Using metadata to improve spatial dataset quality during updates

Christelle Pierkot

Using metadata to improve spatial dataset quality during updates

Christelle Pierkot
LIRMM, D'OC Team
161 rue ADA
34392 Montpellier Cedex 5, France
christelle.pierkot@lirmm.fr

1. Introduction

Nowadays, several different actors are involved in acquiring, distributing and updating spatial data. Such a situation naturally leads to a multiplication of heterogeneous data of different types, formats, abstraction levels and qualities. Blind integration of this data can undermine the consistency of the database and degrade the dataset's quality.

Within this broad issue, we limit our scope to the updating of geographic data by different actors spread out over a communications network – with some data liable to change while the actor is disconnected from the network. Data is replicated on different sites and there is no server holding centralized information. Moreover, updating is conducted simultaneously by the different actors, sometimes in disconnected mode. The evolutions are therefore not necessarily relevant to a particular user, sometimes contain errors, can be conflicting and consistency issues can arise during their integration into different datasets.

In this context, our objective is to propose solutions to allow coherent integration, as far as possible automatically, of spatial-data updates in a multi-master and asynchronous optimistic replication environment.
To achieve this, we propose a global integration strategy, based on standardized metadata, to provide solutions to improve the quality of a dataset during updates by multi-source evolutions. We will have to, on the one hand, deal with the relevance of the proposed evolutions and, on the other, ensure dataset consistency commensurate with the final user's requirements.

This strategy depends on the prior establishment of a spatial data infrastructure in which a communications network has been defined, the users and their different roles are known, and updates can be exchanged using a common strategy [Pierkot et al., 2006]. In this infrastructure, a metadata model that allows formal relationships between data, actors and the evolutions can be set up.

This paper is structured as follows: We start by defining consistency and quality in spatial databases. Then, we present the different parts of the metadata model which we have defined in the infrastructure and we cover in some detail the quality metadata relating to the evolutions. Subsequently, we present the integration strategy and, in particular, the modules for verifying the relevance and consistency. Then, we analyze the results obtained for the consistency-checking process. Finally, we conclude and provide some perspectives for this work.

2. Quality and consistency of spatial databases

In the domain of geographic information, the concept of consistency is tightly linked to the dataset quality and can be impacted by numerous sources of errors. In the context of updating, it is the capacity to satisfy a set of criteria during the integration and/or the propagation of new data and evolutions.

[David and Fasquel, 1997] distinguish between two types of spatial-data quality:

- Internal quality is the set of properties and characteristics of a product or service which confers on it the ability to satisfy the specifications of the content of this product or service. It is measured by the difference between the data which should have been produced and the data which has actually been produced. It is linked to specifications (and, in particular, to errors that can be committed during data production) and is evaluated in terms of the producer.
- External quality is defined as the suitability of the specifications to the user's requirements. It is measured by the difference between data wished for by the user and the data actually produced. It is linked to the users' needs and thus varies from one user to the next.

Quality evaluation thus comes down to verifying the conformity between data of the database and the data considered correct (from the producer's or user's point of view).

[Bel-Hadj-Ali, 2001] emphasizes that geographic-data quality is so complex that it is impossible to use a single global measurement; one has to rely on several components for determining it. Consequently, several criteria have been defined to be able to define a spatial dataset's internal quality such as the lineage, the geometric accuracy, the semantic accuracy, completeness, actuality, logical consistency and semantic consistency [Moellering, 1987], [Clarke and Clark, 1995], [Drummond, 1995], [Goodchild et al., 1992], [Bicking, 1994], [Brassel et al., 1995], [Guptill, 1995], [Salgé, 1995]. All these criteria have been widely tested and are nowadays used in several standardization works [CEN, 1998], [FGDC, 1998], [ISO19115, 2003]. They form a basis for evaluating a geographic dataset's quality.

Data relevance is a concept that we can link to the concept of fitness for use and, in particular, to the external quality [Dassonville et al., 2002]. In these last few years, a lot of research work has been done for an improved taking into account of the external quality [Bruin et al., 2001], [ReV !Gis, 2004] [Vasseur, 2004], [Devillers, 2004], [Devillers and Jeansoulin, 2005].
Two broad approaches have been proposed, one is based on the evaluation of the risk incurred by using unsuitable data [Agumya and Hunter, 1998], [Bruin et al., 2001], the other on the use of metadata to analyze the similarity between data produced and users' needs [Frank, 1998], [Hunter, 2001], [Devillers, 2004], [ReV !Gis, 2004], [Vasseur, 2004]. These methods differ in the way they are developed but both lead to the appraisal of the data quality with respect to the use that it will be put to. They thus allow better targeted interpretations and less risky decision making.

Evaluating the quality of a geographic dataset is not a task to be undertaken lightly; a database of poor quality can lead to numerous errors that can imperil data consistency. Furthermore, different work on geographic-data quality ([David and Fasquel, 1997], [Vasseur et al., 2005], [Devillers and Jeansoulin, 2005], [Harding, 2005]) has repeatedly shown the necessity of considering the user's point of view so that the data is fit for his use. This is all the more true in a context of decision making or where the data will serve as basis for the actions of various actors.

