GSDI-9 Conference Proceedings, 6-10 November 2006, Santiago, Chile

Using Metadata to Help the Integration of Several Multi-source Sets of Updates

Christelle Pierkot
EADS – IGN – IRIT
Université Paul Sabatier,
IRIT, équipe Pyramide
118, Route de Narbonne
31062 Toulouse cedex 9, France, pierkot@irit.fr

Sébastien Mustière and Anne Ruas
IGN – Laboratoire Cogit, France, {sebastien.mustiere, anne.ruas}@ign.fr

Abdelkader Hameurlain
Université Paul Sabatier – IRIT, hameur@irit.fr

Abstract

Today, spatial data are increasingly available on the web and users can update their datasets more easily. Different sets of updates result from diverse sources are furnished to the user, each containing updates acquired in different manners, with different quality and at different times.
A special context where the data and updates could come from different sources is a military mission. Indeed, the actors are distributed between different sites and one particularity is that they can be either a producer or a user of the data. They have their own dataset and can update them in several ways but must regularly supply their evolutions to the others actors in order to guarantee the success of the mission. Therefore, each actor receives many heterogeneous sets of updates and must integrate them in their own dataset in accordance with their needs.
In this context, the user receives several set of heterogeneous updates which can have different quality, which can contain errors due to the manner they were acquired and they have to integrate them in their personal dataset.
Thus, all the evolutions are not necessarily interesting for the user, and conversely one set of updates may not cover all the user needs. These heterogeneous sets of updates could also be concurrent each others and be concurrent with the user dataset.
In this context, how can a user efficiently update his spatial dataset with some evolutions which are not necessarily pertinent and probably concurrent?
This is the essential question to answer if we want to improve the update of spatial data by different sets of evolutions which are coming from multiple sites.
In this paper, we will study the main problem arising when we integrate concurrent and heterogeneous updates and we will propose a process which helps the user to integrate efficiency multi-source updates into his dataset.
This process comprises several steps : Firstly, we classify the evolutions to remove the heterogeneity, secondly we take into account the user needs and exclude the non pertinent data, thirdly we check the concurrency control between all the updates, and finally we reconcile the data if a conflict was detected.
This process uses metadata to choose the “best” evolution to be integrated in the dataset. The metadata used are structured in accordance with the ISO 19115 standard specifications.

Introduction

Spatial data are increasingly available allowing users to update their datasets easily. Data clearinghouses, warehouses or libraries provide a way to consult metadata which give some information about the available data and updates. Mediated or federated systems provide access to data or updates that come from multiple sources. Users can also connect to a portal and search for updates and choose some set of updates which correspond to their needs. Several sets of updates from diverse sources can be furnished to the user, each containing information acquired in different ways, with varying quality and at different times.

One example of a context where the data and updates could come from different sources is a military mission. It is the general context of our study where the actors are distributed between different sites. The units are deployed in France and on the ground of action, and one particularity is that they can be either a producer or a user of the data.
All the actors taking part in the mission have a reference dataset (the data are replicated at each site) which can evolve according to their needs and according to local analyses. The updates must be available for the other actors so that close cooperation can take place. Thus, the units must regularly supply their evolutions to the other actors to guarantee the success of the mission. Problems could arrive when users would like to integrate these numerous evolutions sets in their personal dataset.

Several reasons lead us to think that it is necessary to add information to help the integration process.
Firstly, the updates are heterogeneous, have different quality, can contain errors due to the manner they were acquired. They cannot be integrated directly in the target dataset. It is thus necessary to provide additional information to perform the integration.
Secondly, the updates input by a user on a particular site are not necessarily relevant for a user located at another site (different zones, topics…). Thus, information should be added to allow the exclusion of those evolutions which are finally useless for the end-user.
Finally, the updates coming from multiple sources can be in conflict because they can be ingested by different actors at several times. It is then necessary to detect if these updates are different and if the conflict can create inconsistency. If this is the case, the integration process must be able to propose the updates which are the most suitable according to the user needs and to the evolutions quality. Additional information should be provided to help the process to propose adequate choices.

