulic, or mechanical systems, we can name this phenomenon the "user-provider impedance mismatch". In the case of direct current (DC), impedance is simply a resistance, but it is more complex for alternating current (AC). Without digging too much deeper in this analogy, we may mimick some physical quantities:

- **Intensity**: I can be the amount of data (flow),
- **Voltage**: V is a difference of potential, what, in the case of information, can be the diversity range within data,
- **Transformers**: can modify the voltage, only for AC,
- **Impedance**: complex number made of a resistance R and a reactance X: $Z = R + \imath X$,
- **Resistance**: R can be the bias between the referent and its representation performed by the information system,
- **Reactance**: if X is positive, it is an inductive reactance, and can be interpreted as a tendency to preserve older representations, legacy systems, a backward translator for minimizing comparison distance. The consequence of this preservation is that external energy is produced, for instance in the form of warning, appendices, annotations etc, not to be used directly by the user main decision system,
- If negative, X is a capacitive reactance, which can act as energy storage, an information cache or an information filter (e.g.: reducing coordinates), and possibly transferring the intentional part (e.g.: a partial order) of the information, rather than its extensional part, it can also help to compute aggregates (e.g.: average value).

When two systems are connected, it is important to control the difference between their impedances Z_{in} and Z_{out}, and to take the appropriate answer depending on the objective:

- **Impedance matching**: if the objective is to provide a maximal power transmission, for a best use of total information. It is obtained with $Z_{in} = Z_{out}^{*}$, the impedance of the user system is the conjugate to the impedance of the source. Hence we need to equalize the resistances $R_{out} = R_{in}$, and to oppose the reactances $X_{out} = -X_{in}$ (An inductor against a capacitor or vice-versa).
- **Impedance bridging**: if we want a best control on the use of a small part of the information: for instance, a rich structure with a limited amount of data. It can be obtained with $Z_{in} >> Z_{out}$.

When using remote information systems, some flexibility is allowed, but only within some usability limits, which we will try to characterize, and to understand in this paper. To overcome the impedance mismatch, always entails some overhead, due to the necessary introduction of some intermediate device, to help the translation or transfer process.

2.2 Impedance Mismatch between information systems

Information Systems (IS) are situated in an organization and include several levels (application software, support software, computer hardware). When designing it, multiple mismatches can arise at each level and each task (early and late requirements, global and detailed architecture). Organizational level, and early requirements are usually recognized as critical for a successful design of a target IS, and become particularly crucial for IS modernization projects (i.e. e-government), because requirements concern simultaneously organization it-self and levels below.
(Castro &al. 2002) relates IS impedance mismatches to the gap between the system-to-be and its operational environment. They argue that system-to-be is determined by development driven methodologies, based on concepts different from those related to the operational environment. To reduce impedance mismatches between all phases of development process, they propose a requirement driven development framework, starting from early requirements about several actors, their respective goals and the dependences between these goals.

2.3 Impedance Mismatch between geographic data sets

When considering geographical information, we generally face several issues that we can relate to the impedance mismatch of the GIS.

Vector versus Raster (or object versus field issue), where objects do not commensurate with field parts, neither pixel aggregate with objects. This kind of mismatch looks similar to what has been identified as the "object-relational impedance mismatch" (Ambler, 2001): access by pixel (set of data) and not by object behavior (boundaries, topology). Example: the user wants to overlay any information layer as a grid of pixels, then we should introduce a vector-to-raster transformer that will act as an inductor, and will regularize the data flow as a constant flow of pixels, whatever the input format.

Geometry versus topology mismatch: though we can theoretically derive the topology from a perfect geometry, in most real situations, to give too many geometric details, can hamper the topology clarity, and even consistency, if details are slightly inaccurate. It can be seen as a particular case of the previous mismatch. Example: if the target application must rely on topological constraints, we should build a "capacitor" to filter the overflow of geometric coordinates and retain only their topological consequences.

Space scale issue, where an apparent gradual continuous change in space resolution does not match with a linear (homothetic) zoom factor. This is related to the bandwidth issue in signal processing, the bandwidth used in the Shannon theorem to establish a channel capacity. The theoretical "Nyquist rate" teaches us that the relevant object can be distinguished if they are twice as large as the channel bandwidth. Here, the analogy concerns the resistance rather than reactance. In fact, we must in general adapt our requirements to a range compatible with the input range (resisting to derive more than what the input allows). When several sources are mixing several scales, an aggregation-disaggregation process can help to harmonize the data to a unique scale, and this is similar to a tuning effect produced by a combination of inductors and capacitors.