In our context, a user could integrate several sets of evolutions originating from multiple sources. This implies that the updates are not necessarily all relevant and that these updates could be in conflict or lead to inconsistencies with his data.
For proper integration of these updates, their quality has to be evaluated taking the user's needs into account. Nevertheless, to the best of our knowledge, no study has been conducted on the quality of evolutions; all research into updates has always assumed that the evolutions are relevant to the user. Our study thus has to answer this question – which has so far remained unasked in the literature: How to evaluate the quality of updates so as to retain only those that are consistent and relevant to the user?

3. Metadata model

One solution to evaluate data quality is to use metadata. In fact, from the numerous pieces of information that we can find in metadata, there are those that relate to the resource's quality. However, the quality metadata generally indicates the products' internal quality. In fact, metadata has been developed from the producer's point of view and contains little information for the user to judge the uses it can be put to [Bucher, 2002]. This, in particular, poses the problem of the relevance of data for the final user, especially as far as fitness for use is concerned.

Yet, in our study, metadata has to, on the one hand, filter out irrelevant evolutions and, on the other, ensure the consistency of the dataset, all the while taking the user's expectations into account. We thus have to consider both types of quality and have to add criteria to metadata to be able to define the suitability of the evolutions to the user's requirements.

It also seems relevant to us to define the quality at different granularities, i.e., not only for the entire evolution set but also for the evolutions themselves. In fact, quality evaluation at the level of the entire set allows us to quickly filter out the collections which are irrelevant to the user, whereas quality evaluation at the level of the evolutions helps in reconciling two conflicting evolutions.

3.1. ISO 19115: Metadata for data

To supply shared and consistent knowledge of data between different communities, we have to standardize the metadata [Luzet, 1998], [Gunther and Voisard, 1997], [Spéry and Libourel, 1998]. Standards for describing metadata have been developed and provide a common base to the users [CEN, 1998], [FGDC, 1998], [ISO19115, 2003]. The use of normalized metadata in the infrastructure thus promotes interoperability between different actors and systems.

The standard that draws the most attention nowadays is [ISO19115, 2003]. The ISO 19115 standard defines a large set of metadata elements so that it can be used by several different types of users. However, a community of a given set of geographic-data users normally uses only a part of the different metadata elements defined in the standard and, in spite of the wide variety of elements, often needs to add elements not specified in the standard. ISO 19115 allows this thanks to the possibility of defining community profiles. A profile thus allows use of the standard restricted to a subset of mandatory elements and also extended by the addition of missing sections, entities and elements.

3.2. MUMSDI: Metadata for evolutions

ISO 19115 has been designed to provide information on the use and exchange of datasets. The evolutions sets and basic evolutions (creation, deletion or modification) are not included in the standard. It is therefore not currently possible to provide metadata relating to a set of evolutions that we would want to provide to a user who already possesses the reference dataset.
Consequently, metadata elements have to be added to take these new requirements into account. We thus propose to extend ISO 19115 to encompass evolutions. Towards this end, we have created a metadata profile which we call MUMSDI (Metadata for Updating a Military Spatial Data Infrastructure). It is specifically designed to manage the evolutions of military data[1].

[1] Our study's specific context is that of a military mission wherein the actors are spread out over different sites, use spatial data and exchange update information. MUMSDI was thus specified for the military community.

Quality metadata in the MUMSDI profile

Quality metadata defined by ISO 19115 are used to describe the data quality from the producer's point of view, which is not sufficient for our context. We therefore propose to add quality elements to take the user's point of view into account. We also restrict some elements of the standard which are not useful to us.
Figure 1 gives an overview of quality information in our MUMSDI profile. The first modification we have made concerns the cardinality of the *dataQualityInfo* role which we have restricted because we think that to evaluate the evolutions correctly, quality information has to be made mandatory. In the same way, we have modified the cardinalities of the *lineage* roles and the report of the *dataQualityInfo* section to make them also mandatory in the profile.
We have also modified the scope of the quality elements (attribute *level* of the *DQ_Scope* class). This information is provided in the standard in a list (*MD_ScopeCode*) which contains some elements that allow the description of the level to which the information will apply. We have modified this list in the MUMSDI profile to take into account only those elements that relate to the quality information of the evolutions.
We have also deleted a certain number of attributes that we did not consider useful in an updating context. Our reason for doing so derives from the observation that the main difficulty in the use of metadata comes from the very large number of metadata elements to fill in. We therefore think that fewer the unnecessary elements in the MUMSID profile, easier will be their entry and better their interpretation.