Douglas Nebert defines an infrastructure as a “concept used to promote a reliable, supporting environment that facilitates the access to geographical information using a minimum set of standard practices, protocols and specifications” (Nebert, 2004, p.8).
Using this definition, our environment of work can be modelled as an infrastructure where the actors and data are known.
« Spatial Data Infrastructure (SDI) is an initiative intended to create an environment in which all stakeholders can co-operate with each other and interact with technology, to better achieve their objectives at different political/administrative levels” (Rajabifard et al. 2001, p.2).
Thus, a military operation can be defined as a SDI because all the actors must cooperate to accomplish the different tasks of the mission.
Figure 1 illustrates a model of SDI hierarchy developed at different political-administrative levels introduced by (Chan and Williamson 1999) and (Rajabifard, 2000). The hierarchy is made up of interconnected SDIs at different levels: global, regional, national, state, local and corporate.

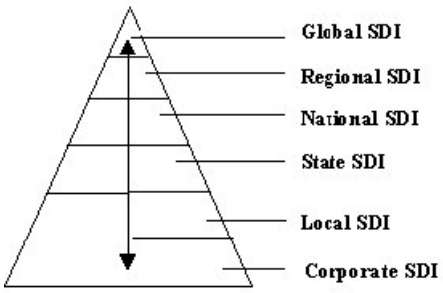

Figure 1 : SDI hierarchy defined by Rajabifard (Rajabifard, 2000)

Using this model, we can define a hierarchy relating to military missions where, for example, a global SDI is used by a multinational army like United Nations missions, a regional SDI is used by European army, a national SDI by the French army and local SDI by units on the ground or by the headquarters.
The main components of a spatial data infrastructure should include data providers, databases and metadata, data network, technologies, institutional arrangements, policies and standards, and end users (Coleman et Nebert, 1998).
A military infrastructure contains all of these components and we can therefore establish a global policy inside the infrastructure, allowing the management and the exchange of data and updates.

Among the most important components of a SDI, there is metadata.
“Metadata helps people who use geospatial data find the data they need and determine how best to use it” (Nebert 2004, p.25). In respect to that point of view, we can assert that metadata can also help people who exchange geospatial evolutions find the updates they need and determine how best to integrate them. Indeed, there is a close link between actors, datasets and evolutions distributed over multi-sites and the use of metadata is one solution to manage this efficiently (Pierkot et al., 2005).
The main objectives of metadata are to organize and maintain the investment made by an organisation, to provide information to data catalogs and clearinghouses and finally to provide information to aid data transfer (Nogueras et al. 2005). To reach this point, organisation must use similar metadata in content and style. This can be done thanks to metadata standards. Some metadata standard dedicated to spatial data exists to ensure the interoperability between all the users handling geographic information.
The Content Standard for Digital Geospatial Metadata was developped by the Federal Geographic Data Committee (FGDC) of the United States to be used in the National Spatial Data Infrastructure (FGDC, 2000).
The European Committee for Standardization (CEN/TC287) also established a standard for the European Geographic Information (CEN, 1998).
But the standard that holds more the attention nowadays is the ISO 19115 ones, defined by the 211 committee of the International Organisation for Standardisation (ISO, 2003). This standard defines the schema for describing geographic information and associated services. It provides much information such as identification, quality or distribution of the spatial dataset. This standard has more than 350 elements. It is very useful because it allows describing numerous resources. But it is very difficult to exploit because of the very great number of elements to manage. Nevertheless, it is possible to create profiles by extend and restrict the

ISO 19115 and as the French Army, more and more organisations use it to create their community profile.
METAFOR is the metadata format for the French Army. It is an ISO 19115 profile that taken into account the French units' needs to share information about spatial datasets used in military mission (Metafor, 2005).

This paper is organised as follow. First, we present the global strategy allowing the integration of the multi-source sets of evolutions on each site taking part into the infrastructure. This paragraph allows understanding why the use of metadata is essential in such a context. Then, we explain in detail our metadata profile called MUMSDI, based on the ISO 19115 standard. Finally, we conclude and give a overview of future work.

Global strategy for updating a particular/personal/user dataset in a military infrastructure:

The context of our study is a closed military infrastructure where the actors must integrate the updates coming from the other sites into their personal datasets.
Several problems arise in such a situation. In particular, the heterogeneity of the evolutions, the relevance of the updates compared to the need for the end-user and the concurrency between the multiple updates.

To aid the update of a spatial dataset by several multi-source sets of evolutions, we thus propose a global strategy which can be applied on each site. Our method takes into account the different problems of heterogeneity, irrelevant of updates and concurrency between the evolutions and with the help of metadata, proposes some solutions to raise them.
This global strategy can be implemented through a process which is dedicated to help a user in the integration of multiple sets of evolution into his personal spatial dataset. This process contains some steps, each one allowing raising one of the problems caused by this context of update.