Time scale issue: similar to space scale issue.
Fitness for use: The fitness for use is sometimes referred to as the external quality, or user-defined quality, and opposed to the internal quality, which is defined for each data set, by its producer. This can be linked to the signal-to-noise ratio, where signal is a producer-side notion, and noise a user-side one. It isn't easy to reduce the fitness for use to a single ratio, because several independent quality components are involved, but we can use the impedance analogy to confirm that the impedance matching is becoming more difficult when the signal power goes down to the noise level.

Granularity of the description (specialization hierarchy): It is an issue, because the number of detail levels does not increase necessarily with the number of words in a vocabulary, and discrepancies can be provoked by various scopes and range of uses of similar words intended to designate similar things.

We will focus on the last two types of impedance mismatch: granularity and quality.

3. A THREE STOREY STORY

3.1 Existence, Quality, and Contents aspects

Data collection and selection process is a huge, and a costly part of the work, the next part being the activation of the user model in order to make a decision. In small applications it is reasonable to group that into a unique process, whose impedance should be adapted to the variety of available sources. In large applications it may go out-of-control, and we should rather consider the data selection itself as one task. Structuring and implementing that, necessarily results in some impedance. Then issue is that this impedance must fit to the sources and to the user models as well. And the new question is: how many user models can we manage with only one data selection system? (more than one?).

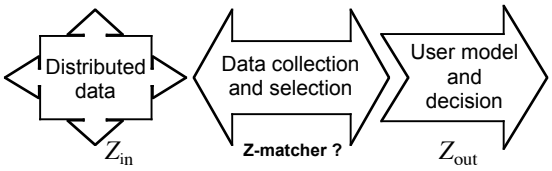

Figure 1: processes and impedance matching

Public Health bodies, for instance, are collecting significant amounts of data, produced by several heterogeneous and distributed systems, which work at different spatial levels (international, national, regional, departmental, communal), and with different perspectives (epidemic monitoring, health expenses control, hospital management…). For any peculiar purpose (e.g.: disease analysis), the integrated view of the whole set of information, should be broken down into three aspects:

- Existence of relevant data;
- Quality, sufficient for fitting the intended use;
- Content, for a consistent and effective use.

3.2 Catalogues, Metadata, and Data storeys

Questioning relevance, fitness, and consistency, contributes to the overall impedance issue, between the user system, and the whole distributed system of all potential data. But do we need to analyze this whole input system, to build the impedance that should oppose it?

In order to break down this task, let's consider what can be gradually learned about the input system. First, come the **Catalogues**, which identify data sets, give their location, plus some description (space coverage, format). Second come the **Metadata** for each catalogued data set, which give a richer description of the content, about various aspects of its quality, about the vocabulary and its granularity. Finally, come the **Data**. This suggests us to use these three levels step by step, to gradually design the components of Z_{out}.

3.3 A three steps impedance builder

3.3.1 Step1 Existence

(a) **Geometry**. To find relevant data looks easy on a simple geometric aspect: to query a Catalog with a country or region name, then to ask the *getCapabilities* OGC feature for intersection with a rectangular zone. It simulates resistance equalization.

(b) **Time**. To meet the time requirements can be done by interval equalization (easy), or by accepting a much larger time interval, with an additional 'inductive' processing. Here, the word inductive makes sense as an electrical analog and a logical term, meaning that regularization is necessary.

(c) **Theme**. To find relevant data is very approximate for thematic aspects. For instance, the answer to: "is the layer *SRTM* relevant for my purpose?" is uneasy, if you ignore what this acronym means (*Shuttle Radar Topographic Mission*). A good choice is necessary at the Catalogs level, and even at the Catalog of Catalogs level, and we need smart operators, able to match terms (same resistance), extracted from titles, from textual description of data sets, but also able to establish similarities (inductive regularization), and possibly to build some hierarchy between related terms (capacitor). This should be helped by a new generation of Internet robots, combining direct or reverse geo-location, with smart text-processors. Rapid browsing can select too few, or cautious approach can select too much, depending on the strength or permissiveness of the impedance of our operators, but we can expect to avoid most of the totally irrelevant sources (in space and theme).

3.3.2 Step2 Quality

Fitness for use is a multi-dimension quality question, which is not easily convertible into quality metadata elements; but it is mandatory to have a reliable software support here, in order to avoid to proceed to the next step with too many data, or to miss a small set of data that may be very significant. Though the standard quality elements (ISO19115) do not represent the user quality vision, it makes sense to use them in the description of the incoming impedance Z_{in}.

