

N° d'ordre 2010-ISAL-0032

Année 2010

Thèse

Qualité des données capteurs pour les systèmes de surveillance de phénomènes environnementaux

Présentée devant

L'Institut National des Sciences Appliquées de Lyon

Pour obtenir

Le grade de docteur

Formation doctorale

Informatique

École doctorale

École doctorale Infomaths

Par

Claudia Catalina GUTIERREZ RODRIGUEZ

Soutenue le 4 juin 2010 devant la Commission d'examen

Jury MM.

Pr. Robert LAURINI	Professeur à l'INSA de Lyon – Directeur de Thèse
Dr. Sylvie SERVIGNE	Maître de conférences à l'INSA de Lyon – Encadrante de Thèse
Pr. Thérèse LIBOUREL	Professeur à l'Université Montpellier II - Rapporteur
Pr. Tullio TANZI	Professeur à l'ENST (TélécomParisTech) – Rapporteur
Pr. Hervé MARTIN	Professeur à l'Université Joseph Fourier de Grenoble - Examineur

Laboratoire d'Informatique en Image et Systèmes d'Information

A Jean-Pedro, la luz de mi vida.....

Remerciements

Je tiens tout d'abord à exprimer toute ma reconnaissance à mes directeurs de thèse Dr. Sylvie Servigne et Pr. Robert Laurini pour leur encadrement, leurs conseils et leur soutien tout au long de ma thèse. Un grand merci au Pr. Laurini pour m'avoir permis d'entreprendre mes travaux de recherche au sein du laboratoire LIRIS. Merci à Sylvie pour leur temps consacré à me guider et à « décoder » mes idées. Elle a su me conduire dans ce travail avec ténacité, patience et conviction.

Je remercie également aux membres du jury pour avoir bien voulu consacrer une partie de leur temps à ma thèse. Merci au Pr. Thérèse Libourel et Pr. Tullio Tanzi d'avoir accepté d'être mes rapporteurs.

Je souhaite également remercier le Pr. Jean Marc Petit, pour m'avoir accueilli dans l'équipe BD « Base de Données » et au sein du projet Optimacs. Merci pour votre soutien et compréhension dans cette dernière année de thèse.

Un grand merci au Conseil National de Sciences et Technologies (CONACYT) et à la Faculté d'Ingénierie Mécanique et Electrique de l'Université Autonome de Nuevo Leon, grâce à leur soutien économique cette thèse à été possible.

Merci à tous les membres du laboratoire LIRIS, à toutes les personnes que j'ai connues tout au long de ces années de thèse et qui m'ont soutenu à tout moment. Plus particulièrement « les filles » : Sana, Salma et Reim. Vous êtes partie importante de cette réussite, votre amitié, soutien et compagnie dans les bons et mauvais moments resteront dans mon cœur pour toujours.

Mes remerciements s'adressent aussi spécialement à Karine Rouyre, Karla Lopez, Frédéric Flouvat, Pablo Mendez, Darina Dosamantes, Dr. Alan Castillo, Dr. Rogelio Rivera, Dr. Moises Hinojosa, Pr. José Angel Mendoza et Severine Fantini.

Egalement, je remercie cordialement Yann Gripay, Roberto Veiga, Diego Lemos, Sandro Bimonte, Dr. Frédérique Laforest, Dr. Marian Scaturucci, Dr. Elod Egyed-Zsigmond. Aussi à Ahlem Zayati, Arnaud Renard, Julien Canquelain, Simon Geffrey, Brice Chardin, Daphne Parramond, Caroline Ferri et Mabrouka Gheraissa. Sans oublier mes grands amis mexicains Ivette Mendoza, Agueda Garcia, Aldo Martinez, Fernando y Santiago Banda, Gerardo Gonzalez, Yessica Gallardo, Delfina Casique y Gerardo Cabrera.

Je remercie très particulièrement ma famille. Même en étant loin, ils ont eu toujours les mots pertinents pour me soutenir et me motiver pour continuer dans mon objectif. Sans leur soutien, cette réussite n'aurait pas eu lieu. Merci à ma mère pour tout son amour, ses prières, sa force pour me pousser à aller de l'avant, à ma sœur et mon frère qu'on toujours cru en moi. Un pensé très spécial à mon père. Et finalement je remercie à Ti, qui a su supporter mes angoisses, qui a cru en moi, m'a encouragé à tout moment, merci de tout mon cœur « *Só enquanto eu respirar, vou me lembrar de você* ».

Qualité des données capteurs pour les systèmes de surveillance de phénomènes environnementaux

Résumé

De nos jours, les applications émergentes dans le domaine géographique, exploitent de plus en plus d'informations géolocalisées provenant de capteurs, notamment pour la gestion de crises, la gestion de véhicules en temps réel, la gestion de risques urbains ou environnementaux, etc. Plus particulièrement, l'utilisation de capteurs au sein du domaine de la surveillance, notamment environnementale (i.e. inondations, avalanches, volcans...), permet une interprétation plus simple du monde réel. Cependant, les grandes quantités de données issues de ces capteurs, à fréquences et positions variables dans le temps, acquises dans des environnements hostiles avec une capacité d'énergie et de communication limitée, rendent les données plus imprécises et incertaines. Ceci pose le problème de la qualité des données issues de capteurs.

La qualité des données est fortement intégrée dans les Systèmes d'Information Géographiques. En effet, grâce aux standardisations existantes dans le domaine (i.e. série ISO 19100), l'évaluation et l'amélioration de la qualité ont permis une meilleure représentation de l'information géographique facilitant ainsi sa compréhension et interprétation par les utilisateurs. Ces standards proposent des éléments, critères et méthodes d'évaluation dédiés aux données spatiales (données statiques exploitées dans des applications conventionnelles), qui sont bien connus de la communauté géographique internationale. Dans cet esprit, certaines organisations internationales dans le domaine géospatial et environnemental comme l'Open Geospatial Consortium (OGC), le National Aeronautics and Space Administration (NASA) ou encore le National Oceanic and Atmospheric Administration (NOAA), ont soulevé le besoin de réaliser des études plus approfondies sur le sujet de l'évaluation et le contrôle de la qualité de données issues de capteurs. Ces organisations, nécessitent de plus en plus d'informations liées à la qualité de données en temps réel pour les différents acteurs (i.e. géologues, météorologies, etc.).

Le but de notre travail est donc de relever le challenge de la qualité des données issues de capteurs en temps réel, afin d'aider les utilisateurs lors de la prise de décision en situations critiques. Ainsi, nous proposons une méthodologie pour la définition, l'évaluation et la communication de la qualité des données issues de capteurs. Cette méthodologie s'inspire de méthodologies existantes traitant de la qualité de données dans différents domaines et est structurée en trois phases (*la définition, l'évaluation et la communication*).

Cette méthodologie, à la différence des méthodologies existantes dans le domaine des systèmes d'information, a nécessité la spécification de caractéristiques des données issues de capteurs liées à la qualité des données. L'évaluation de la qualité, au sein de la méthodologie prend en compte les facteurs du contexte d'acquisition et du traitement qu'ont un impact sur la qualité des données, ainsi que la gestion des métadonnées qualité et l'évaluation de la qualité multicritères en temps réel.

La méthodologie proposée est supportée par un prototype de visualisation appelé MoSDaQ (Monitoring Sensor Data Quality) qui permet de visualiser via le web les

données issues des observations environnementales, conjointement aux informations liées à la qualité des données exploitées en temps réel.

Mots-Clés: Qualité des données - méthodologie qualité de données - métadonnées - données capteurs - réseaux de capteurs – temps réel - SIG – aide à la décision.

Sensor data quality for environmental phenomena monitoring systems

Abstract

Nowadays, emerging applications in the geographic domain increasingly explore geolocalized information provided by sensors, specially for crisis management, real-time vehicle management, urban or environmental risks management, etc. More particularly, the use of sensors in the surveillance domain, specially environmental surveillance (i.e., floodings, avalanches, volcanoes...), allow a simpler interpretation of the real world. In the meanwhile, the great volume of data coming from these sensors, at variable frequency and positions, acquired in hostile environments with limited battery and communication power turn the data imprecise and uncertain. We have thus a problem regarding the quality of the data provided by these sensors.

The quality of these data is strongly integrated in the Geographic Information Systems. Actually, thanks to the existent standards in the domain (i.e., ISO 19100 series), the evaluation and improvement of data quality allowed a better representation of the geographic information, making its comprehension and interpretation by users easier. These standards propose elements, criteria and methods of evaluation dedicated to spatial data (static data explored in conventional applications), that are quite known by the international geographic community. Following this idea, some international organizations in the geospatial and environmental domain, like the Open Geospatial Consortium (OGC), the National Aeronautics and Space Administration (NASA) or the National Oceanic and Atmospheric Administration (NOAA), have raised the need to perform deeper studies over the subject of evaluation and control of sensor data quality. These organizations increasingly need information related to real-time data quality for different actors (i.e., geologists, meteorologists etc).

The goal of our work is to raise the challenge of sensor real-time data quality, in order to help users in the decision making during critical situations. This being, we propose a methodology for the definition, evaluation and communication of sensor data quality. This methodology is inspired by existent methodologies dealing with data quality in different domains and is structured in three phases (*the definition, the evaluation and the communication*).