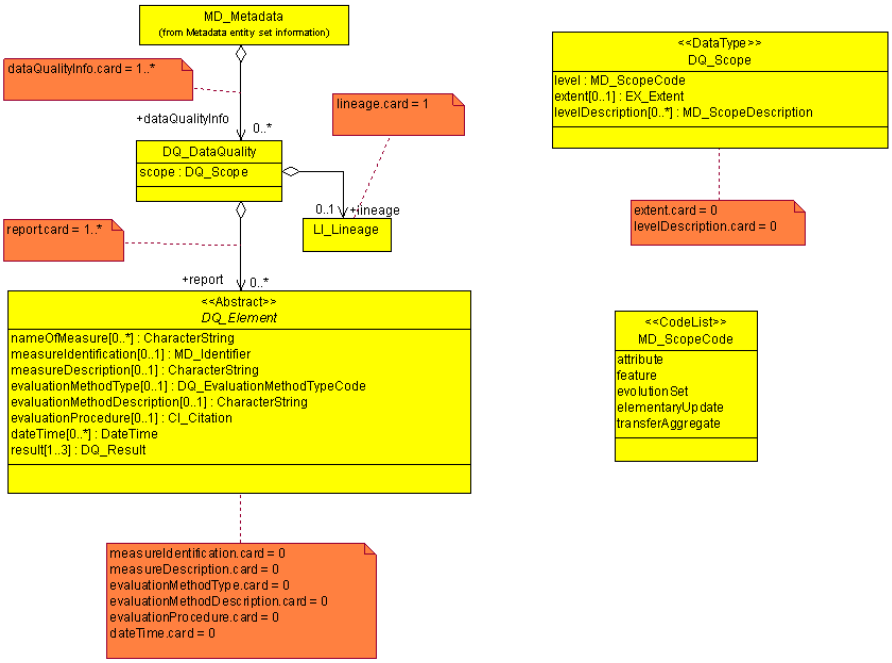

Figure 1: Quality information in the MUMSDI profile

All the information on the quality is available thanks to a set of quality measurements accessible via the *DQElement* class of which only the name (attribute *nameOfMeasure*) and the result (attribute *result*) of quality measurements are specified in the MUMSDI profile (see figure 2).
We have deactivated the sub-classes of *DQElement* that are not relevant to our context (shown with white backgrounds in the figure). We have then added a sub-class *MU_Usability* to be able to judge the evolutions' ability to satisfy the usage the user may put them to (shown with orange backgrounds in figure 4). This element indicates the degree of the evolutions' conformity with the usage that a given user type may put them to. This class has two attributes to indicate the type of user concerned by the quality measurement (attribute *finalUserRole*) and the site where the user is located (attribute *finalUserLocation*).

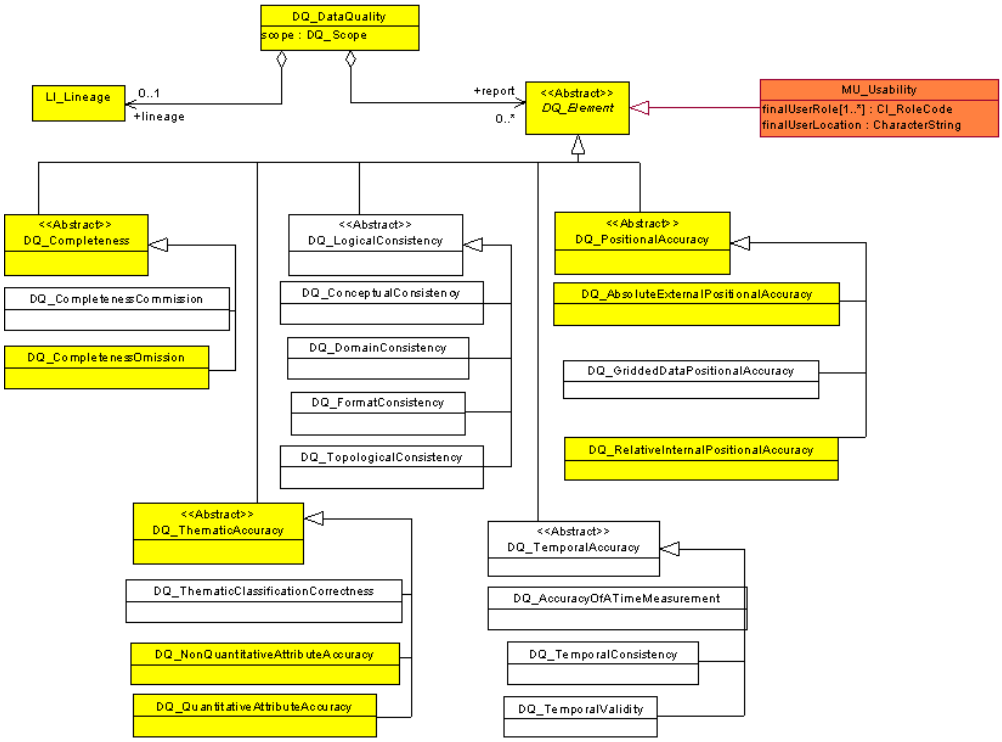

Figure 2: Quality elements in the MUMSDI profile

Finally, the expression of the result of the quality in MUMSDI is shown in figure 3. The results available in the standard (*DQ_QuantitativeResult* and *DQ_ConformanceResult*) can be easily supplied by a producer but with difficulty, or not at all, by the other users. In fact, some users do not even have the technical means to evaluate the evolutions accurately and have to judge by themselves the quality of the updates they have made. We therefore think that the quantitative results are not sufficient to describe the information on the evolutions' quality and we have thus added qualitative elements to be able to judge the evolutions' quality in a more flexible manner. To this end, a class *MU_QualitativeResult* class has been created. It allows us, on the one hand, to quickly see whether non-spatial information attached to the basic evolutions has been correctly documented (attribute *documentation*) and, on the other, to see what are the types of errors that the updates may contain (attribute *errorType* of type *MUerror)*.