Positional accuracy. Absolute accuracy is the closeness of the coordinate values in a dataset to values accepted as true [i.e.: by the producer]: to meet user requirements can be achieved by adapting the output resistance, if undershoot, or by preparing to downsize the data (next step 3), if overshoot. Relative accuracy is the closeness of the relative positions of features to their respective relative positions accepted as true: this is important for topology preservation, and it will need an appropriate capacitor for the next step, in order to compute, from the data, the required topological constraints, and to confront them with expected constraints.

Attribute accuracy. Similar operations: to include a resistance at this very level (metadata analysis), or to prepare an inductor to downsize the data in step 3.

Completeness. In a vector representation, it refers to the degree to which the geographic features, their attributes and their relationships are included or omitted in a dataset. A combination of resistances (undershoot case), inductors (overshoot), and capacitors (if some positive or negative inference should be derived), is the solution. And again, this is a preparation for the operators required in the next step.

Time accuracy, time completeness, time validity. Similar operations are expected there.

Lineage. Logical consistency. This information must be collected (capacitor) for further processing in step 3, where it will be combined with additional constraints created by other impedance elements, e.g.: a topologic capacitor, an integrative resistance, etc.

3.3.3 Step3 Contents

Once a reduced list of data sets has been selected, it is necessary to confront them to the integrity constraints of the global schema. This task is cost effective because we are now at the data level, which is much larger. Moreover, the probable detection of conflicts, between the actually merged data, makes the whole process intractable. It is mandatory to reduce a-priori the size of the exploration space by using an appropriate preference order. The same approach can be used to order confidence levels for the retained solution.
Such partial orders are built by machine learning algorithms that can be based on statistics (e.g. Bayesian), or qualitative ranking (e.g. Formal Concept Analysis), or a mix. Then a decision must be taken about accepting the data, possibly issuing an associated warning. If data are rejected, a new query must follow, back to the upper levels. This is where the AC analogy makes sense: structural information retained by the various capacitors, will help to rephrase the next query, with a limited expectation, but with a preserved focus. And the process may loop until a decision.

4. MEDIATION ARCHITECTURE

The first two steps can be achieved, most of the time, at an early stage, by appropriate browsing of data catalog, and metadata examination. This approach complies with the global structure adopted by ISO standards and INSPIRE (Figure 2). Using catalogs and metadata we can anticipate a reliability level for the decision outcome, and we can take the decision to go further, with an a priori best selection of data, before completing the fitness for use assessment (third step of queries), by accessing actual data from heterogeneous and distributed systems.

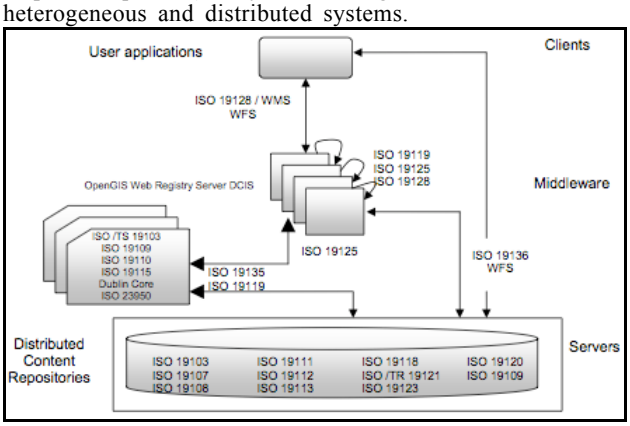

Figure 2: the INSPIRE architecture reference model, and associated standards (INSPIRE, 2002).

4.1 A requirements driven view of the three steps

We model an integrated view of the three steps as represented on the Figure 3.

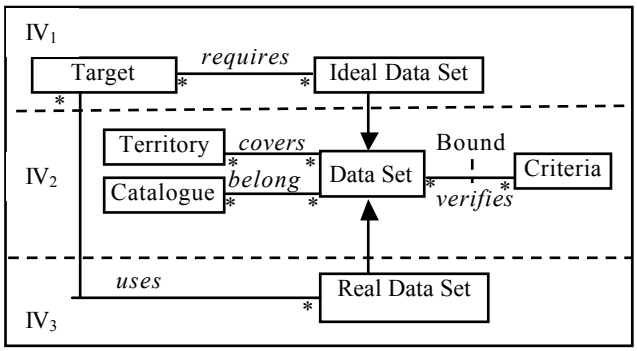

Figure 3. Integrated view of the existence, quality and content aspects

Top and middle levels, IV_1 and IV_2, of this integrated view concern requirements for step1 (existence of relevant data) and step2 (quality of the available data). Level IV_3 is only necessary to the step3 questions on contents of real sources. We give now the general lines for these steps.