This methodology, differently from existent methodologies in the information systems domain, needed the specification of sensor data characteristics linked to data quality. The evaluation of the quality, in the core of the methodology, takes into account the acquisition context factors and treatment that can impact the data quality, as well as the metadata management and the evaluation of multi-criterion quality in real-time.

The proposed methodology is supported by a visualization prototype called MoSDaQ (Monitoring Sensor Data Quality) that allows us to visualize via web the data provided by environmental observations, together with information related to the data quality in real-time.

Mots-Clés: Data quality – data quality methodology – metadata – sensor data – sensor networks – real-time – SIG – decision-making support

Table des matières

<i>Introduction générale</i>	1
1.1 Problématique.....	3
1.2 Contributions	4
1.3 Organisation du manuscrit.....	6
Première partie : État de l'art	8
<i>Chapitre 2 La qualité : données, métadonnées, services</i>	9
2.1 La Qualité de données dans les Systèmes d'Information	9
2.1.1 Dimensions et caractéristiques de la qualité de données.....	11
2.1.2 Modélisation de la qualité de données	14
2.1.3 Méthodologies pour l'évaluation et l'amélioration de la qualité de données	16
2.2 La Qualité de données dans les Systèmes d'Information Géographiques.....	19
2.2.1 Modélisation de la qualité de données géographiques	20
2.2.2 Métadonnées qualité et données géographiques	22
2.2.3 Spécification des normes et la qualité de données géographiques.....	24
2.2.4 Évaluation de la qualité de données géographiques	26
2.2.5 Communication et visualisation de la qualité des données géographiques.....	28
2.3 La qualité de service – QoS	29
2.3.1 Modélisation de la Qualité de Service	29
2.3.2 Gestion de la Qualité de Service.....	30
2.3.3 La qualité de service et la qualité de service spatiale.....	31
2.4 Conclusion	32
<i>Chapitre 3 Les réseaux de capteurs : description et applications</i>	34
3.1 Les capteurs et réseaux de capteurs	34
3.1.1 Les capteurs.....	34
3.1.2 Les réseaux de capteurs.....	36
3.2 Les données et flux issus de réseaux de capteurs.....	37
3.2.1 Traitement de données de capteurs	37
3.2.2 Stockage des données capteurs	38

3.2.3	Réseaux de capteurs et domaines d'application	39
3.3	Les réseaux de capteurs et la qualité de données	43
3.3.1	Les réseaux de capteurs et le contexte environnemental	44
3.3.2	Les réseaux de capteurs et la QoS.....	45
3.4	Conclusion	46
Deuxième partie : Contribution		48
<i>Chapitre 4 Modélisation des données capteurs dans les systèmes de surveillance.....</i>		<i>49</i>
4.1	Du phénomène au système de surveillance	49
4.1.1	Couches de traitement dans les systèmes de surveillance environnementaux	50
4.1.2	Spécificités des systèmes de surveillance environnementaux.....	51
4.2	Données spatio-temporelles issues de réseaux de capteurs : données observation.....	53
4.2.1	Gestion de données spatio-temporelles issues de réseaux de capteurs	54
4.2.2	Formalisation des objets issus de capteurs	57
4.3	Traitement des données spatio-temporelles issues de réseaux de capteurs.	61
4.3.1	Données capteurs et temps réel.....	61
4.3.2	Spécificités du traitement des données spatio-temporelles issues de capteurs	62
4.4	Conclusion	63
<i>Chapitre 5 Définition et évaluation de la qualité des données capteurs.....</i>		<i>66</i>
5.1	Phase 1 : définition de la qualité des données issues de capteurs.....	68
5.1.1	Analyse des facteurs d'impact de la qualité des données issues de capteurs	69
5.1.2	Spécification des critères d'évaluation de la qualité de données capteurs en temps réel	92
5.1.3	Définition du modèle de qualité pour les données issues de capteurs	101
5.2	Phase 2 : l'évaluation de la qualité des données issues de capteurs	107
5.2.1	Métadonnées spatiotemporelles en temps réel : sources d'information qualité	108
5.2.2	Évaluation de la qualité des données issues de capteurs.....	126
5.2.3	Gestion de la qualité de données issues de capteurs	129
5.3	Phase 3 : Communication de la qualité des données issues de capteurs	134
5.3.1	Stratégies de communication de la qualité des données issues de capteurs : indicateurs qualité et rapport qualité	135
5.3.2	Appréciation de la qualité des données capteurs	140
5.3.3	Spécification du rapport qualité.....	147

5.4	Conclusion	151
<i>Chapitre 6 MoSDaQ : Un prototype pour la visualisation de qualité des données dans les systèmes de surveillance environnementaux.....</i>		
6.1	Description du prototype.....	152
6.2	Cas d'utilisation	155
6.3	Visualisation des données d'observation et de l'information qualité.....	159
6.4	Conclusion	170
Troisième partie : Conclusion et perspectives		171
7.1	Synthèse	172
7.2	Discussion et perspectives	174
7.2.1	Evaluation de la qualité de données capteurs et les experts	174
7.2.2	La qualité des données : dynamicité et hétérogénéité	175
7.2.3	La découverte de connaissances et la prise de décision.....	175
7.2.4	L'extension du prototype MoSDaQ.....	176
<i>Bibliographie</i>		<i>177</i>
<i>Annexe</i>		<i>187</i>

Index des figures

Figure 1 - Principales problématiques de la qualité des données (adapté du [Batini'06])...	10
Figure 2 - Dimensions de la qualité de données proposées par l'approche empirique [Wang'96].....	12
Figure 3 – Dimensions proposées dans une approche intuitive [Redman'97]	13
Figure 4 – Classification de méthodologies [Batin'09]	17
Figure 5 – Éléments dans les méthodologies pour l'évaluation et l'amélioration de la qualité des données	19
Figure 6 – Diagramme d'entités de métadonnées standardisé par l'ISO 19115 [ISO'03a]	23
Figure 7 – Critères qualité compris par la norme ISO 19113 [ISO'02].....	25
Figure 8 – Concepts de la qualité de service (QoS) [ISO'99]	29
Figure 9 – Capteurs au terrain d'observation [Delin'00]	35
Figure 10 – Système de surveillance de phénomène naturel [Noël'06]	41
Figure 11 – Diagramme schématique d'un réseau de capteurs environnemental [Hart'06]	41
Figure 12 – Architecture technique d'un système de surveillance environnementale	50
Figure 13 – Couches de traitement dans les systèmes de surveillance environnementaux	50
Figure 14 – Modèle conceptuel d'observation.....	53
Figure 15 – Modèle de données spatio-temporelles temps réel [Gutiérrez'07].....	55
Figure 16 – Éléments statiques du modèle de données spatio-temporelles temps réel	56
Figure 17 - Éléments dynamiques du modèle de données spatio-temporelles issues de capteurs.....	56
Figure 18 – Méthodologie d'évaluation de la qualité des données spatio-temporelles issues de capteurs	67
Figure 19 – Phase 1 : Définition de la qualité des données issues de capteurs.....	68
Figure 20 – Phase 1 – Analyse de facteurs d'impact de la qualité des données capteurs.....	69
Figure 21 - Étapes de sélection des facteurs qualité dans le système de surveillance.....	69
Figure 22 – Analyse des facteurs qualité selon les processus de traitement de données capteurs.....	70
Figure 23 - Processus d'acquisition et composants	71
Figure 24 - Composants du processus de transformation	75
Figure 25 - Composants du processus de transformation	80
Figure 26 - Modèle conceptuel des erreurs dans un système de surveillance	83
Figure 27 – Énergie du capteur vs Exactitude Figure 28 – Temps de vie vs Précision	88
Figure 29 – Phase 1 – Spécification des critères qualité	92
Figure 30 – Erreurs et critères qualité au sein du système de surveillance	93

Figure 31 - Catégories de la qualité des données issues de capteurs	93
Figure 32 – Structuration des critères qualité des données capteurs	94
Figure 33 – Phase 1 – Définition du modèle de qualité	101
Figure 34 – Modèle de qualité de données capteurs	102
Figure 35 – Interaction du modèle de données capteurs et modèle de qualité.....	103
Figure 36 – Phase 2 : l'évaluation de la qualité des données issues de capteurs	107
Figure 37 – Phase 2 – Spécification des sources qualité	108
Figure 38 – Axes d'étude des métadonnées spatio-temporelles temps réel [Gutiérrez'09]	110
Figure 39 – Modèle global des métadonnées spatio-temporelles proposées	114
Figure 40 – Composition des Métadonnées Observation.....	115
Figure 41 – Composition des Métadonnées capteur.....	115
Figure 42 - Composition des Métadonnées Capteur Statique.....	116
Figure 43 - Composition des Métadonnées Capteur Dynamique.....	117
Figure 44 - Composition des Métadonnées Qualité.....	118
Figure 45 – Modèle intégral de données capteurs et la qualité de données.....	122
Figure 46 – Diagramme de séquence pour l'extraction et mise à jour des MD_Observation	124
Figure 47 - Diagramme de séquences pour les métadonnées capteurs	125
Figure 48 - Diagramme de séquences pour les métadonnées qualité	126
Figure 49 – Phase 2 – Evaluation de la qualité des données capteurs.....	126
Figure 50 – Phase 2 – Gestion de la qualité des données capteurs	129
Figure 51 – Relations dans le traitement de la qualité de données capteurs (adapté [Klein'07]).....	130
Figure 52 – Exemple d'interaction entre les données et les métadonnées en forme de table relationnelle.....	131
Figure 53 – Modèle de traitement de la qualité des données issues de capteurs	131
Figure 54 – Architecture du système de gestion des données issues de capteurs	132
Figure 55 – Méthodologie SDAQ – Phase 3 : la communication	134
Figure 56 – Phase 3 – l'analyse des stratégies de communication	135
Figure 57 – Exemple des données en temps réel issues de bouées d'observation océaniques (TAO/TRITON)	136
Figure 58 – Tableau de bord proposé par [Devilleers'02].....	138
Figure 59 – Tableau de bord proposé par [Huth'07]	138
Figure 60 – Phase 3 – Appréciation de la qualité de données capteurs.....	141
Figure 61 – Symboles slider, plus/moins et compteur dans la complétude.....	144
Figure 62 – Exemple de symboles pour l'exactitude, la précision et la fiabilité	145
Figure 63 – Exemple de symboles pour l'actualité et la ponctualité.....	145
Figure 64 – Exemple du symbole indicateur pour le critère volatilité.....	145
Figure 65 – Exemple de notre approche de représentation des indicateurs qualité.	146