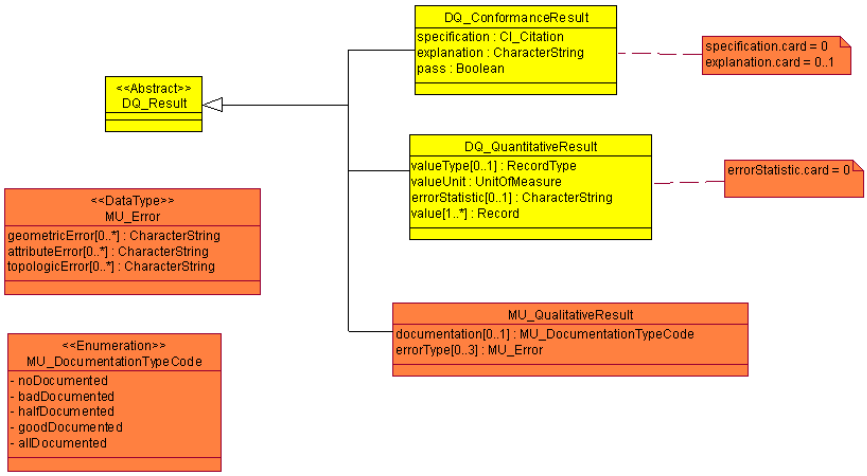

Figure 3: Expression of the quality result in the MUMSDI profile

3.3. Metadata for actors

Infrastructure users do not all have the same requirements as far as evolutions are concerned. These requirements depend on several factors such as the user's role, the site where he is located, etc.

Indeed, a certain number of consistency constraints linked to the dataset used and to the technical methods available have to be defined for each actor. In fact, an infrastructure actor has a dataset that is, admittedly, derived from a unique reference set, but which may have been transformed so that it can be used by the user's system.
This entire information (needs and constraints) constitutes the actors' metadata. The needs are not fixed and can change over time. On the other hand, the constraints are imposed at the beginning and cannot be changed.

We have specified several criteria defining the entirety of user needs, such as the maximum spatial extent, the minimum occurrence date, the different thematic layers and the type of evolutions required. Also specified are the minimum geometric and semantic accuracies, as well as the reliability.
The list of consistency constraints has, on its part, been defined by spatial constraints (geometric accuracy, minimum resolution, etc.), semantic constraints (quantitative and qualitative accuracy) and context constraints (type of sources allowed, occurrence date and maximum extent of the dataset, etc.).

4. Integration strategy for multi-source updates

4.1. General approach

Figure 4 shows the general approach for the evolution integration strategy. The integration strategy consists of three stages and applies to the dataset of an infrastructure user for whom evolution sets originating from multiple sources are destined. The execution of the three stages of the integration strategy leads to the updating of the user's dataset.

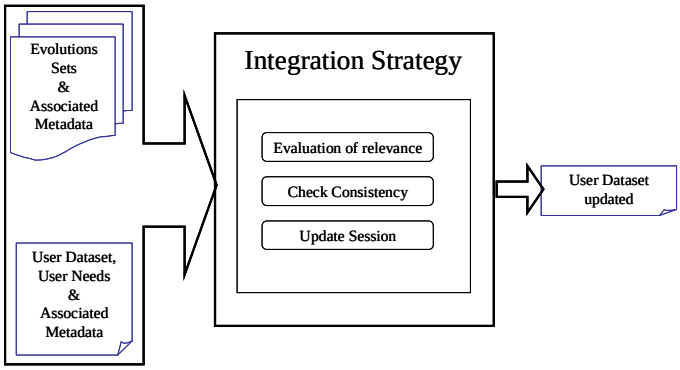

Figure 4: Strategy of integrating updates into a user dataset

The first stage consists of the evaluation of the relevance of the evolutions. Here, we filter out those evolutions that are not useful to the user. To do this, we use metadata associated with evolutions and metadata associated with the users' needs. At the end of this stage, the evolution set proposed for the integration contains only those evolutions that are suitable for the usage the user wants to put them to.
The second stage concerns the checking of the consistency. Its aim is to detect and process any possible conflicts that can arise from the fact that the evolution sets originate from multiple sources. This stage which executes in a two-step process consists, firstly, of consistency verification followed by a phase to reconcile conflicting data. We use here the structure of the evolutions and geometric processing to detect possible conflicts. We then use metadata associated with the evolutions, with the data and with the user's needs to define the reconciliation methods. At the end of this stage, we obtain a set of relevant and non-conflicting evolutions that we can integrate into the user's dataset.
The third and final stage of the strategy is closely linked to the second and permits the integration of the evolutions already processed into the user's dataset.