Step1. For a given target T, be **rids(T)** = $\{S_1, S_2, \ldots S_m\}$ the set of ideal data set required by T. Step 1 must determine the 'usable' data sets **uds(T)** = $\{S'_1, S'_2, \ldots S'_k\}$ such that

$$\forall\, i = 1, \ldots m \;\; \exists\, j = 1, \ldots k \quad \exists\, c_{ij}\,(S_i, S'_j,) > tc \qquad (1)$$

where $S_i \in$ Ideal Data Set, $S'_j \in$ Real Data Set, c_{ij} is a correspondence between S_i and S'_j, better than a threshold *tc*, eg: a minimal number of constraints to satisfy.

The c_{ij} are defined in the sense of (Parent, Spaccapietra, 2000). To find it, we should (i) explore different dimensions: theme, space and time, and (ii) use semantic, geometric and topological relations between the metadata of the catalogues and sources (usability study).

Capacitive action: when computing **uds(T)** for *tc*, some near-to-*tc* sets can be memorised into **uds'(T)** (capacitor). For instance, if we have the required data for a neighbouring region, or at a more global scale: just keep track of that, for saving time during a possible next call.

Step2. We note **Δ(T)** = **d**(rids(T), uds(T)) the gap between required and usable data sets for a target T, with respect to some distance function d. Step2 consists to:

- evaluate the quality of uds(T) with respect to some criteria and bounds derived from those related to the corresponding data sets rids(T),
- choose one optimal uds(T), noted **ouds(T)** that minimizes Δ(T) (w.r.t. organizational, conceptual or technical aspects ≈ impedance matcher).
- if no **ouds(T)** can be found, release threshold *tc* and go back to step1, emptying the capacitor **uds'(T)** and possibly recharging it. Some extra processing (inductor) must be activated to enhance the new, but less fitting **uds'(T)**: for instance a similarity model or an aggregation/de-aggregation model to compute approximate data from **uds'(T)**.
- repeat until **Δ(T)** become acceptable, with respect to a quality balance, or abort and report failure reasons.

Step3. It integrates data sets of ouds(T), if they exist.
The "best" situation arises when rids(T) ≡ ouds(T) (i.e. (1) is satisfied with c_{ij}= identity, $\forall$ i,j = 1, … m), while the "worst" situation arises when ouds(T) = ∅.

Role of quality issues
Steps 1 and 2 reduce the volume of data sources addressed to step 3, compromising between (i) queries and quality needs expressed by the target system, and (ii) existing data and their quality. If necessary, they mutually adapt (i) and (ii), to obtain acceptable transformation of (i) and/or acceptable costs for inductor and capacitors (filters, caches …) on (ii).

Classification of data sets and targets
Steps 1 and 2 also generate classifications of data sets and targets. For instance: "required and available", or "qualified" data sets (resp.: **rads(T)**, **qds(T)**), "described" and "well described" targets (resp.: **DT**, **WDT**):

$$rads(T) = rids(T) \cap uds(T)$$
$$qds(T) = \{ds \mid ds \in rads(T) \text{ and } ds \text{ satisfies some criteria}\} \quad (2)$$
$$DT = \{T \mid rids(T) \subseteq rads(T)\}$$
$$WDT = \{T \mid rids(T) \subseteq qds(T)\}$$

These concepts introduce orders at each level (existence, quality) for Target and Data Sets classes. The order relation between application's hierarchies depends on the order relation between data sets hierarchies related to it.

4.2 Architecture: a LAV approach for the mediator

The proposed architecture couples a reasoning system and an integration system (mediator). The first operates on application ontology, while the second is based on (i) a global schema, (ii) a set of sources, containing real data, and (iii) a set of relations between the global schema and the local sources. Integration architecture assures a maximum control on the activities related to the information management: to acquire, evaluate, use and diffuse information, as required e.g. in e-government contexts.

To model relation between global schema and local sources two approaches have been proposed: the *Global_As_View* "GAV", and the *Local_As_Views* "LAV". A theoretical survey of these is given in (Lenzerini, 2002). Our view follows a LAV approach, to characterize the local sources as views over a global schema, to be built first by taking care of the user needs and expectations. Thus, priority is given to the global requirements, in terms of concepts and objects (ontology), as well as quality of data necessary for a reliable decision. In opposition a GAV approach would have determined the global schema and possible requirements only from the local ones (the available data sources). Though the GAV seems more natural, and easier to implement, it contradicts the priority to the global requirements (extensibility of the system, quality of the sources), and we prefer LAV instead.