Figure 66 – Phase 3 – Spécification du rapport qualité.....	147
Figure 67 – Concept du rapport qualité au sein du traitement de données issues de capteurs	148
Figure 68 – Processus de génération du rapport qualité à la demande de l'utilisateur.....	149
Figure 69 – Exemple du rapport qualité généré.....	150
Figure 70 – Système de surveillance environnemental.....	153
Figure 71 - Architecture générale du prototype MoSDaQ.....	153
Figure 72 – Surveillance volcanique.....	155
Figure 73 – Architecture physique d'un système de surveillance volcanique.....	155
Figure 74 – Plan cartographique du volcan Popocatépetl et ses environs [Cenapred'06]	156
Figure 75 – Surveillance du Cenapred au volcan Popocatepetl [Quaas'06].....	156
Figure 76 – Image du volcan Popocatepetl	Figure 77 – Rapport de l'état de la surveillance.....
Figure 78 - Aperçu du prototype d'interface web MoSDaQ.....	160
Figure 79 – Section de mapping des objets géolocalisés (capteurs) au sein du volcan.....	161
Figure 80 – Affichage du type cartographique	Figure 81 – Affichage du type terrain.....
Figure 82 – Symboles utilisés pour représenter les capteurs et signaux d'alerte au sein des capteurs.....	162
Figure 83 – Information sur le capteur sélectionné en temps réel.....	162
Figure 84 – Exemple de représentation en temps réel de la valeur mesurée du capteur de température.....	163
Figure 85 – Section d'information observation.....	163
Figure 86 – Section Information des capteurs.....	164
Figure 87 – Indicateurs de conditions environnementales.....	165
Figure 88 – Section d'indicateurs sur la qualité des données issues de capteurs.....	166
Figure 89 – Configuration des indicateurs qualité.....	166
Figure 90 – Indicateurs sur la qualité de données issues de capteurs.....	167
Figure 91 – Rapport qualité généré lors d'une évaluation.....	168
Figure 92 – Section d'interrogation.....	169
Figure 93 – Composition des Métadonnées Observation.....	188
Figure 94 – Composition des Métadonnées capteur.....	189
Figure 95 - Composition des Métadonnées Capteur Statique.....	190
Figure 96 - Composition des Métadonnées Capteur Dynamique.....	192

Index des tables

Tableau 1 – Dimensions récurrentes dans la qualité de données	14
Tableau 2 – Approches générales sur la modélisation de la qualité de données [Wang'96] [Naumann'99]	15
Tableau 3 – Critères de référence pour la qualité de données géographiques [ISO'02]	21
Tableau 4 – Composantes des mesures qualité [ISO'06]	26
Tableau 5 – Dimensions orientées qualité de l'information et QoS	32
Tableau 6 – Catégorisation d'application des réseaux de capteurs	40
Tableau 7 – Facteurs d'impact qualité : Processus d'acquisition	73
Tableau 8 – Facteurs d'impact qualité : Processus de transformation	78
Tableau 9 – Facteurs d'impact qualité : Processus d'exploitation	81
Tableau 10 - Facteurs qualité du processus d'acquisition vs critères qualité au domaine géographique.....	86
Tableau 11 - Facteurs qualité du processus d'acquisition vs critères QoS et temps.....	88
Tableau 12 - Facteurs qualité du processus de traitement vs critères qualité au domaine géographique.....	89
Tableau 13 - Facteurs qualité du processus de transformation vs critères QoS et temps....	90
Tableau 14 – Facteurs qualité du processus d'exploitation vs critères qualité au domaine géographique.....	91
Tableau 15 – Facteurs qualité du processus d'exploitation vs critères qualité de service et liées au temps	91
Tableau 16 – Valeurs de température caractérisées par deux critères qualité	105
Tableau 17 – Valeurs d'un flux de température et informations qualité.....	106
Tableau 18 – Information contenue dans le rapport qualité proposé par l'ISO 19114	140
Tableau 19 – Tableau comparatif des symboles et l'utilité pour la description des critères qualité des données capteurs (inspiré de [Huth'07] et [Devillers'07]).....	143
Tableau 20 – Proposition du contenu du rapport qualité orientée données issues de capteurs	148
Tableau 21 – Extrait du schéma de métadonnées observation	188
Tableau 22 - Extrait du schéma des métadonnées capteur	189
Tableau 23 – Extrait de la composition des métadonnées capteur statique.....	190
Tableau 24 – Extrait du schéma de métadonnées capteur dynamique.....	192

Introduction générale

De nos jours, les systèmes d'information ont connu une évolution importante dans plusieurs domaines (médical, géographique, etc). Cette évolution a entraîné la diversité des données produites et traitées ainsi que de leurs sources. Face à cette constante évolution, les utilisateurs sont de plus en plus intéressés par la qualité des données exploitées [Peralta'04] [Wang'96]. Le grand challenge est de fournir aux utilisateurs une vision bien spécifique sur la fiabilité des données et de leurs sources. En effet, une qualité de données médiocre expose les utilisateurs et organisations (scientifiques, médicales, d'aménagement, etc.) au risque de compromettre leurs décisions et impacter l'exécution des systèmes. Il est donc nécessaire de pouvoir déterminer la qualité de ces données ainsi que celle de leurs sources à fin de permettre ainsi un accès et une gestion de la qualité des connaissances produites, en temps réel.

La qualité de données, a fait l'objet de nombreux travaux de recherche [Wang'95], [Strong'97], [Smith'98], [Veregin'99], [Naumann'04], [Berti-Equille'04], [Akoka'07] qui ont abouti à la définition de *modèles et méthodologies orientés qualité* [Batini'09], à la proposition de *techniques d'évaluation et d'amélioration* [Naumann'02], ainsi qu'à *des standardisations* [ISO'00].

Dans le domaine géographique, la qualité de données représente également un enjeu important. Elle a fait l'objet de plusieurs analyses, qui ont abouti principalement aux normes de la série 19100, dédiées à la qualité des données géographiques (i.e. [ISO'02]). Plus précisément, les applications émergentes dans le domaine du contrôle et de la surveillance, la gestion de risques, de crises (dans les milieux urbains ou environnementaux) exploitant ces données géospatiales (spatio-temporelles) issues de capteurs, ont soulevé le fait que la qualité des données demeure un élément critique dans l'utilisation des nouveaux systèmes d'information [OGC'09] [NOAA'09], [TAO'06], [SWE'07], [Sherwood'04], [MMI'07]. Les données issues de ces capteurs sont généralement collectées, prétraitées et ensuite envoyées vers un centre d'observation pour être analysées et mises à disposition des utilisateurs. L'exploitation de ces données est souvent faite en ligne, via le web ou au sein d'un Système d'Information Géographique. Disposer d'informations complémentaires, notamment sur la qualité de données dans ces systèmes permettra aux utilisateurs d'améliorer leurs actions.

Les données traitées par ces systèmes sont souvent hétérogènes provenant de divers capteurs et ayant des types différents (géographique, géospatiale, géoréférencée ou encore géolocalisée). L'utilisation de capteurs et de réseaux de capteurs au sein des applications pour la surveillance environnementale est en constante augmentation. Les capteurs permettent une acquisition plus simple et plus performante de renseignements sur le monde réel, en fournissant de grandes quantités de données à fréquences et positions variables dans le temps. Ces dispositifs d'acquisition présentent une capacité d'énergie et de stockage limité et sont souvent déployés dans des conditions périlleuses (changements climatiques, sinistres, actions malveillantes etc.). Ces aspects, analysés dans divers travaux de recherche, notamment dans le domaine des réseaux de capteurs [Xu'09], [Wilson'05], [Xu'04], [Worboys'06], [Nittel'09], impactent principalement l'exactitude, la précision et la fiabilité des données acquises. En effet, ces données issues de capteurs peuvent être erronées ou encore manquer de fiabilité nuisant ainsi à des prises de décision adéquates, et due aux objectifs des utilisateurs. Ces utilisateurs

peuvent être des volcanologues, des urbanistes, des biologistes..., qui élaborent des plans et des stratégies basés sur l'analyse de ces données. Ils sont ainsi souvent exposés au risque de traiter des données non fiables s'ils ne possèdent pas de moyens permettant d'évaluer ou d'assurer le niveau de qualité de l'information. Par conséquent, la prise de décisions basée sur des données d'une qualité médiocre peut entraîner des actions incorrectes comme le déclenchement intempestif d'une alerte, des représentations cartographiques erronées (i.e. routes inexactes, croisements manquants,...), une mauvaise reconnaissance du terrain et, jusqu'à l'endommagement du système d'information en général. Ceci pose la question de la qualité de données issues de capteurs.