The first stage of our method concerns the categorisation and classification of the evolutions. Categorisation is the action which consists to organize elements according to predefined categories. Classification is the action which consists to formally group some elements according to their type.
In our context, the evolutions can come from many sites, located either inside (for example, updates coming from the ground of action), or outside the local infrastructure (for example, updates coming from allies already in place). These updates are heterogeneous. Indeed, they do not have necessarily the same format, they were not collected in the same way, neither with the same tools, nor under the same conditions. A common structure must be used to facilitate the integration of such evolutions into the infrastructure. It is thus necessary to transform (categorize and classify) the original updates to be able to exploit them. This step involves establishing associations between the different types of evolutions that can be found in the various set of updates and the type of evolutions which are specified in the infrastructure. Metadata connected to the evolutions allow knowing the manner the updates were collected and thus help us to define the correspondences.

The second stage of our process relates to the exclusion of the non relevant evolutions. Indeed, the changes coming from the various sites can be collected in various contexts and according to different policies of updates. They are not necessarily adequate for the requirements of the final user. For example, the update of a dataset located at the military

headquarters can relate to a geographical zone larger than the zone actually covered by the units, or the updates relating to certain thematic layers can also not interest some actors of the infrastructure. Similarly, an update created with a high level of detail can be unusable in a less detailed dataset. These changes are not finally useful for the end user and must be excluded from the evolutions sets before any integration. Metadata connected to the evolutions and user needs define which evolutions are relevant or not for the end user.

The third step and the most important one, concerns the concurrency check between all the changes to be integrated. Indeed, because of multiple sources of the updates, it is possible to find in different evolutions sets, evolutions concerning the same geographical object or having been placed at the same spatial position. Evolutions are thus concurrent.
The following figure shows the taxonomy used to define the concurrency between the updates in our method:

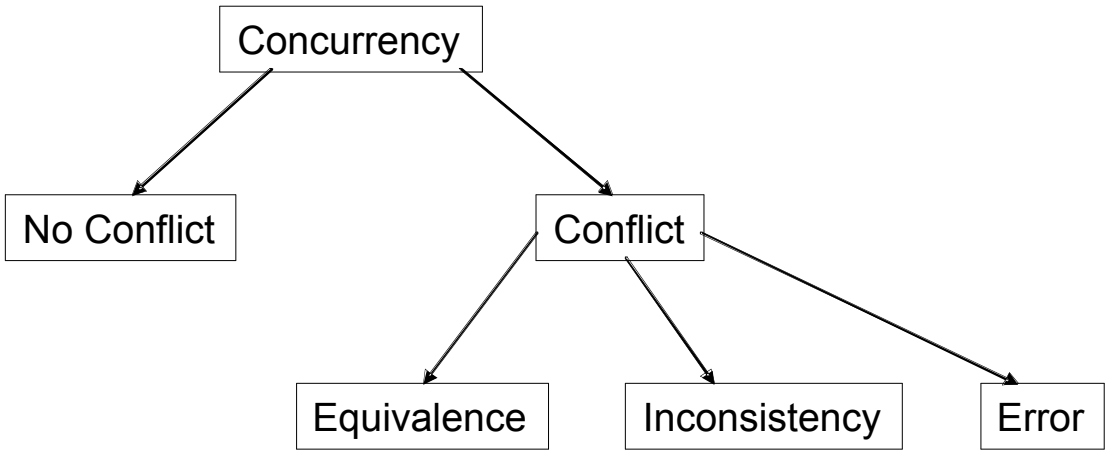

Figure 2 : Taxonomy used to define concurrency between evolutions

Definition 1: Evolutions are **concurrent** if they are located at the same spatial position and/or concern the same entity in the real world. The concurrent updates can be in conflict or not.

Definition 2: Concurrent evolutions are **not in conflict** if they concern the same entity in the real world and are located at the same spatial position. It is thus the same evolution and no conflict must be detected.

Definition 3: Concurrent evolutions are in **conflict** if they do not concern the same entity in the real world but are located at the same spatial position or if they concern the same entity but are located at two different positions. The conflicting updates can be equivalent, inconsistent or erroneous.