4.2. Relevance of the evolutions

The sources used in the infrastructure originate from the same reference dataset but have evolved (updates and transformations) depending on the specific requirements of the users and the technical methods they have at their disposal. In addition, the evolutions could have been entered in different ways, under different conditions and at different locations. They are thus heterogeneous, of varying qualities and do not necessarily relate to the same geographical coverage zone.
For all these reasons, an evolution set originating from an infrastructure actor does not, all by itself, fulfil the final user's needs. In fact, for a given particular requirement, an evolution set usually contains more information than really needed by the user. Finally, an infrastructure actor has to retrieve several evolution sets originating from different sources to completely satisfy his requirements but has to exclude evolutions which are definitely not relevant to the usage he wants to put them to.
The goal of a consistency checking process is therefore to filter out, from all the evolution sets proposed, the evolutions that are not relevant to the user and which would needlessly risk deteriorating his dataset's external quality. The solution we advance to do this filtering is based on the metadata associated with evolutions and to the users' needs. The idea is to proceed by analysis and comparison of metadata to determine if the evolutions are relevant or not to the user. To do this, we use the work by [Jeansoulin and Wilson, 2002] and [Vasseur, 2004] as a basis. They evaluated a dataset's external quality as a function of the use it could be done by a user community. We provide here the general method of doing so without going into each module's details.

Three stages are necessary to evaluate the external quality:

- Creation of problem and product ontologies in a common reference base [Jeansoulin and Wilson, 2002]
- Definition of expected and internal quality matrices [Vasseur, 2004]
- Evaluation of suitability to needs using utility calculations [Vasseur, 2004]

Product and problem ontologies provide a real world view from the points of view, respectively, of data producers and of the final user, by formalizing the data characteristics and the user needs. In our study, the product ontology is specified by the evolutions proposed for integration and the problem ontology by the user's needs. The metadata associated with evolutions and to user needs is used to define these ontologies.
From the problem and product ontologies, matrices of internal quality and expected quality are created in a common reference base. In our study, the matrix of internal quality corresponds to the evolutions' quality and the matrix of expected quality to the user needs. Since the metadata associated with evolutions and with actors is normalized, they can be used without problems in the different matrices.
The utility calculation then evaluates the data in terms of user expectations and therefore provides an estimate of the data adequacy to different needs. In our study, the utility calculation informs us if the evolutions' quality is suitable for the user's needs. If not, the evolution is excluded from the evolution set.

4.3. Consistency checking

We restrict ourselves to studying the data consistency and not that of models or schemas which are assumed handled within the infrastructure.

Different consistency levels are required depending on the needs and roles of the actors in the infrastructure. In fact, heavy producers who have to provide reference information and prepare future datasets have access to system resources for updating and sharing their data. These actors' objective is therefore to retrieve quality information to be able to obtain a reliable and accurate dataset. This means that care has to be taken to limit as far as possible the inconsistencies that can be produced while integrating evolutions into their datasets. The consistency level should therefore be high because we focus on the quality and consistency of data rather than the quantity of information.
Users, on the other hand, use data for purposes of decision making. Their primary objective is to rapidly react and take initiatives as soon as required. The goal is here to retrieve a maximum amount of information as fulfilling a specific need irrespective of its quality. We therefore accept that inconsistencies will arise when evolutions are integrated into the dataset. The consistency level desired can therefore be considered weak because we prefer the data quantity over data quality and consistency.
Light producers have to supply evolutions to the users and also transmit upstream the information retrieved to the heavy producers. The light producers have simplified system resources which nevertheless allow them to update their data and to then share it with the other actors. The goal for these actors is to retrieve a maximum amount of information and to transmit to the other actors (heavy producers and the users), based on their needs, while preserving a certain quality level. We try to limit but accept some inconsistencies during the integration of evolutions into the dataset. The consistency level desired here is intermediate between the other two cases.

We therefore propose a consistency checking protocol which, on the one hand, allows the detection of conflicts that could lead to inconsistencies in the dataset during integration and, on the other, offers reconciliation routines for conflicting evolutions as a function of the desired consistency level.
We handle concurrency by dividing the conflicts into 3 distinct types (update conflicts, topological conflicts and conflicts of creation). We then apply methods to detect the conflicts (semantic and geometric matching, tests on the spatial relationships between the objects).

The reconciliation protocol starts when one or more conflicts are detected during the stage of concurrency checking. This process's goal is to offer the most suitable solution for resolving the conflict(s). The result depends on the desired consistency level and the desired balance between quality and quantity.
The reconciliation protocol we propose is original in its use of available metadata to offer a result commensurate with the actor's expectations.
In fact, the metadata included with the evolutions and data provides information such as on its quality or on its origin. This information allows the process to compare the items with the actor's expectations (constraints or needs).

The reconciliation process consists of several stages:

- First, a comparison of metadata associated with the items in conflict is made with the actors' metadata and a calculation for measuring the quality of each of the items' characteristics (geometry, attributes, reliability, etc.) is made. This part is based on the work done on the utility calculation by [Grum and Vasseur, 2004] and [Frank et al., 2004].
- Then, we calculate a measurement of overall quality for each item in conflict to be able to obtain a result that is a function of the final user's expectations and of the desired consistency level.
- Finally, a comparison of the overall quality measurements of the items in conflict is made and the item found to be the most relevant is selected for future integration. This part can be executed automatically, semi-automatically or interactively.