Cette thèse se focalise sur les enjeux de la qualité des données issues de capteurs exploitables dans des applications de surveillance de phénomènes environnementaux. Dans ces applications, où les contraintes de temps pour la prise de décision sont plus importantes par rapport aux applications traditionnelles, l'information qualité représente un complément nécessaire à l'utilisateur lors de sa prise de décision. Ce type d'applications exploite des informations géolocalisées (voire des données spatio-temporelles temps réel) soit à partir de capteurs filaires fixes, agiles (capteurs avec une mobilité restreinte) ou mobiles, soit par des réseaux de capteurs sans-fil (notamment les réseaux de géocapteurs), soit à partir des deux. Dans ce type de systèmes, les données issues de capteurs sont variables dans le temps, voire dynamiques, contrairement aux données traitées par les applications traditionnelles. En effet, ces données sont collectées dans des situations souvent difficiles, possèdent des caractéristiques spatiales et temporelles, et évoluent au cours du temps. Ces spécificités nécessitent donc un traitement particulier en termes de qualité prenant en considération l'hétérogénéité et la « dynamique » des données capteurs.

1.1 Problématique

Actuellement, les applications émergentes dans le domaine géographique, nécessitent de plus en plus l'exploitation des informations géolocalisées provenant de capteurs, notamment pour la gestion de crises, la gestion de véhicules en temps réel, la gestion de risques urbains ou environnementaux, etc. Plus particulièrement, la surveillance de phénomènes environnementaux (i.e. inondations, avalanches, volcans...) est un processus complexe qui consiste à évaluer de façon continue les flux de données provenant de capteurs. Le principal objectif est de détecter un comportement inhabituel et de communiquer les informations aux experts afin d'éviter les dommages graves au sein des populations ou/et de l'environnement.

L'utilisation des capteurs, permet une interprétation plus simple du monde réel. Cependant, les grandes quantités de données issues de ces capteurs, à fréquences et positions variables dans le temps, acquises dans des environnements hostiles avec une capacité d'énergie et de communication limitée, rendent les données plus imprécises et incertaines.

La qualité des données est fortement prise en compte dans les Systèmes d'Information Géographiques. En effet, grâce aux standardisations existantes dans le domaine (i.e. ISO19100), l'évaluation et l'amélioration de la qualité ont permis une meilleure représentation de l'information facilitant ainsi sa compréhension et interprétation par les utilisateurs. Ces standards proposent des éléments, critères et méthodes d'évaluation dédiées aux données spatiales, qui sont bien connus dans la communauté géographique internationale (i.e. IGN¹, OGC²). Dans ce contexte, les métadonnées géographiques sont aujourd'hui définies selon ces normes (ISO 19115), et leur principal objectif est de faciliter l'échange de données entre différents utilisateurs et les renseigner sur la qualité de ces données. Cependant ces métadonnées ont été définies pour des données statiques exploitées par des applications conventionnelles, qui ne prennent pas en compte la dynamique des données. Une analyse vers la structuration et l'exploitation des métadonnées en temps réel sur des données dynamiques et plus précisément celles issues de capteurs semble nécessaire.

Dans ce contexte, certaines organisations internationales dans le domaine géospatial et maritime comme l'OGC, la NASA³, la NOAA⁴, ont soulevé le besoin de réaliser des études plus approfondies sur le sujet de l'évaluation et le contrôle de la qualité de données issues de capteurs. Ces organisations, nécessitent actuellement de plus en plus d'information liée à la qualité en temps réel, pour les différents acteurs (organisations météorologiques, géologues, la société elle-même...). Ces informations doivent être complètes, fiables, compréhensibles réutilisables, partageables et obtenues selon des contraintes de temps et d'espace bien spécifiques.

Le but de notre travail est donc de relever le challenge de la qualité en temps réel sur des données issues de capteurs, afin d'aider les utilisateurs (experts, gestionnaires de l'environnement, gestionnaires du patrimoine urbain, etc.) dans la prise de décision, notamment en situation de crise.

C'est dans ce contexte que plusieurs questions se posent : *Quel est l'impact des réseaux de capteurs et du temps réel sur la qualité de données ? Quelles sont les informations à prendre en considération dans les métadonnées ? Comment distinguer données et métadonnées dans un contexte dynamique ? Comment définir leur granularité ? Comment les gérer et les stocker en temps réel ? Sur quels critères évaluer la qualité ? À quelle périodicité ?...*

Ce travail de recherche vise à apporter des solutions aux questions précédentes dont les contributions seront décrites plus en détail dans la section suivante.

1.2 Contributions

L'objectif principal de notre travail de recherche est de fournir en temps réel, des informations sur la qualité de données exploitées dans les systèmes de surveillance de phénomènes environnementaux, afin d'aider les utilisateurs lors de la prise de décision

¹ Institut de Géographie National

² Open Geospatial Consortium

³ National Aeronautics and Space Administration

⁴ National Oceanic and Atmospheric Administration

périodes critiques (délais de temps restreints). Plus précisément, nous étudions les spécificités de données issues de capteurs afin d'évaluer leur qualité. Dans ce cadre, nous considérons que la qualité des données exploitées par l'utilisateur final au sein d'un système de surveillance dépend principalement de la qualité des sources de données et des caractéristiques des processus de communication et de traitement. Plus spécifiquement, la qualité dépend de la qualité du contexte d'acquisition (i.e. la précision et fiabilité des sources) et des types de processus qui traitent ces données. Ainsi, nous présentons une méthodologie pour la définition de la qualité des données issues de capteur, qui aboutit à la proposition d'une solution d'évaluation et de communication de la qualité de ces données. Cette méthodologie de définition, prend en compte les méthodologies existantes [Batini'09] traitant la qualité de données dans différents domaines, et propose un contexte d'évaluation de la qualité en considérant les facteurs contextuels des capteurs, les métadonnées qualité et la communication de ces aspects en temps réel.

Au cœur de cette méthodologie une spécification des caractéristiques des données issues de capteur est présentée. Dans ce cadre, nous proposons un modèle de données capteur [Gutiérrez'07], qui se base sur le langage de modélisation UML⁵, afin de mettre en évidence les éléments potentiellement exploitables pour constituer l'information qualité. Dans ce modèle de données capteurs, nous avons mis en place diverses classes concernant l'environnement d'observation d'un phénomène, par exemple, une observation déterminée par des zones, des périodes et des éléments d'observation. Ainsi, nous avons mis en relation les éléments statiques et dynamiques (changeant au cours du temps) caractérisant le contexte d'acquisition. Une formalisation concernant les objets issus de capteurs et leurs mesures est également proposée.

Le besoin d'une évaluation de la qualité de données capteurs, fédère cette méthodologie qui permet principalement de définir ce que la qualité des données représente et qui propose à l'utilisateur un moyen d'évaluer et d'exploiter la qualité des données capteurs en temps réel. Cette méthodologie est composée de trois phases principales : *la définition, l'évaluation et la communication*.

- **La phase 1 : la définition**, consiste à analyser et à définir les facteurs qui impactent la qualité de données issues de capteurs dans le contexte des systèmes de surveillance environnementaux. Nous représentons de manière formelle les critères et indicateurs de qualité optimales pour modéliser la qualité de données issues de capteurs. Ces critères et indicateurs issus du contexte d'acquisition, de la donnée elle-même et du traitement, seront gérés dans des métadonnées qualité et basés sur une approche orientée utilisateur pour leur spécification. Ces informations obtenues aboutissent à un modèle de qualité, qui comprend notamment des critères en termes d'exactitude, fiabilité et temporalité.
- **La phase 2 : l'évaluation**, consiste en la spécification des critères qualité proposés dans la phase 1, en se basant sur des fonctions mathématiques et l'agrégation pour leur calcul. Une analyse et une structuration de

⁵ UML – Unified Modeling Language

métadonnées qualité sur les données capteurs est proposée [Gutiérrez'08], [Gutiérrez'09], afin de gérer les informations qualité en temps réel. Nous décrivons également dans cette phase, l'approche d'évaluation de la qualité de données capteurs en temps réel. Cette approche se repose sur l'évaluation multicritères et l'exploitation des métadonnées associées.

- **La phase 3 : la communication**, comprend la représentation de l'information qualité de façon la plus adéquate pour son exploitation en temps réel. Nous nous basons sur l'exploitation des informations qualité contenues dans des métadonnées qualité, pour fournir des rapports qualité aux utilisateurs, et pour visualiser en temps réel, à l'aide des indicateurs, les mesures relatives à la qualité des données exploitées.

1.3 Organisation du manuscrit

Ce mémoire s'articule autour de trois parties organisées de la façon suivante.

La première partie est dédiée à l'état de l'art, laquelle présente les concepts et définitions touchant les deux grands domaines de notre recherche, la qualité de données et les réseaux de capteurs dans les systèmes de surveillance environnementaux.