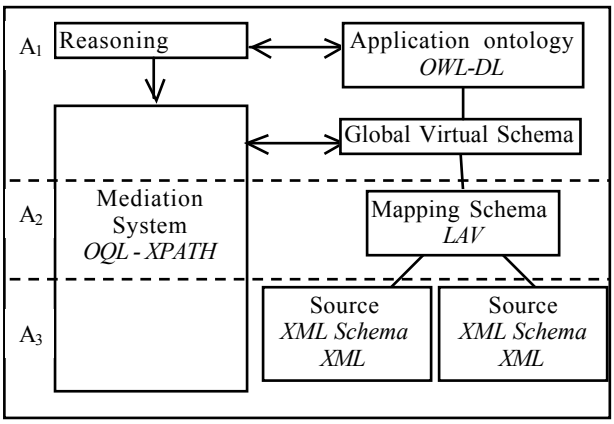

Figure 4. Three-level architecture

4.3 Three levels: global, mediation and local

As in Figure 4, this architecture presents three levels.

4.3.1 Level A_1: Global ontology and virtual schema

The global level $\mathbf{A_1}$ is the application level, where decision-making requirements are defined. It contains application ontology and a global virtual schema.

Application ontology: in the (Gruber, 1993) terminology, it is a formal and explicit conceptualization of the application requirements. It starts with a few concepts, properties and roles, extracted by the top and middle levels of the class model (Figure 3). In order to represent the quality of the decision process (step1) and the quality of the data sources (step2), it is often necessary to derive new classes from the set of primitive classes, as well as their complementary classes. These classes are based on formulas (2).

Description Logics Formalism: concepts and relations (like (2)) are formally specified in a Description Logics (DL) language (Calvanese et al., 2004), and operated by a reasoning system. DL belongs to a family of knowledge representation formalisms based on first order logic and the notion of class (concept). DLs make it possible to use various constructors to build complex concepts based on previous ones. In their great majority, they are decidable, and provide complete reasoning services. A DL knowledge base consists of a set of terminological axioms (called TBox) and a set of assertional axioms (called ABox): see table 1.

Name	Syntax	Signification
Concept inclusion	$C_1 \sqsubset C_2$	C_1 is a sub class of C_2
Role inclusion	$P_1 \sqsubset P_2$	P_1 is sub property of P_2
Concept equality	$C_1 \equiv C_2$	C_1 is equivalent to C_2
Role equality	$P_1 \equiv P_2$	P_1 is equivalent P_2
Concept assertion	C(a)	a: an individual belongs to C
Role assertion	R(a, b)	a has a role R with b

Table 1. Terminological and assertional axioms (example)

Table 2 presents some DL constructors, where C, C_1, C_2 are classes, P a role, x and I an individual and n an integer.

Name	Syntax	Signification
intersection	$C_1 \sqcap C_2$	{x, belong to both C_1 and C_2}
union	$C_1 \sqcup C_2$	{x belong to C_1 or C_2}
negation	$\neg C$	{x don't belong to C}
existential quantification	$\exists P.C$	{x having some values of P in C}
v. restriction	$\forall P.C$	{x having all values of P in C}
existential quantification	$\exists P.\{I\}$	{x having I as a value of P}
Unqualified number restriction	$\geq n\ P$	{x with at least n role of P}
	$\leq n\ P$	{x with at most n role of P}
	$= n\ P$	{x with exactly n role of P}
Qualified number restriction	$\geq n\ P.C$	{x with at least n role of P in C}
	$\leq n\ P.C$	{x with at most n role of P in C}
	$= n\ P.C$	{x with exactly n role of P in C}
Nominal	$\{I_1, ..., I_n\}$	A list of individuals

Table 2. Some Description Logic concept Constructors

Global schema: is the domain ontology, independently developed from data sources, which provides a unified view of the heterogeneous, distributed and autonomous sources, to describe their semantics and formulate user global queries (Visser, 2004). In our case, it is an object-oriented schema

describing concepts, with typed attributes, connected by binary relations. This conceptual model can be implemented in order to perform syntactical and lexical verifications of global query. Application and domain ontologies have common concepts (e.g.: Target and Territory).

4.3.2 Level A_2: Mediation schema

Mediation level **A_2** is described according to an academic integration technique (Amann &al., 2002). The correspondences between the concepts defined by the global primitive or derived classes, and the data sources are expressed, according to LAV approach, by a set of mapping rules. The queries on the global concepts are formulated in an OQL variant. A global query is broken into a set of local queries, executed by the local systems, and the results are merged.

4.3.3 Level A_3: Local Sources

The local level **A_3** is made of the existing data sources, such as registered in catalogues, and such as described by their schemas. They are completed by some metadata, corresponding to the quality criteria (step2). Only data sources corresponding directly or indirectly, to global requirements are marked. The other data sources, which are definitely out of scope, are ignored. Next section describes this process.