Definition 4: Conflict evolutions are **equivalent** if they do not contain errors and if they concern the same entity but are located at two different positions and if they respect the same intentions. For example, two updates of the same road which were be collected at two different levels of detail but respecting the intention “extend the road RN2 till the road RD5” are consider to be equivalent.

Definition 5: Conflict evolutions are **inconsistent** if they do not contain errors, and if they concern the same entity but are located at two different positions and do not respect the same intentions, or if they do not concern the same entity in the real world but are located at the

same spatial position. For example, two updates, one of a road and the other of a building where the road crosses the building. They are not erroneous (in fact they have been entered at different level of abstract) but are inconsistent evolutions.

Definition 6: Conflict evolutions are **erroneous** if at least one contains some errors. The error can concern the value of the attribute, the geometry or the semantic of the evolution (bad classification for example). For examples, two new roads located at the same place, one having the attribute "tarred" and the other "not tarred" or two new roads, one classified as a primary road and the other as a secondary one.

According to the type of conflict, the reconciling method is different. Indeed, in the case of equivalent evolutions, the integration of one or other update will not create inconsistency in the target dataset but one of both will be probably "better" for the user; the choice will be done according to the user needs. In the case of erroneous update, the wrong evolution must be excluded to guarantee the coherence of the dataset. Finally, in the last case, it depends on the nature of the inconsistency. Thus, in step with the nature of the conflict (equivalent, inconsistent or erroneous), the reconciliation process must be able to propose the "best" evolution to preserve, or to suggest the creation of a new update with different parts of the evolutions in conflict. Metadata connected to evolutions and user needs allow proposing the best choice when updates are concurrent.

At each stage of this method, metadata provide the required information to raise the problem. As it is mentioned in a study order by the French government, metadata are important elements for the access, the diffusion and the good utilisation of spatial data. (ADAE, 2006). European project INSPIRE (INfrastructure for SPatial InfoRmation in Europe) insist as well on the importance having harmonized metadata to facilitate the use of geographic data in a spatial data infrastructure.
But as it is mentioned in the GINIE's project final report, spatial data are often badly or not documented (GINIE, 2004). The main reason is that there are too many fields to define and mostly users do not fill them. This is truer in a military context where deployments often happen in crisis situations and where the data must be quickly provided to all the units. Communities must create profiles to avoid having to handle too many sets of metadata and thus facilitating the filling of the metadata fields.

We propose to define a profile which is dedicated to inform the units about the updating of spatial dataset in a military mission context. This profile is described in the following section of this paper.

The MUMSDI profile: Metadata for Updating in a Military Spatial Data Infrastructure

The French army uses metadata structured as the ISO 19115 standard to share information about spatial data between military units. Thus, they created a community profile conforming to the international standard, called METAFOR.
Our context is to exchange updates in a French military mission. In accordance with the French army policy, we must use ISO 19115 metadata to share information and particularly metadata defined by the METAFOR profile. But, the METAFOR profile does not take into account metadata for evolutions and considers all data including raster data. Hence, this profile is not really adapted to our specific problem.

We thus defined another profile called MUMSDI to share information about evolutions resulting from vector datasets updates.
MUMSDI restricts on the one hand METAFOR's profile (and also ISO 19115) by deleting any information not adequate to evolutions data and by imposing a more stringent obligation or a more restrictive domain on some existing metadata elements. On the other hand, it extends the French army profile by adding more information about evolutions quality, particularly information which can be used to determine fitness for use.

The MUMSDI profile contains voluntarily a minimum of elements but the core ISO mandatory elements are preserved in accordance with the ISO 19115 recommendations.
We use UML schema to show how we have extended and restricted METAFOR according to our study needs. Classes represented in yellow (light gray in printed document) referred to the ISO 19115 and Metafor classes. Classes or notes in salmon (dark gray in printed document) referred to the MUMSDI profile.

MUMSDI for the évolutions

Evolutions are not taken into account in the ISO 19115 or in the METAFOR specifications. We propose to add a class named MU_EvolutionsSet to consider updates as well. As the DS_DataSet is defined by the standard, MU_EvolutionsSet has at least one metadata which describe evolutions stored in the set. Class MD_Metadata is the starting point of any information about the evolutions.