Comparison of metadata and computation of quality measurements

The first action in the reconciliation process is therefore to compare metadata. A comparison is only possible if the two metadata are standardized and if correspondences are explicitly established. Metadata sets that we have specified in the infrastructure are in formats that are very close because they have been specified conforming to the requirements of the ISO 19115 standard. We can thus establish correspondences between the different items of each metadata set.

In addition, we classify the metadata items according to the type of information they carry. We distinguish five principal characteristics for measuring the quality of a resource with respect to a user's expectations: geometric characteristics (accuracy and resolution), semantic characteristics (quantitative accuracy and qualitative accuracy), lineage characteristics (sources and processes of the construction of resources), reliability characteristics (error type and the trust accorded to the actors) and other, more general, characteristics such as timeliness, completeness and extent of the set containing the resources. We then calculate the quality measurements for each characteristic taken individually.
We calculate the quality measurement of each characteristic using a distance calculation which measures the difference between the quality that the actor desires and the actual quality of the resource. This calculation is possible for the majority of characteristics because quality values are numeric values or percentages. For other types of values, we have to create a metric associated with the items.

Computation of the measurement of overall quality

At the end of this second stage, we wish to obtain a measurement of the overall quality which corresponds to the quality measurement of the resource in conflict with respect to the expectations of the user who will use it. This measurement depends on the desired consistency level and will, therefore, differ from actor to actor (even if every item's quality measurements are identical). This stage is divided into three phases:

- First, we normalize the quality measurements of each characteristic obtained from the previous stage so that all the measurements are in the same unit and are, thus, comparable. The normalization should lead to values between -1 and 1.
- Then, we allocate a weight to each of these individual quality measurements so as to take into account the consistency level and expectations of the user who will use the data. The value of the weight depends on the desired consistency level and the actor's requirements. It is thus contextual in nature.
- Finally, we aggregate the quality measurements of each characteristic to obtain an overall measurement. To do this, we take the mean value. This provides us with an overall idea of the resource's external quality with respect to the user's needs and to the expected consistency level.

Selection of the relevant item

The third and final stage of the reconciliation process allows an item to be chosen from conflicting ones.
The final selection can take one of three forms:

- The evolution corresponds best to the actor's expectations; it is selected.
- The data corresponds best to the actor's expectations; it is retained and the evolution is deleted.
- Neither of the two resources, considered globally, entirely satisfies the actor's expectations but some characteristics taken separately may be relevant. In such a case, a new evolution can be built with these characteristics. This happens when one part of the evolution and one part of the other resource provide a better solution if they are combined rather than if either is taken individually.

This third stage can be executed automatically, semi-automatically or interactively depending on the results obtained. It is always preferable to automate as far as possible the reconciliation protocol so that the results obtained may be harmonized. This is possible when there is sufficient information (in quantity and of quality) available with the evolutions and the data for the quality measurements to be calculated accurately. However, when a new evolution is proposed by aggregation of the relevant characteristics of two resources, it is sometimes difficult to choose between several solutions, all of which may appear acceptable. In such cases, the protocol proposes a list of choices to the user for his selection. As a last recourse, when no satisfying solution has been found, the process defers to the user's choice of a solution he deems best. No doubt this situation has to be avoided as far as possible because the actors, even if they are handling a dataset with identical constraints, have different viewpoints of the real world and their analyses of the same situation can lead to diverging interpretations, which is obviously not desirable.

5. Implementation and evaluation

The simulation context we have selected is that of an actor with a dataset derived from a reference set but which he has changed to suit his environment and his requirements. Evolutions originating from other infrastructure actors are proposed to our reference actor for possible integration.
We have focused on the consistency checking part of the integration strategy. In particular, we have wanted to show that the case is convincing for the use of metadata for the reconciliation of conflicting data.
We have considered the situation where some data and an evolution have been declared as conflicting and have to be processed. Several metadata elements describing the quality of the evolutions and of the data (geometric accuracy, completeness, reliability, actuality date, etc.) are supplied with the data and the evolution.

To be as close as possible to our study's contextual reality and to estimate as best as possible the reconciliation results, we have simulated the following two cases: in the first, the reference actor has requirements that can change over time. In the second case, the reference actor is a producer whose requirements were fixed and cannot change. Proceeding in this manner, we can illustrate the reconciliation process as a function of each actor's own constraints and as a function of the consistency level inherent to their roles within the infrastructure.

The first stage in the reconciliation process consists of retrieving the information contained in the metadata, then of calculating a quality measurement for each item of information. The quality measurement is calculated as a function of the reference actor's needs and of the desired consistency level.
For each characteristic, the process calculates a quality measurement, and then normalizes this measurement to obtain a result between -1 and 1. Closer the quality measurement is to -1, better is the external quality.
The table below shows the result of the normalized quality measurement calculation that we have obtained for a few characteristics attached to the evolutions (geometric accuracy, semantic accuracy, reliability, etc.).