Le chapitre 2, illustre le principe de la qualité, ainsi que les diverses approches et techniques existantes utilisées pour déterminer la qualité de données dans les systèmes d'information, notamment dans le domaine géographique. Nous détaillons également les spécificités des métadonnées qui représentent de nos jours, la source d'information qualité la plus couramment utilisée dans ce domaine. Nous abordons à la fin de ce chapitre, les concepts concernant la qualité de service, vis-à-vis de la qualité de l'information et leur apport à notre recherche.

Dans le chapitre 3, nous passons en revue les concepts concernant les réseaux de capteurs et les spécificités de leurs données, en nous intéressant plus particulièrement aux réseaux dédiés à la surveillance de phénomènes environnementaux.

La deuxième partie de ce mémoire est composée par les chapitres 4, 5 et 6 qui représentent nos contributions.

Dans le chapitre 4, nous décrivons les spécificités des systèmes de surveillance environnementaux, sur lesquelles nous avons basé notre approche. Ici, nous détaillons les caractéristiques des données capteurs et notre proposition concernant notamment, leur structuration et leur gestion. Un modèle et une formalisation des données capteurs dans les systèmes de surveillance environnementaux sont proposés.

Le chapitre 5, présente une méthodologie pour la définition de la qualité des données issues qui propose à l'utilisateur un moyen d'évaluer et visualiser la qualité des données issues de capteurs en temps réel. Nous détaillons dans ce chapitre les différentes phases de la méthodologie proposée. Les trois phases proposées s'adaptent

aux spécificités des données capteurs et les besoins de l'utilisateur exploitant un système de surveillance en temps réel.

Ensuite, le chapitre 6, détaille les caractéristiques du prototype MoSDaQ (Monitoring Sensor Data Quality), qui consiste à représenter les informations qualité dans un environnement de surveillance volcanique en temps réel.

Finalement, nous concluons et donnons quelques perspectives à ce travail de recherche dans la dernière partie de ce mémoire.

Première partie :
État de l'art

Chapitre 2

La qualité : données, métadonnées, services

La qualité est un concept souvent associé à des aspects comme les produits, les services, les données, etc. Quand nous parlons de qualité, nous pensons souvent au niveau de satisfaction que nous pouvons avoir avec les bonnes caractéristiques d'un produit, d'un service ou de l'information. C'est-à-dire l'adéquation aux besoins ou le « *fitness for use* » [Veregin'99].

Cependant, certains aspects que nous expérimentons chaque jour dans le monde réel engendrent diverses inquiétudes en termes de qualité. Le fait d'avoir une bonne ou une mauvaise qualité nous pousse à réfléchir sur les différentes causes possibles de cette dégradation de la qualité, et comment fournir aux utilisateurs des informations qualité qui lui permettent de prendre des décisions de façon plus fiable.

Dans ce chapitre, nous introduisons les principaux concepts liés à la qualité de données et les différents éléments nécessaires pour la mesurer, l'évaluer et la communiquer. Nous abordons dans la Section 2.1, la qualité de données au sein des systèmes d'information, notamment dans le domaine géographique (Section 2.2). La Section 2.3 décrira les métadonnées, lesquelles représentent de nos jours, la source d'information qualité la plus utilisée pour l'évaluation de la qualité de données.

Nous décrivons également les concepts essentiels concernant la Qualité de Service (QoS), afin de compléter l'aperçu général sur les différentes façons de gérer la qualité (Section 2.4). Nous concluons ce chapitre dans la Section 2.5.

2.1 La Qualité de données dans les Systèmes d'Information

Dans la littérature, la qualité des données est souvent définie comme « *les données qui sont adaptées à l'usage des consommateurs de données* » [Wang'96] ou plus formellement comme énoncé dans la norme ISO 8402 [ISO'94], « *la qualité représente l'ensemble des propriétés et de caractéristiques d'un produit ou d'un service qui lui confèrent l'aptitude à satisfaire des besoins exprimés ou implicites* ». Toutefois, pendant des années, la qualité des données a été perçue comme étant une simple exactitude. Pourtant la qualité représente plus que ça, elle est un concept à multiples aspects, dont différents éléments (i.e. dimensions, attributs, critères...) sont utilisés pour la décrire.

Actuellement, le domaine de la qualité des données fait partie intégrante de différents sujets de recherche, la statistique [Lohningen'99], la représentation de connaissances [Davis'93], la fouille de données [Gama'07], [Gaber'05], la gestion des systèmes d'information, et l'intégration des données [Lenzerini'02], ainsi que divers domaines d'application (militaire, environnemental, sciences de la vie, médical, etc.) cherchant à améliorer la qualité de données qu'ils produisent (Figure 1).

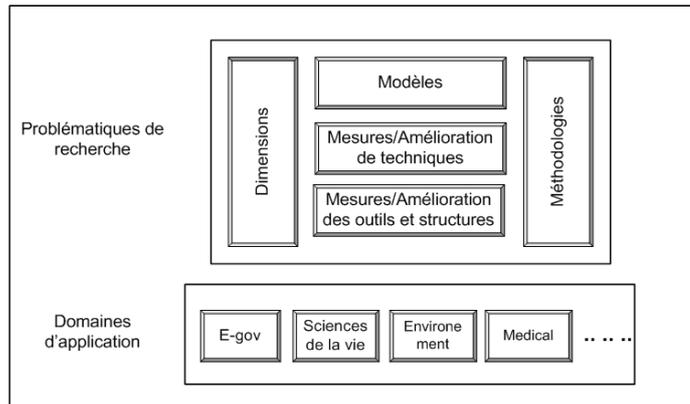


Figure 1 - Principales problématiques de la qualité des données (adapté du [Batini'06])

Cet intérêt a abouti à la définition de différentes méthodologies d'évaluation et d'amélioration de la qualité des données [Batini'09], lesquelles incluent des méthodes et des techniques adaptées aux besoins en termes de qualité (i.e. règles, dimension, modèles, processus, etc.).

Dans le cadre de ces méthodologies, la qualité peut être orientée vers divers systèmes d'information génériques comme, les systèmes d'information *monolithiques*, les *entrepôts de données*, les *systèmes distribués*, les *systèmes coopératifs*, *peer-to-peer*, etc., [Batini'06]. Dans ces différents systèmes, les données peuvent être des *données structurées* (i.e. tables relationnelles) ou/et *semi-structurées* (i.e. schéma XML) ou/et *non-structurées* (données en langage naturel), et sont principalement stockées dans des bases de données [Berti-Equille'04].

Les données peuvent être également caractérisées par leur niveau de traitement, par exemple, les *données brutes*, les *données composantes* et les *produits d'information* [Wang'98]. On retrouve également des données qui sont *stables*, *variantes à long terme* (données qui changent à une base fréquence), ainsi que des *données hautement changeantes* (données qui présentent un changement intensif) comme les mesures d'un capteur de température, l'information routière, etc., [Peralta'04]. Celles-ci sont caractérisées par la variation dans le temps (i.e. flux de données).

Ces spécificités (i.e. structure, variation...) dans les données et dans les systèmes d'information caractérisent de façon différente leur qualité. Par exemple, dans les systèmes monolithiques, le contrôle de la qualité de données est beaucoup plus simple en raison de l'homogénéité et la stabilité des données, et également grâce à la centralisation des processus. Cependant, dans le cadre de systèmes d'information plus

complexes et hétérogènes tels que les systèmes distribués, géographiques, ou pair-à-pair, la gestion de la qualité des données devient plus compliquée.

De façon générale, nous remarquons que la problématique de la qualité de données concerne principalement les dimensions qualité (i.e. précision, exactitude, cohérence, etc.), les modèles, les techniques, les outils, ainsi que les méthodologies adaptées aux nouveaux types de données et systèmes d'information :

- Les *dimensions* sont normalement appliquées dans les modèles qualité, ainsi que dans les techniques, les outils et les structures.
- Les *modèles* généralement utilisés dans les bases de données, ont été enrichis afin de représenter les dimensions et autres aspects liés à la qualité des données.
- Les *techniques* sont un ensemble d'algorithmes, d'heuristiques, et de processus pour répondre à un problème spécifique sur la qualité des données.
- Les *méthodologies* fournissent des directives pour choisir, à partir des techniques et des outils existants, la mesure de la qualité de données la plus efficace pour améliorer un système d'information spécifique.
- Les *outils*, qui sont nécessaires aux techniques et aux méthodologies, représentent des processus automatisés avec une interface qui permettent à l'utilisateur l'exécution manuelle de certaines techniques pouvant être intégrées dans des structures (*frameworks*).

De nos jours, divers travaux [Pipino'02], [Redman'01] et standardisations (i.e. [ISO'02]) ont abordé ce sujet, et ont formalisé des dimensions, des indicateurs, des techniques et des procédures pour une évaluation efficace de la qualité des données.

Dans les sections suivantes, nous détaillons les concepts et les travaux liés à ces aspects.

2.1.1 Dimensions et caractéristiques de la qualité de données

Les travaux proposés jusqu'à aujourd'hui, s'accordent sur le fait que la qualité de données peut être décrite en termes de facteurs, critères, dimensions, catégories, attributs, etc. [Wang'93], [Carter'94], [Wang'96], [Redman'01], [Brodie'80], [Berti-Equille'04] lesquels représentent les différents aspects qui définissent une caractéristique particulière ou qui forment la qualité de données elle-même.