4.4 Technical implementation choices

Knowledge representation formalisms:
The application ontology is implemented in OWL DL (Smith et al., 2004), a sub-language of OWL that allows use of Description Logic reasoning services, since it is decidable. It permits to formulate new concept definitions based on previous ones.
OWL DL is based on $\mathcal{SHOIN}(\mathcal{D})$ (Baader et al., 2005), and supports transitive properties, role hierarchies, nominals, unqualified number restrictions and data types.

Links between DL and OWL:
Table 3 presents correspondences between the DL syntax constructors, presented in tables 1 and 2, and the OWL DL language. One can use graphical ontology tool, like Protégé, to generate OWL DL code from DL descriptions:

DL	OWL
$\forall P.C$	allValuesFrom
$\exists P.C$	someValuesFrom
$\exists P.\{i\}$	hasValue
$\geq n\ P$	minCardinality
$\leq n\ P$	maxCardinality
$= n\ P$	Cardinality
$C_1 \sqcap \ldots \sqcap C_n$	intersectionOf
$C_1 \sqcup \ldots \sqcup C_n$	unionOf
$\neg C$	complementOf

DL	OWL
$C_1 \sqsubseteq C_2$	subClassOf
$C_1 \equiv C_2$	equivalentClass
$P_1 \sqsubseteq P_2$	subPropertyOf
$P_1 \equiv P_2$	equivalentProperty
$C_1 \sqsubseteq \neg C_2$	disjointWith
$\{i_1\} = \{i_2\}$	sameAs
$\{i_1\} \sqsubseteq \neg\{i_2\}$	differentFrom
$P_1 = P_2^-$	inverseOf
$\{i_1, \ldots, i_n\}$	oneOf

Table 3. Correspondences between DL and OWL

Tools:
The technical infrastructure is based on Protégé OWL editor (Knublauch et al. 2004), from Stanford University, and Racer reasoning system (Haarslev, Möller, 2001).
Protégé is an open-source development environment for ontology building in OWL DL, and knowledge-based systems. It supports $\mathcal{SHOIN}(\mathcal{D})$ and it has an extensible architecture allowing use of several useful plug-ins. Protégé can be used with a DL reasoning system through a standardized XML common interface.
Racer is a description logic reasoning system that automatically classifies ontology and checks for inconsistencies. It implements highly optimized algorithms for the very expressive description logic $\mathcal{SHIQ}(\mathcal{D})$. It offers reasoning services for multiple TBoxes and ABoxes.
The knowledge base permits to perform existence (step1) and quality (step2) interrogations. The application ontology corresponds to the TBox part, while the ABox part contains individuals and their descriptions.
We consider that the Metadata elements describing data sources belong, at the same time, to the data sources and to the knowledge base, as part of catalogues descriptions.

Towards Local levels:
Local queries are translated into the local supported query language i.e. XQuery.
At local level, XML Schema and XQuery are respectively used to represent data source schemas and constraints and to query it.

5. APPLICATION

5.1 A simple example

We consider a set of target applications related to the health risks management. For each risk, we want correlate, over a geographic territory (GT) concerned by the risk, the demands of services, for dependent older people (DOP) and other vulnerabilities, with the offer of services (hospital, beds, …). A lot of these social data (as data related to DOP) are collected on administrative territories (departments), while scientific data are in general related to geographic territories. An example of query, essential for each target, is:

Q: For a GT concerned by a risk, and for each department in GT, how many are DOP?

Q is formulated from a fragment of the global schema represented in the figure X.

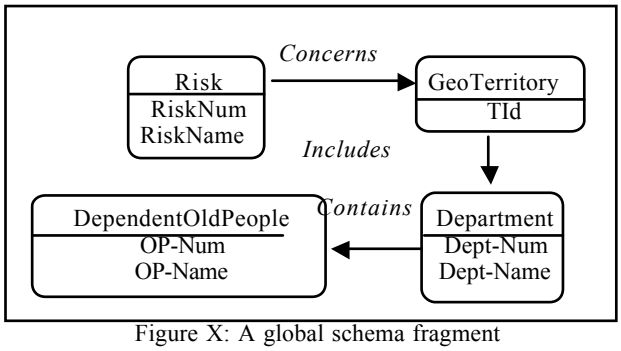

Figure X: A global schema fragment

Running data

T_1, T_2 and T_3 represent respectively heat wave, cold wave and inundations. With notation of §3.1, we suppose rids(T_1) _ {S_1, S_2, S_3, S_4, S_8}, rids(T_2) _ {S_1, S_2, S_3, S_4, S_5, S_6, S_8} and rids(T_3) _ {S_1, S_2, S_3, S_4, S_7 , S_8}. These sources are linked to the risk management activities; for instance: S_1 relays risks and departments, S_2 represents hospitals, S_3 and S_4 DOP on two departments, respectively AT_1 and AT_2,, S_5 and S_6 homeless on the same departments, S_7 camp-sites on all departments. S_8 gives relations between geographic and administrative territories: for instance, GT contains AT_1 and AT_2.
We note that:

- all sources, except S_6, belong to the catalogues

- existing sources verify quality criteria, except S_7 which violates freshness criteria.