Figure 3 : MUMSDI for evolutions

Quality of the updates in the MUMSDI profile

Quality metadata defined by the ISO 19115 standard are dedicated to describe the quality of the data according to the producer point of view.
Sometimes, information about quality provided by the producer is not suitable for the end user because the user needs the data in a particular context and not in a global one. It is the case, for example in military missions, where units located on the ground of action or at the headquarters have not the same needs of data, and can use them at different levels of details, or with different quality. Metadata supply with the producer point of view, are then not adequate for all of the units and the user point of view must be taken into account in the metadata qualities elements.
We propose to add some elements into the standard to consider the user point of view and to restrict the quality elements not necessary for the military context.

The following figure shows a global view of the quality information in the MUMSDI profile. The first changes concern the dataqualityInfo, the lineage and the report roles cardinalities which are mandatory in our profile. We think that it is the simplest manners to know and to evaluate the evolutions which must be integrate into different user's datasets. Indeed, if there is no available quality information with the updates, how the process can know which evolutions is better than the other at the reconcile moment?

At the opposite, we have suppressed several attributes by restricting the cardinality. The main reason is that we think that, the less there are of unnecessary elements in the MUMSDI profile, the easier is to fill them; and more the elements will be fill in, less difficult will be the integration of the evolutions. For example the dateTime attribute in the DQ_Element class has no interest in our context, because the evaluation of the quality of one evolution is made simultaneous with the update of the dataset. The date is then the same than the date of the update.

Moreover, there are some elements in the ISO 19115 which can be filled in by a producer but not by a user on the ground. This user has no technical way to precisely evaluate the evolutions and must evaluate by himself the quality of his update. We think that quantitative results are not sufficient to describe quality information's about evolutions and we have thus added qualitative elements by the MU_QualitativeResult class. This new class allows one to quickly describe if the attributes linked to the data were well or badly documented and what type of errors could be contained in the updates.

The last change in this global aspect concerns the elements of the level of the scope. It is a list which has been restricted and extended in the MUMSDI profile. Components of this list are those for which quality information is available and concern only elements relative to the updating.