Les *dimensions* qualité peuvent être considérées comme une extension de la donnée, et sont souvent définies de façon qualitative [Naumann'04]. En effet, les dimensions ne fournissent pas des mesures quantitatives, et sont donc associées à une ou plusieurs *métriques*. Pour chaque *métrique*, une ou plusieurs *méthodes de mesure* sont fournies en considérant le moment de la prise de mesure, la donnée incluse, le dispositif de mesure, etc. Cependant, dans certains travaux, les termes dimension et métrique ne sont pas différenciés ou seules les métriques sont utilisées pour quantifier la qualité [Piattini'00].

Il existe dans la littérature un nombre considérable d’approches définissant les différentes dimensions au sein d’un modèle qualité (i.e. [Wang’96], [Redman’97], [Kahn’02]). Nous distinguons deux types d’approches : les approches génériques qui sont liées aux systèmes d’information en général et les approches spécifiques qui sont liées à des systèmes d’information dans des domaines bien précis.

2.1.1.1 Les approches génériques

Ces approches peuvent être considérées comme théoriques car elles adoptent un modèle formel afin de justifier les dimensions, ou considérées aussi comme empiriques car elles construisent un ensemble de dimensions en accord avec des expérimentations, des entretiens et des questionnaires, ou intuitives car elles se basent sur un sens commun ou l’expérience pratique [Wand’96], [Wang’96], [Redman’97].

a) Les approches théoriques

Les approches théoriques comme celle proposée par [Wand’96], considère le système d’information comme la représentation du système du monde réel. Le monde réel est représenté si un système d’information existe, et est caractérisé par le fait qu’en aucun cas deux états doivent être mis en correspondance au même état dans un système d’information (**RW → IS**). Toute déviation dans cette représentation peut engendrer des déficiences, c’est-à-dire, des contraintes en termes de qualité. Les auteurs distinguent des déficiences, notamment liées à la conception et les opérations du système. Plus spécifiquement, les représentations du système peuvent être *incomplètes*, *ambiguës* ou avec des états *sans signification* et les opérations, peuvent être fausses (*garbling*). En accord avec ces déficiences, cette approche a de plus identifié certaines dimensions sur la qualité : *l’exactitude, la fiabilité, la ponctualité, la complétude et la cohérence*.

b) Les approches empiriques

Au niveau des approches empiriques, notamment dans [Wang’96], les dimensions ont été choisies en interviewant les consommateurs de données. Pour la sélection des dimensions, ils ont proposé autour de 179 dimensions de la qualité des données, parmi lesquelles les auteurs en ont sélectionné quinze.

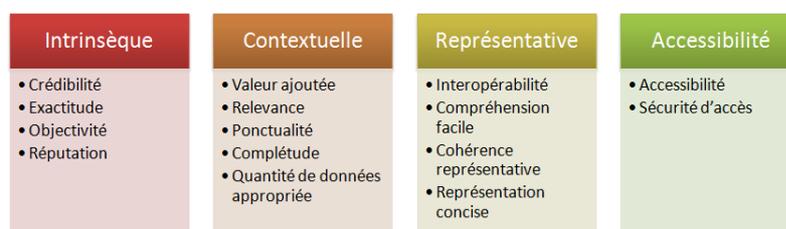


Figure 2 - Dimensions de la qualité de données proposées par l’approche empirique [Wang’96]

Ces dimensions ont été classifiées selon quatre catégories (Figure 2) : qualité de données intrinsèques (la qualité concernant la donnée elle-même), qualité des données

contextuelle (considère le contexte de l'utilisation de la donnée), qualité des données représentative (considère les aspects liés à la qualité de la représentation de la donnée) et qualité d'accessibilité des données (liée à l'accessibilité de données et aux propriétés non-fonctionnelles de l'accès aux données).

c) Les approches intuitives

Concernant les approches intuitives, [Redman'97] a proposé une classification des dimensions de la qualité selon le schéma conceptuel, les valeurs de données, et le format des données. Le schéma conceptuel correspond aux dimensions de la qualité, qui sont liées aux données et sont indépendantes des représentations internes de la donnée. Ces représentations sont gérées par les dimensions dédiées au format des données (Figure 3).

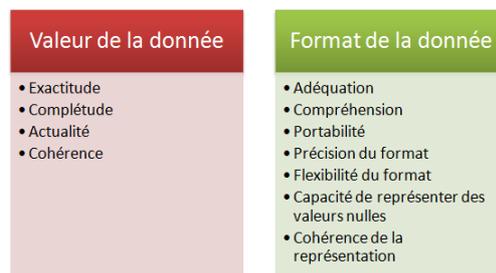


Figure 3 – Dimensions proposées dans une approche intuitive [Redman'97]

La série d'approches proposées dans le domaine des systèmes d'information [Wang'96], [Redman'97], [Wand'96], [Bovee'01], [Berti-Equille'02], que nous avons présentée, nous amène à sélectionner les dimensions citées de façon récurrente (avec certaines divergences) par les travaux dans le domaine. Ces dimensions récurrentes représentent le socle de dimensions utilisées pour caractériser la qualité des données (Tableau 1).

2.1.1.2 Les approches spécifiques

Dans des domaines plus spécifiques (i.e. le web, la statistique, le domaine géographique), des dimensions plus adéquates ont été proposées afin de mieux décrire les spécificités de leurs domaines (*précision géométrique, cohérence temporelle, etc.*). D'autre part, le changement et la mise à jour des données ont conduit à des dimensions liées au facteur temps, c'est-à-dire, la qualité des données stables, variables à long terme, ou encore sur les données hautement « changeables ». Les principales dimensions proposées dans ce contexte sont : *l'actualité, l'instabilité, et la ponctualité.*

Dimension	Description
Exactitude	Liée à « l'inexactitude », ce qui implique que l'état du monde réel est différente de ce qui est représenté.
Ponctualité	Définies le délai entre le changement de l'état du monde réel et la modification résultant de l'état du système d'information.
Complétude	L'habilité d'un système d'information à représenter chaque état du système du monde réel.
Cohérence	La cohérence apparait quand plus d'un état du système d'information correspond à un état du système du monde réel.
Compréhensibilité	Concerne la documentation et les métadonnées qui sont disponibles pour interpréter correctement la signification et les propriétés des sources de données.
Accessibilité	Mesure l'habilité que l'utilisateur accède aux données sans importer leur culture, leur statut/fonctions et les technologies disponibles.
Utilisabilité	Mesure l'effectivité, l'efficacité, la satisfaction avec laquelle les utilisateurs spécifiés perçoivent et utilisent les données.
Crédibilité	(D'une organisation) Mesure combien est fiable l'organisation qui fournis les sources de données.

Tableau 1 – Dimensions récurrentes dans la qualité de données

- *L'actualité* représente la rapidité avec laquelle la donnée est mise à jour ; elle peut être évaluée en considérant la dernière mise à jour (information gardée dans des métadonnées), c'est-à-dire, la dernière fois que la donnée à été actualisée.
- *L'instabilité*, concerne la fréquence avec laquelle la donnée varie dans le temps, et peut être évaluée par la période du temps dont la donnée reste valide.
- *La ponctualité* exprime à quel point les données sont actuelles pour une tâche, mais ceci n'implique pas seulement les données actuelles, mais aussi les périodes de temps pour lesquelles une donnée à été planifiée. Pour mesurer la ponctualité, il s'agit de mesurer l'actualité de la mesure et contrôler si la donnée est disponible pour la période de temps planifié.

Compte tenu de la variété des dimensions proposées dans les travaux existants, nous pouvons déduire qu'il n'existe pas une définition générale de l'ensemble de dimensions relatives à la qualité des données. Comme le cite [Berti-Equille'04], chaque dimension peut prendre du sens selon les spécificités du domaine d'application. Parmi les propositions analysées, nous apprécions le fait que des similarités fortes existent concernant la perception globale de la donnée et les problèmes qualité. Cependant, au moment de la détermination des dimensions dans chaque domaine particulier, les divergences surviennent.

Après avoir présenté les principes des dimensions et des caractéristiques de la qualité de données, nous allons présenter, dans la section suivante, les approches proposées en termes de modélisation de la qualité des données.

2.1.2 Modélisation de la qualité de données

La modélisation de la qualité des données permet de représenter la structure, les dimensions et autres aspects liés à la qualité des données. Les approches de modélisation dans ce domaine, ont été fortement influencées par la représentation que

les données subissent dans un système de gestion de bases de données relationnelles (SGBD), notamment en termes de modèles, de définition des données et de langage de manipulation. La modélisation de la qualité de données prend également en considération les propositions concernant les dimensions associées à la qualité elle-même.

En effet, plusieurs travaux ont repris ces fondements et les ont orientés suivant différents domaines d'application, comme pour les *documents*, le *web*, le *e-business*, les *bases de données*, etc., [Wang'95], [Rocker'04], [Redman'01], [Berti-Equille'02]. Chacune des propositions de ces travaux possède une perspective différente et prend en compte divers aspects pour la caractérisation et l'analyse de la qualité de leurs données, comme par exemple, les approches orientées sémantique [Pipino'02] ou celles orientées traitement de la donnée [Naumann'99] (Tableau 2).