Then, we can answer Q for all targets, but we also detect automatically that:

- T_1 is well described (all sources exist and verify the quality constraints)
- T_2 is not described because S_6 is not available
- T_3 is only described, because of the violation of quality criteria on S_7

So, T_2 and T_3 require actions to reduce impedance mismatches linked to requirements for S_6 and S_7 These actions could concern one or more aspects (theme, geography, time) of the required data sets

5.2 Step by step queries

The first iteration of the approach transforms Q into queries associated to each step:

Step1 - Q1 (Existence): Which are described targets on a geographic territory GT? For not described targets, reduce the mismatch impedance in the next iterations.

Step2 - Q2 (Quality) : Which are well described targets on GT? For badly described targets, reduce the mismatch impedance in the next iterations.

Step3 - Q3 (Contents): For well described targets on GT, give contents about DOP and departments in GT.

We illustrate classification issues for steps 1 and 2, and integration aspects for step 3.

5.2.1 Classification of targets (Steps 1 and 2)

In this first iteration, we check for correspondences in formula (1) (§4.1), which are identities. Hence, rids(T) ⊇ ouds(T) and Step1 limits to check for for existence, in the catalogues, of sources required by the targets.

Step1 - Q1 requires a sequence of concepts:

- *AvailableSource* ≡ *Source* ⊓ ∃ *belongs.Catalogue*
- *MissingSource* ≡ *Source* ⊓ ¬ *AvailableSource*
- *TerritoryGT* ≡ *Territory* ⊓ ∃ contains-.{GT})
- SourceGT ≡ Source ⊓ ∃ covers.TerritoryGT
- AvailableSourceGT ≡ AvailableSource ⊓ SourceGT
- DescribedTargetGT ≡ Target ⊓ ∃ manages.(Risk ⊓ ∃ concerns.{GT}) ⊓ ∃ requires.(AvailableSourceGT ⊓ ¬SourceGT)

As result, DescribedTargetGT = {T1, T3}, because, for all administrative territories AT in GT, each S in rids(T3) and in rids(T1) are available. T2 doesn't belong in this class since S_6 is not available.

Step2 - Q2 requires a sequence of concepts:

- *QualifiedSource* ≡ *AvailableSource* ⊓ ∀ *satisfies.RespectedCriteriaBound*
- *NotQualifiedSource* ≡ *AvailableSource* ⊓ Å*QualifiedSource*
- *QualifiedSourceGT* ≡ *QualifiedSource* ⊓ *SourceGT*
- *WellDescribedTargetGT* ≡ *DescribedTargetGT* ⊓ ∀†*requires.(QualifiedSourceGT* ⊔ ¬ *SourceGT)*

As result, *WellDescribedTargetGT*$_1$ = {T_1}, because each S in rids(T_3) is available and verify quality criteria, for all AT in GT. T_3 doesn't belong to this category since data source S_7 violates freshness criteria.

5.2.2 Queries about contents (Step3)

This query is valid for each target. We give general lines of an integration approach (step 3).

Meta data and Data sources Structures

Metadata elements are the same for all sources. They are described using XML Schema, as below:

```
<xs:element name="MetaData">
<xs:complexType>
<xs:all>
    <xs:element name="SId" type="xs:anyURI"/>
    <xs:element name="Object" type="xs:string"/>
    <xs:element name="PubDate" type="xs:date"/>
    <xs:element name="Extension" type="xs:string"/>
    <xs:element name="Coverage" type="xs:string"/>
</xs:all>
</xs:complexType>
</xs:element>
```

where *xs* refers to the XML Schema namespace.

For instance, XML Schema description for a local data source S_3 (related to the DependentOlderPeople of figure 6) is

```
<xs:element name="LOP" maxOccurs="unbounded">
    <xs:complexType>
    <xs:sequence>
    <xs:element name="PNum" type="xs:integer"/>
    <xs:element name="PName" type="xs:string"/>
    </xs:sequence>
    </xs:complexType>
</xs:element>
```

Mappings

The correspondences between the local data sources schemas and the global schema, according to LAV approach, are expressed by a set of mapping rules. Each rule associates path in the source schema, using XPath, with a conceptual path in the global schema. For example, rules below map paths in the source S_3 (augmented with metadata) to paths in the global schema (Fig. 6).