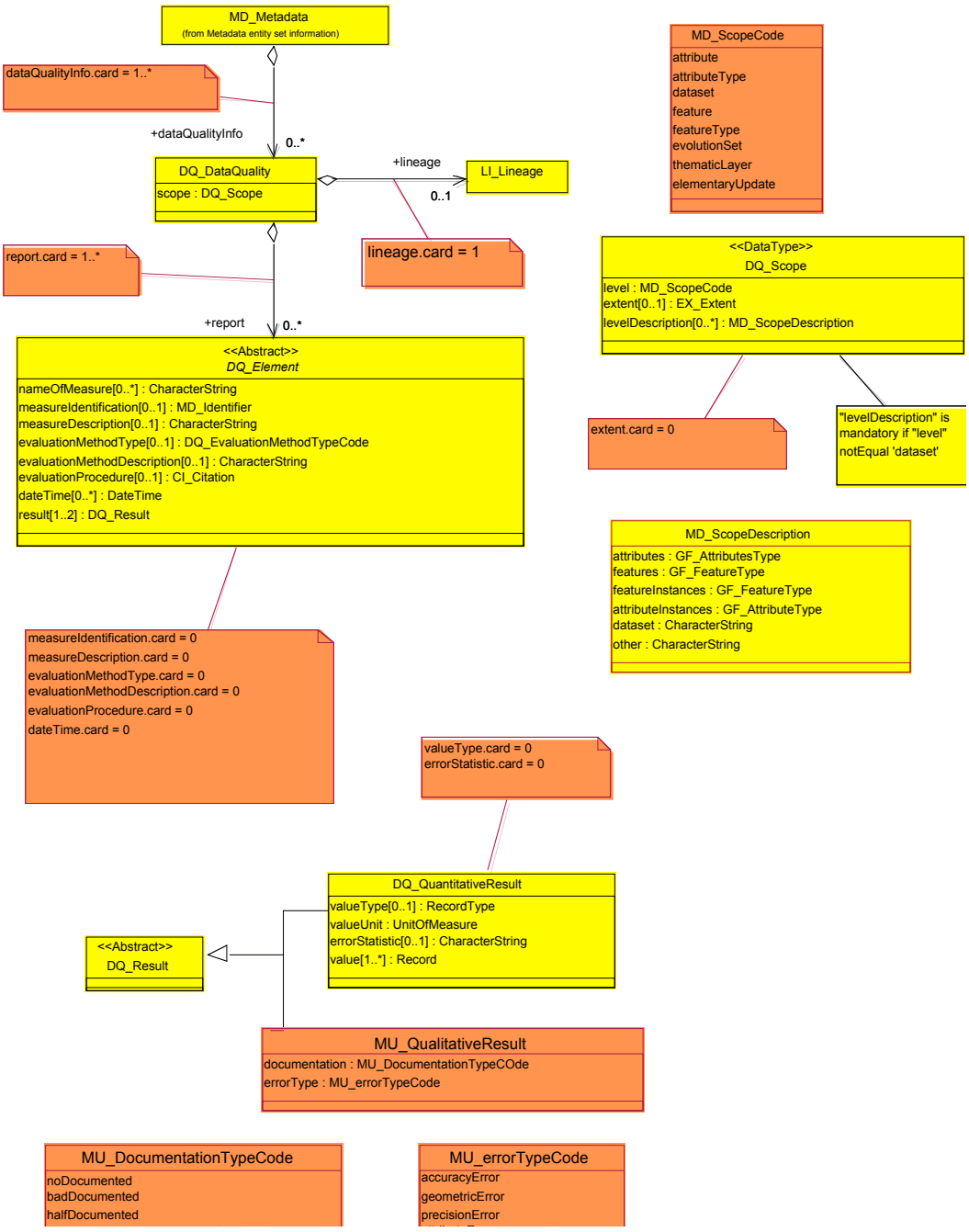

Figure 4 : Quality information in MUMSDI's profile

The following figure shows the quality elements which can be found in the MUMSDI profile. Compared with the standard, we have added and suppressed some elements.
Indeed, the MU_Usability class was inserted in order to evaluate the usability of an evolution in step with the user needs.

Elements about the logical consistency or the temporal accuracy were suppressed because they are not useful for our study and as we noted before, less there are of elements, more metadata are generated.

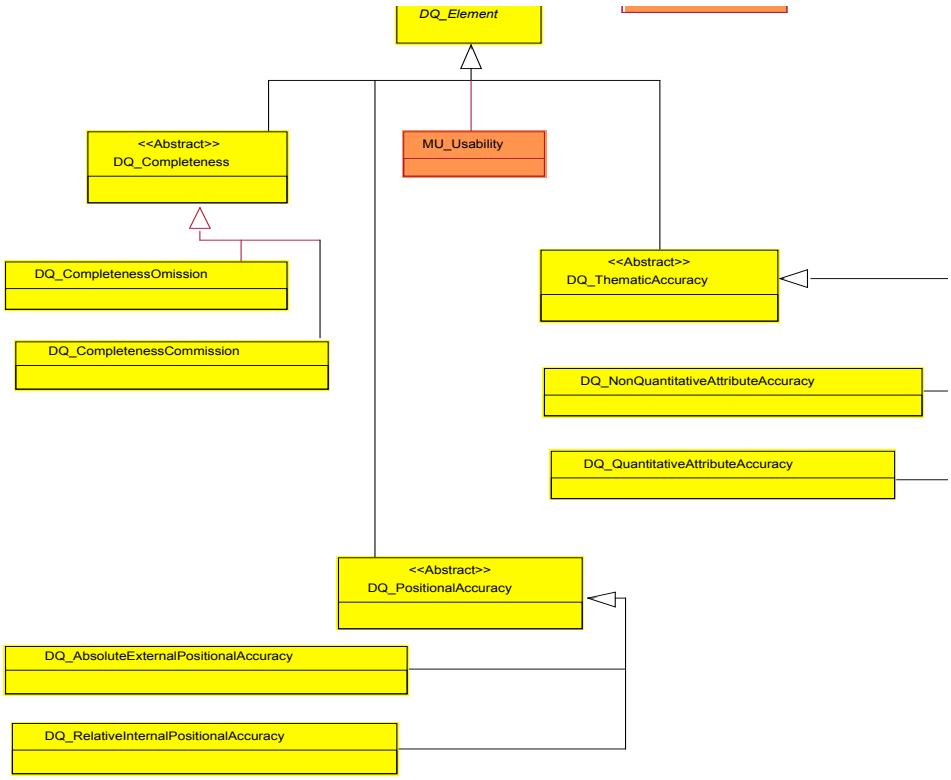

Figure 5 : Quality elements in MUMSDI's profile

Some few others changes in the MUMSDI profile

The others changes concern especially restrictions about elements cardinalities (attributes or classes) or the numbers of elements in the code lists.
For example, the only way to know the extent (attribute extent of the MD_DataIdentification class) of an evolution set is to ask about the geographic bounding box (class EX_GeographicBoundingBox); we do not take into account the temporal extent and we have suppressed the others way.
Another example concerns the numbers of possible values for the different roles of the actors (CI_RoleCode of the attribute role of the CI_ResponsibleParty class); we have extended this list to add the specific roles of the French military actors (included allies, operational actors and producers) and restricted to eliminate roles not essential in a military context (custodian, publisher…).