Référence	Type d'approche	Organisation	
		Catégories	Dimensions
Wang et al. 1997	Orientée sémantique -4 catégories -13 dimensions	• Qualité intrinsèque	• exactitude, objectivité, crédibilité, réputation
		• Qualité d'accessibilité	• accès, sécurité
		• Qualité contextuelle	• pertinence, complétude, quantité de données
		• Qualité de la représentation	• concision, cohérence, intégrité, compréhension

Référence	Type d'approche	Organisation	
		Classe	Critère
Naumann et al., 1999	Orientée traitement -3 classes -11 critères	• Spécifique source	• facilité de compréhension, réputation, fiabilité âge
		• Spécifique requête	• disponibilité, prix, consistance de représentation, temps de réponse, exactitude, relevance
		• Spécifique attribut	• complétude, qualité des données

Tableau 2 – Approches générales sur la modélisation de la qualité de données [Wang'96] [Naumann'99]

Afin de modéliser les dimensions liées à la qualité des données, diverses extensions des modèles ont été proposées en accord avec différents types de données existantes (structurées, semi-structurées...).

Concernant les données structurées, l'extension des modèles conceptuels comme le modèle entité-objet ou le modèle relationnel ont été soulevées [Storey '98], [Storey'01]. Des modèles logiques ont également été étendus à la description de la qualité de données avec des valeurs qualité associées à chaque attribut donnant comme résultat un modèle des attributs qualité [Wang'01]. Des modèles pour tracer l'origine de données comme le modèle *Polygen* ont été également considérées dans la littérature [Wang'90].

À l'égard des données semi-structurées, les modèles pour l'association des valeurs qualité aux documents-XML orientés données ont été proposés (i.e. D2Q [Scannapieco'04]). Ce type de modélisation est plus flexible et les valeurs des dimensions de la qualité des données peuvent être associées à divers éléments du

modèle de données (classe de données, classe qualité, fonction d'association qualité). Le modèle résultant peut être traduit en un modèle de données XML et par la suite être interrogé par une extension du langage ad-hoc qui aura accès aux valeurs qualité dans le modèle.

Concernant la gestion de ces modèles, diverses extensions ont été aussi suggérées. Par exemple, l'IP-MAP¹ [Shankaranarayan'00], qui est un modèle graphique basé sur des blocs d'opération (source, consommateur, qualité, traitement...), part du principe que les données peuvent être considérées comme un produit résultat d'un processus quelconque. Ce langage de description a été extensible à une formalisation appelé IP-UML (extension du UML). Ceci permet de créer un profil de qualité (*data quality profile*) qui consiste en trois différents modèles dédiés à l'analyse des données, à l'analyse de la qualité et à la conception de la qualité. Dans ce modèle, le langage UML est utile pour décrire l'interaction entre les données, les processus et l'organisation des éléments, ainsi que pour la définition des profils dédiés à la qualité.

Ces diverses propositions de modélisation fournissent une source de base pour représenter et organiser les dimensions et les mesures de la qualité des données. Elles représentent également une façon de mesurer et améliorer les profils qualité qui sont utilisés pour l'évaluation des ensembles de données, des processus, ainsi que pour des organisations entières.

2.1.3 Méthodologies pour l'évaluation et l'amélioration de la qualité de données

Due à l'évolution des systèmes d'information actuels, les travaux de recherche dans le domaine de la qualité de données se sont focalisés dernièrement, sur la définition des méthodologies qui aident à la sélection, l'adéquation et l'application des techniques d'évaluation et d'amélioration de la qualité de données [Batini'09]. Une méthodologie dans ce contexte, peut être considérée comme l'ensemble de directives et des techniques décrivant le contexte d'application de l'information, ainsi que la définition de processus pour évaluer et améliorer la qualité des données.

Les différentes méthodologies existantes [Batini'09], [Berti-Equille'04] divergent selon les dimensions utilisées, les diverses phases ou étapes méthodologiques, ainsi que les stratégies et techniques utilisées. Ces aspects ajoutés à la diversité des types de données et des systèmes d'information font que chaque méthodologie est adaptée à son domaine et type d'application.

Comme le montrent [Batini'09] dans leur analyse comparative, les méthodologies existantes, par exemple, TDQM² [Wang'98], TIQM³ [English'99], AIMQ⁴ [Kahn'02], DQA⁵ [Pipino'02], IQM⁶ [Münzenmaier'02], CIHI⁷ [CIHI'09]

¹ IMAP – Information Product Map

² Total Data Quality Management

³ Total Information Quality Management

⁴ A Methodology for Information Quality Assessment

⁵ Data Quality Assessment

⁶ Information Quality Measurement

présentent différentes spécificités orientées d'un côté vers l'évaluation et l'amélioration de qualité des données, de l'autre vers l'audit de la qualité de données.

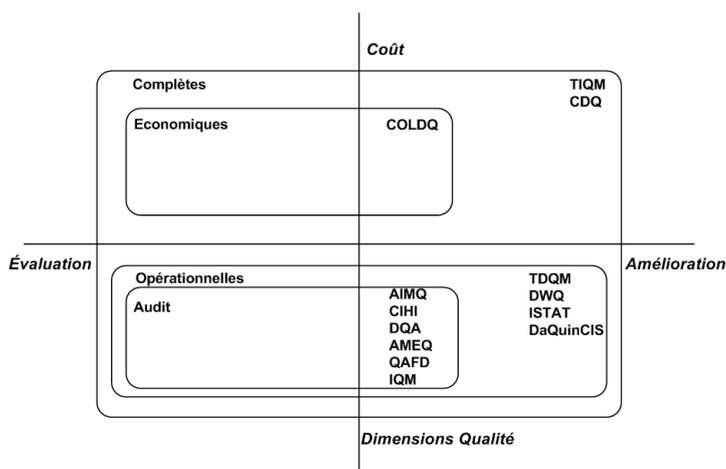


Figure 4 – Classification de méthodologies [Batin'09]

D'autres méthodologies se focalisent plus sur des problèmes techniques liés à l'évaluation et l'amélioration de la qualité, ou encore se concentrent sur les coûts des évaluations (Figure 4).

2.1.3.1 Structuration des méthodologies pour la qualité des données

Généralement, les méthodologies appliquées à la qualité des données comprennent *une phase de reconstitution, une phase d'évaluation et une phase d'amélioration* [Batini'06]. L'objectif de ces étapes est de rassembler toutes les informations contextuelles liées aux collections des données, leurs processus de gestion, les problèmes de qualité, les besoins, ainsi que les coûts.

➤ Phase de reconstitution

Cette phase est dédiée principalement à la collection d'information contextuelle sur les processus, services, acquisition de données, gestion, etc. Cette étape est souvent optionnelle car certains systèmes d'informations possèdent des études approfondies sur les caractéristiques et processus autour des données. Les modèles qualité, avec leurs dimensions et métriques sont le résultat de cette phase.

➤ Phase d'évaluation de la qualité des données

Dans cette phase, les ensembles de données sont évalués selon des dimensions qualité relatives aux domaine et besoins. Implicite dans cette phase, nous retrouvons divers pas qui amènent à l'évaluation de la qualité des données :

⁷ Canadian Institute for Health Information Data Quality Framework

- *L'analyse de données et des besoins qualité*, dont les données sont analysées afin de les comprendre et pour identifier les problèmes;
- *L'identification des secteurs critiques*, dont la sélection des sources de données à évaluer est effectuée ;
- *L'identification des processus*, produisant et gérant les données ;
- *L'évaluation de la qualité*, dont les dimensions de la qualité affectées par les problèmes qualité sont identifiées et la définition des métriques adéquates est effectuée. Cette évaluation, peut être réalisée de façon *objective* (métriques quantitatives) ou *subjective* (métriques qualitatives).

➤ Phase d'amélioration de la qualité des données

Cette phase consiste en la sélection de stratégies, processus et techniques pour cibler de nouvelles contraintes en termes de qualité. Dans cette phase, l'utilisation des métadonnées est souvent mise en évidence, car les métadonnées nous permettent de stocker des informations complémentaires pour comprendre et évaluer les données. Pour améliorer la qualité des données, des aspects comme l'identification des erreurs et leurs causes, des techniques de contrôle de qualité, etc., sont prises en compte.

2.1.3.2 Techniques d'évaluation au sein des méthodologies pour la qualité de données

Parmi les approches existantes de l'évaluation de la qualité des données décrites auparavant, certaines se basent sur des techniques diverses comme l'utilisation des métadonnées [Dasu'03], [Berti-Equille'02], [Akoka'07], l'intégration des données [Scannapieco'04], l'utilisation de la statistique [Bovee'01] et la fouille de données [Redman'97], ou encore sur les techniques correctives utilisées dans des bases de données [Bouzeghoub'04] pour évaluer les données.

Les approches qui exploitent les métadonnées (i.e. TDQM), exploitent de cette façon l'information nécessaire pour comprendre et/ou évaluer les données. Ceci facilite le processus d'estimation et la maintenance de la qualité des données, ainsi que la sélection de stratégies et techniques qui peuvent l'améliorer.

D'autres approches reposant sur la caractérisation des sources et l'intégration des données, exploitent les préférences sur les sources de données afin d'homogénéiser l'information selon le degré de confiance des sources [Scannapieco'04], [Berti'99]. D'autres approches encore s'appuyant sur des techniques statistiques ou de fouille de données exploratoire, permettent de mesurer la qualité des données en proposant des résumés et des visualisations pour détecter les problèmes dans un ensemble de données, avant l'exécution d'analyses plus coûteuses [Bouzeghoub'04].