R1 : http://www.pa01.fr/S3.xml/Source3 as *u1* → Department
R2 : *u1*/MetaData/Coverage as *u2* → DeptNum
R3 : *u1*/LOP as *u3* → Contains
R4: *u3*/PNum as *u4* → OPNum
R5 : *u3*/PName as *u5* → OPName

Algorithms

With these mapping rules, we formulate the Step3-Q3 query:

```
select    d, count(g)
from      GeoTerritory a,
          a.Includes b,
          a.TId c,
          b.DeptNum d,
          b.Contains f,
          f.OPNum g
where     c="GT1"
```

This query must be decomposed since the data set returns only partial answers for it: it gives (i) a prefix query $Q3_1$, to find (from S_8) administrative territories included in the

geographic territory GT and (ii) a suffix query $Q3_2$, to find dependent older people on two departments (from S_3 and S_4)

$Q3_1 \rightarrow S_8$	$Q3_2 \rightarrow S_3, S_4$
select d from GeoTerritory a, a.Includes b, a.TId c, b.DeptNum d where c="GT"	Select d, count(g) From Department b, b.DeptNum d, b.Contains f, f.OPNum g

Rewriting the local queries $Q3_1$ and $Q3_2$, in XQuery language, gives respectively:

For $a in doc("S9")/Source9/LGT, $b in $a/LDept, $c in $a/TId, $d in $b/DNum where $c="GT1" return $d	for $b in doc ("S3")/Source3, $d in $b/MetaData/Coverage return {$d, {for $f in $b/LOP, $g in $f/PNum, return count($g) }}

The final results of the Step3-Q3 are obtained by merging the results of these last queries.

6. CONCLUSION

In this paper we addressed impedance mismatch problems, occurring when two systems are plugged for business activities requiring data sets from heterogeneous systems, within autonomous organizations. We introduced impedance mismatch metaphor in geographic information fusion and we proposed an interoperability approach. In this approach, the impedance mismatch has been broken down in three subclasses, related to existence, quality and contents aspects which drive the design of interoperability architecture. Each step classifies target systems with respect to the requirements about the sources. So we could decide if and how reduce impedance mismatch. The quality aware interoperability is a multidimensional optimization problem. So, we aim to investigate how to build preference relations (partial order or pre-order) to have a guess on best compromises to reduce the impedance mismatch.

REFERENCES

Amann, B., Beeri, C., Fundulaki, I., Scholl, M, 2002. Ontology-based integration of xml web resources. Springer, *LNCS,* 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, October 1st, 2001, Toronto, Canada. LNI series, vol. 7, pp. 7-11.

Baader, F., Horrocks, I., Sattler, U, 2005. Description Logics as Ontology Languages for the Semantic Web. Springer, *LNAI,* 2605, pp. 228–248.

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

Castro, J., Kolp, M., Mylopoulos, J, 2002. Towards requirements-driven information systems engineering: the Tropos project. *Information Systems*, 27(6), pp. 365-389.

Gruber T.R, 1993. *Knowledge Acquisition, chap.: A translation approach to portable ontology specifications*. Academic Press Ltd, London, pp. 199-220.

Haarslev, V., Möller, R., 2001. RACER System Description. Springer, *LNCS,* 2083, pp. 701-705.

Knublauch, H., Fergerson, R.W., Noy, N.F., Musen, M.A, 2004. The Protégé OWL Plugin: An Open Development Environment for Semantic Web Applications. Springer, *LNCS* 3298, pp. 229-243.

Lenzerini, M. 2002. Data integration: A theoretical perspective. In: Proc. of PODS 2002. (2002), pp. 233–246

Parent, C., Spaccapietra, S., 2000. "Database Integration: The Key to Data Interoperability", in Papazoglou, Spaccapietra & Tari (Eds.) *Advances in Object-Oriented Data Modeling*, MIT Press, 2000, pp. 221-254.

Smith, M.K., Welty, C., McGuinness, D, 2004. OWL Web Ontology Language Guide. http://www.w3.org/TR/owl-guide (accessed 23 Jan. 2007)

Visser U. 2004, *Intelligent Information Integration for the Semantic Web*. Springer, Berlin, LNAI 3159, pp. 13-34.

ACKNOWLEDGEMENTS

This work has been partially supported by French Regions PACA and Midi-Pyrénées.