Characteristic	Actual value	Desired Value	Normalized quality measurement
Geometric accuracy	150m < x < 200m	≤ 500m	-0.43
Semantic accuracy	50%	≥ 50%	0
Reliability	50%	≥ 40%	-0.1
Completeness	42%	≥ 10%	-0.32
Actuality	15/03/08	≥ 01/02/08	-0.4

Tab 1 : normalized quality measurement for characteristics attached to the evolution

The process then calculates the overall quality measurement for each conflicting data.
For this simulation, we have considered two types of reference actors: a heavy producer and a light producer. The heavy producer's consistency level is higher than that of the light producer. In addition, the heavy producer prefers obtaining reliable data with a geometric accuracy close to his requirements rather than voluminous data which is of suspect quality. The weight allocated to the geometric accuracy and to the reliability will be higher than to the completeness. On the other hand, the light producer prefers to obtain recent information rapidly irrespective of its quality. The weight allocated to completeness and to actuality will be higher than that to reliability or geometric and semantic accuracies.
Finally, the process calculates the weighted mean to obtain an overall quality measurement for each of the conflicting data. With the values assumed for our simulation, we obtain the following results:

	Heavy Producer	Light Producer
Evolution	-0.2246	-0.3118
Data	-0.437	-0.17

Tab 2 : Overall qualities measurements

The analysis of results shows that we obtain overall quality measurements which differ depending on the actor's role in the infrastructure and therefore depending on the expected consistency level. In fact, we see in particular that the evolution will be better for the light producer whereas the data will more suitable for the heavy producer. The process will therefore have a different preference for conflicting data depending on the actor who will finally use the data after integration.
This result proves that it is possible to develop a reconciliation process which uses information present in metadata to offer a choice between conflicting data by considering, on the one hand, the needs and constraints of the actor – based on his role in the infrastructure – and, on the other, the desired consistency level for the different datasets which will be used.

6. Conclusions and perspectives

The solution we have proposed in this paper is to process evolutions originating from multiple sources relies on a spatial data infrastructure and an integration strategy for evolutions using normalized metadata. The main benefits of this method are:

- The formalization of a metadata model for managing quality evolutions, normalized and thus favouring interoperability.
- The taking of user needs into account to evaluate the suitability of the evolutions for the use they will be put to.
- The elimination of conflicts thanks to a reconciliation protocol using the information contained in the metadata.
- The integration of evolutions that are relevant and of those that will not lead to inconsistencies.
- At the end, we obtain an updated dataset whose quality has been preserved.

The improvements we could bring to this work relate to several aspects of the integration strategy. We believe that automatization of the filling in of metadata during the data-entry of evolutions could lead to an optimal number of information items usable by the reconciliation process. We could study the work of [Libourel, 2003] and, especially, the thesis of [Barde, 2005] which rely on a RDBMS to control the entry of certain metadata elements.
The development of an interface to manage the interactivity also seems to us to be worth considering for improving the integration strategy. We could, for example, draw inspiration from the work of [Devillers, 2004] and use quality indicators. This would help the actor in making the best choice when the reconciliation of conflicting data cannot be done automatically.
Finally, there are many general perspectives to this work since so many possibilities remain to be explored regarding the quality during the updating of distributed geographic databases. In fact, substantial work remains to be done on the fulfilling of user requirements; it would allow us to improve our results regarding the filtering of irrelevant evolutions and the reconciliation of conflicting data.

REFERENCES

[Agumya et Hunter, 1998] Agumya, A. & Hunter, G. (1998). Fitness for use: Reducing the impact of geographic information uncertainty. In URISA 98 Proceedings.

[Barde, 2005] Barde, J. (2005). Mutualisation de données et de connaissances pour la gestion intégrée des Zones Côtières. Application au projet SYSCOLAG. PhD thesis, University of Montpellier II, France

[Bel-Hadj-Ali, 2001] Bel-Hadj-Ali, A. (2001). Qualité géométrique des entités géographiques surfaciques : Application à l'appariement et définition d'une typologie des écarts géométriques. PhD thesis, University of Marne la Vallée, France.

[Bicking, 1994] Bicking, B. (1994). A Formal Approach to Automate Thematic Accuracy Checking for Cartographic Data Sets. PhD thesis, University of Maine, USA.

[Brassel et al., 1995] Brassel, K., Bucher, F., Stephan, E. et Vckovski, A. (1995). Elements of Spatial Data Quality, chapter Completeness. S.C.Guptill et J.L.Morisson. Oxford : Elsevier.

[Bruin et al., 2001] Bruin, S. D., Bregt, A. et de Ven, M. V. (2001). Assessing fitness for use: the expected value of spatial data sets. International Journal of Geographical Information Science,15(5):457-471.

[Bucher, 2002] Bucher, B. (2002). L'aide à l'accès à l'information géographique : un environnement de conception coopérative d'utilisations de données géographiques. PhD thesis, University Paris 6, France.

[CEN, 1998] CEN (1998). Geographic Information European Prestandards, Euronorme Voluntaire for Geographic Information -Data description- Metadata. European Committee for Standardization -CEN/TC287.

[Clarke et Clark, 1995] Clarke, D. et Clark, D. (1995). Elements of Spatial Data Quality, chapter Lineage. S.C.Guptill et J.L.Morisson. Oxford : Elsevier.