Conclusion

In this paper, we have proposed a method to manage the update of specific users' spatial dataset by several evolutions sets coming from multiple sites in a military mission context. Military missions can be considered as closed infrastructures where a global policy can be employed.
We thus present a global strategy which can be applied at each site thanks to a process made in several steps. This process is dedicated to manage evolutions which can be heterogeneous, irrelevant or concurrent. It takes into account the end-user needs and thanks to metadata proposes some solutions to lead the integration of the best updates in the user spatial dataset.
According to the French army, the metadata used in this method must be conforming to the ISO 19115 standard and particularly to the METAFOR profile. We thus propose a profile that extends and restricts METAFOR by adding new information in order to consider the evolutions of spatial data and by deleting information superfluous in an updating context. This profile must be used in the infrastructure defined by the mission and a method to fill the elements the most automatically possible must be applied in order to guarantee the future use of the metadata by the integration process.
Today, our process take several evolutions sets, and after some processing, provides a final set which contain only attractives evolutions to integrate in the user spatial dataset.
Future work could be the adaptation of the process to a continous flow of updates which must be integrated in real time.

References

ADAE, 2006, Information Géographique. Recommandation relative aux métadonnées.Projet 8 DT. TN/05.002,Version 1.0. Agence pour le developpement de l'administration electronique. République Française, Ministere du budget et de la réforme de l'état.

CEN, 1998, Geographic Information European Prestandards, Euro-norme Voluntaire for Geographic Information –Data description- Metadata. ENV 12657, European Committee for Standardization – CEN/TC287.

Chan T.O.and Williamson I.P.,1999, The different identities of GIS and GIS diffusion. International journal of Geographic Information Science, 13:3, p.267-281.

Coleman DJ and Nebert DD.,1998, Building a North American Spatial Data Infrastructure. Cartography and Geographic Information Systems, 25(3):151-160.

FGDC, 2000, Content Standard for Digital Geospatial Metadata, version 2.0. Document FGDC-SDT-001-1998, Federal Geographic Data Commitee, Metadata Ad Hoc Working Group. http://www.fgdc.gov/

GINIE, 2004, Geographic Information Network In Europe. Research project. http://www.ec-gis.org/ginie/

INSPIRE, 2006, INfrastructure for SPatial InfoRmation in Europe. Research project. http://inspire.jrc.it/

ISO, 2003a, Geographic Information - Metadata. ISO 19115:2003, International Organisation for Standardisation.

ISO, 2003b, Geographic Information – Metadata. Implementation specification. ISO/WD 19139, International Organisation for Standardisation.

METAFOR, 2005, Gamme de produits CARGENE. Format de fichiers de métadonnées. République Française, Ministere de la défense. IGN/DT.TN/03.054

Nebert D 2004, p.25, Developing Spatial Data Infrastructures : The SDI Cookbook, version 2.0.

Nogueras-Iso J, Zarazaga-Soria F and Muro-Medrano P.R., 2005, Geographic Information Metadata for Spatial Data Infrastrucutres. Resources, Interoperability and Information Retrieval. Springer editions. ISBN :3-540-24464-6

Pierkot C, Mustiere S, Ruas A, Hameurlain H and Raynal L., 2005, Modelling Heteregeneous and Distributed Spatial Datasets in an Update Context. Proceeding in the 22th International Cartographic Conference ICC 2005, Coruna, Espagne, ICA Publications (International Cartographic Association).

Rajabifard A., Williamson I.P., Holland P. and Johnstone G., 2000, From local to Global SDI initiatives : a pyramid building blocks. Proceedings of the 4th GSDI conference, Cape Town, South Africa.

Rajabifard A. and Williamson I.P., 2001, p.2, Spatial Data Infrastructures: Concept, SDI Hierarchy and Future Directions. Proceedings of GEOMATICS'80 Conference, Tehran, Iran.