[Dassonville et al., 2002] Dassonville, L., Vauglin, F., Jakobsson, A. et Luzet, C. (2002). Spatial Data Quality, chapter Quality Management, Data Quality and Users, Metadata for Geographical Information. Taylor & Francis.

[David et Fasquel, 1997] David, B. et Fasquel, P. (1997). Qualité d'une base de données géographique : concepts et terminologie. Rapport technique, IGN.Bulletin d'information n°67.

[Devillers, 2004] Devillers, R. (2004). Conception d'un système multidimensionnel d'information sur la qualité des données géospatiales, Phd thesis, University of Laval, Québec and University of Marnes la Vallée, France.

[Devillers et Jeansoulin, 2005] Devillers, R. et Jeansoulin, R. (2005). Qualité de l'information géographique. Traités en Information Géographique et Aménagement du Territoire, IGAT, Hermès Sciences, Lavoisier. ISBN 2-7462-1097-5.

[Drummond, 1995] Drummond, J. (1995). Elements of Spatial Data Quality, chapter Positional Accuracy. S.C.Guptill and J.L.Morisson. Oxford : Elsevier.

[FGDC, 1998] FGDC (1998). Content Standard for Digital Geospatial Metadata, version 2.0. Document FGDC-SDT-001-1998. Federal Geographic Data Commitee, Metadata Ad Hoc Working Group.

[Frank, 1998] Frank, A. (1998). Metamodels for data quality Description, pages 15-29. Data Quality in Geographic Information. From Error to Uncertainty. Editions Hermes.

[Frank et al., 2004] Frank, A., Grum, E. et Vasseur, B. (2004). Procedure to select the best dataset for a task. In International Conference on Geographic Information Science.

[Goodchild et al., 1992] Goodchild, M., Guoquing, Y. et Y.Shiren (1992). Development and test of an error model for categorical data. International Journal of Geographical Information Systems, 6(2):87-107.

[Grum et Vasseur, 2004] Grum, E. et Vasseur, B. (2004). How to select the best dataset for a task? In Fourth International Symposium on Spatial Data Quality, ISSDQ'04, pages 197-206.

[Gunther et Voisard, 1997] Gunther, O. et Voisard, A. (1997). Metadata in Geographical and Environmental Data Management. Managing Mutimedia Data : Using Metadata to Integrate and Apply Digital Data.

[Guptill, 1995] Guptill, S. (1995). Elements of Spatial Data Quality, chapter Temporal Information. S.C.Guptill et J.L.Morisson. Oxford : Elsevier.

[Harding, 2005] Harding, J. (2005). Qualité des données vectorielles : perspective d'un producteur de données, chapter 10. Qualité de l'information géographique, Traités IGAT, Hermès Sciences, Lavoisier. ISBN 2-7462-1097-5.

[Hunter, 2001] Hunter, G. (2001). Spatial data quality revisited. In GeoInfo.

[ISO19115, 2003] ISO19115 (2003). Geographic Information : Metadata. ISO/TC211.

[Jeansoulin, 1997] Jeansoulin, R. (1997). Data Quality in Geographic Information : From error to Uncertainty, chapter sing Spatial Constraints as Redundancy Information to Improve Geographical Knowledge. M.Goodchild and R.Jeansoulin. Hermes.

[Jeansoulin et Wilson, 2002] Jeansoulin, R. et Wilson (2002). Model-based semantics for ontologies of geographic information. In Gisciences.

[Libourel, 2003] Libourel, T. (2003). Autour de la conception de systèmes complexes. Modélisation, évolution, infrastructures. Mémoire d'HDR.

[Luzet, 1998] Luzet, C. (1998). Megrin's gddd, moving to distributed metadata. In EOGEO98.

[Moellering, 1987] Moellering, H. (1987). A draft proposed standard for digital cartographic data, national committee for digital cartographic standards. Technical report, American Congress on Surveying and Mapping.

[Pierkot et al., 2006] Pierkot, C., Mustiere, S., Ruas, A. et Hameurlain, H. (2006). Using metadata to help the integration of several multi-sources set of updates. In 9th GSDI Conference, Santiago, Chile.

[ReV !Gis, 2004] ReV !Gis (2004). Revision of the uncertain geographic information. Technical report, MIT Labotary for Computer Sciences and RSA data Security. Projet n°IST-1999-14189, www.lsis.org/REVIGIS/Full/index.html.

[Salgé, 1995] Salgé, F. (1995). Elements of Spatial Data Quality, chapter Semantic Accuracy. S.C.Guptill et J.L.Morisson. Oxford : Elsevier.

[Spéry and Libourel, 1998] [Spéry et Libourel, 1998] Spéry, L. & Libourel, T. (1998). Vers une structuration des métadonnées. In Journées Cassini.

[Vasseur, 2004] Vasseur, B. (2004). Modélisation de l'information de qualité dans les applications Géographiques. Phd thesis, University of Aix-Marseille 1, France.

[Vasseur et al., 2005] Vasseur, B., Jeansoulin, R. et Devillers, R. (2005). Evaluation de la qualité externe de l'information géographique : une approche ontologique. Qualité de l'information géographique, Traités IGAT, Hermès Sciences, Lavoisier. ISBN 2-7462-1097-5.