L'ensemble des concepts décrits dans cette section représentent un aperçu des travaux existants sur la qualité de données dans les systèmes d'information. Nous retenons dans les principales approches existantes les éléments pris en considération,

notamment la description générique d'une méthodologie, les processus et les techniques utilisées pour l'évaluation de la qualité, etc. (Figure 5).

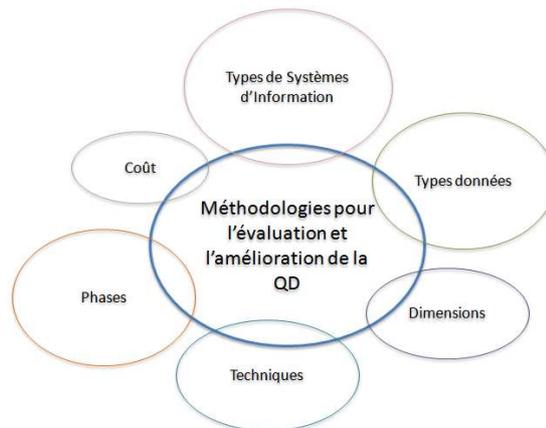


Figure 5 – Éléments dans les méthodologies pour l'évaluation et l'amélioration de la qualité des données

Cependant, d'autres propositions ont été réalisées au sein de domaines particuliers comme dans le domaine géographique qui possède des spécificités plus particulières dans ses données. Dans la section suivante, nous allons décrire les principaux aspects qui caractérisent la qualité des données dans ce domaine.

2.2 La Qualité de données dans les Systèmes d'Information Géographiques

L'intérêt pour la qualité des données a touché tous les secteurs qui exploitent l'information géographique ou géospatiale [Bédard'95], autant pour un conducteur d'une ambulance utilisant un GPS pour retrouver l'endroit de l'accident, que pour un urbaniste qui réalise des mises à jour sur une carte d'une ville, ou pour une personne qui cherche à calculer la distance à parcourir de sa maison jusqu'à l'hôpital le plus proche en utilisant Google Maps⁸. Depuis 20 ans, on constate que la qualité a été traitée et prise en compte dans le domaine de l'information géographique et on ne peut plus ignorer son impact dans le fonctionnement des Systèmes d'Information Géographique (SIG).

Comme dit précédemment, dans le domaine géographique, le terme « Qualité » s'est concentré sur la précision et l'exactitude spatiale des données collectées. Cependant, même dans ce domaine, la qualité est multicritères et est présente dans les processus d'acquisition, de gestion et d'exploitation de l'information géographique. On constate que ces dernières années, les SIGs ont subi des avancées très importantes au niveau des technologies appliquées (i.e. GPS, internet, etc.) et parallèlement, l'analyse de la qualité s'est aussi développée aboutissant à un ensemble d'éléments complémentaires comme, *la cohérence, l'actualité des données, la généalogie, la complétude,*

⁸ <http://code.google.com/intl/fr/apis/maps/>

etc., [Guptill'95]. Ces éléments abordent des incertitudes sur la qualité qui sont importantes dès lors que ces données géographiques vont être utilisées pour la prise de décision dans des tâches complexes, comme l'analyse statistique sur la pollution d'une ville, le trafic routier, etc.

2.2.1 Modélisation de la qualité de données géographiques

La modélisation de la qualité de données du domaine géographique repose sur les spécificités des données dites de façon générale *géographiques* (avec les composants géométriques, descriptives, graphiques et métadonnées). Dans un tel contexte géographique, divers types de données peuvent être distingués selon leurs attributs et leur localisation dans l'espace.

Parmi ces données géographiques, les *données géospatiales* sont des données numériques représentant un territoire quelconque, qui sont ancrées dans un système de coordonnées géographiques ou de projection bien défini, de sorte qu'on puisse les situer exactement sur une portion du globe terrestre (i.e. les orthophotos, les images satellitaires corrigées...) [Devillers'06].

D'autre part, les *données géoréférencées*, sont des données couplant les attributs des données statiques à des données géospatiales via un attribut géographique (i.e. Zone) [Thomas'08].

Finalement, nous retrouvons les *données géocalisées*, qui sont des données ayant une position dans l'espace. Il peut s'agir, dans ce cas, de points, lignes, de surfaces ou encore d'images. Grâce à la géolocalisation, nous pouvons utiliser : des objets ponctuels comme localisation des stations météorologiques, ou un phénomène quelconque (i.e. séisme...); des objets linéaires comme des routes de navigation ou trafic, ou encore des objets représentant une surface comme un département, un pays, etc., [Trevisan'08].

Suite à cette diversité de propriétés qui caractérisent actuellement les données géographiques, la qualité des données géographiques possède à présent, une diversité de concepts, modèles, méthodes et techniques qui la décrivent, mais deux tendances sont principalement perçues [Devillers'07]. Une dédiée aux caractéristiques internes de la donnée, comme les propriétés de leur acquisition, leur modélisation ou leur stockage, et une autre qui correspond à la satisfaction des attentes des consommateurs ou *l'adéquation à l'usage - fitness for use*. Ces deux cas se résument en la *qualité interne* (produits sans erreurs) et la *qualité externe* (satisfaction des besoins des utilisateurs) [Bédard'95].

Ces deux catégories sont également caractérisées par différents critères. Concernant la qualité interne, les critères ont été proposés par les principales normes traitant des données géographiques comme l'ISO⁹, CEN¹⁰, OGC¹¹, etc. Usuellement ces critères sont classifiés en cinq ou sept paramètres selon le domaine et besoins, comme *la généalogie, l'exactitude spatiale d'attribut, sémantique et temporelle*, ainsi que *la cohérence*

⁹ International Organization for Standardization

¹⁰ European Committee for Standardization

¹¹ Open Geospatial Consortium

logique et la complétude [Guptill'95], [USGS'00]... Selon la littérature, les critères de base actuellement utilisés dans le domaine sont décrits comme suit (Tableau 3).

Critère	Définition
Généalogie	C'est le seul critère qualitatif accepté par la communauté (normé). Ce critère décrit l'origine matérielle des données et des méthodes utilisées pour leur acquisition, ainsi que toutes les transformations subies par les données.
L'exhaustivité ou complétude	Vérifie l'adéquation de la représentation donnée aux utilisateurs. Ce critère analyse l'absence ou présence anormale des objets.
La cohérence logique	Vérifie les correspondances de l'ensemble de données avec les caractéristiques des structures du modèle utilisé (spécifications de saisie).
Précision géométrique	Appelée aussi, précision ou exactitude spatiale, ce critère définit les écarts de valeurs de position entre les données collectées dans la base de données et ceux du terrain nominal ou l'univers de discours.
Précision sémantique	Aussi connue comme la précision des attributs non spatiaux, décrit la différence entre la valeur des attributs non spatiaux et leur valeur réelle.
La précision temporelle	Décrit l'aspect temporel des données, voir la gestion du temps en dates, mise à jour, validité des données, etc.
La cohérence sémantique	Indique la pertinence des objets, les relations et les attributs par rapport aux règles et spécifications selon le modèle sélectionnée.
La qualité spécifique	Permet à l'utilisateur de définir ses propres critères de qualité, si les critères officiels ne répondent pas à leurs attentes.

Tableau 3 – Critères de référence pour la qualité de données géographiques [ISO'02]

Ces critères ont été proposés pour fournir une expression de la qualité et ils sont organisés en deux classes : *quantitatifs* ou *paramètres qualité* lesquels permettent de donner une expression quantifiable de la qualité, ainsi que des critères *qualitatifs* qui permettent la description qualitative de la qualité. Chacun de ces critères sont ensuite composés par des sous-éléments décrivant leurs divers critères.

L'information relative à la qualité de données interne est typiquement contenue dans les métadonnées transmises avec les ensembles de données. Dans ce contexte, le principal objectif des métadonnées est de permettre aux utilisateurs d'évaluer l'adéquation à l'usage de leurs données [ISO'03].

D'un autre coté, la qualité externe est identifiée comme la qualité au sens le plus général. Elle représente l'ensemble des propriétés et des caractéristiques d'un produit ou service qui satisfait les besoins. Dans cet esprit, la qualité externe est souvent définie comme *l'adéquation à l'usage (fitness-for-use)* [Chrisman'83], [Veregin'99].

Cependant, les informations nécessaires pour l'évaluation de la qualité externe manquent dans les spécifications actuelles [Servigne'06], même si certains auteurs proposaient l'agrégation de certaines sources d'information et critères (selon leurs besoins) pour rendre possible leur évaluation [Bédard'95], [Aalders'98].

Dans la section suivante nous présentons plus en détails les métadonnées qui sont un élément clé dans les approches d'évaluation de la qualité des données géographiques.

2.2.2 Métadonnées qualité et données géographiques

Dans le domaine géographique, toute source de données ne se limite pas seulement à son contenu attributaire et géographique, mais elle s'accompagne d'informations caractérisant la source elle-même, c'est-à-dire des données sur les données qu'on appelle *métadonnées*.

Les métadonnées fournissent l'information descriptive sur le producteur, le contenu, la qualité et d'autres caractéristiques d'un objet. Ces objets peuvent être divers comme des plans, des photos, ensembles de données numériques, etc., et peuvent être décrits par une liste d'éléments qui vont les caractériser et les décrire.

L'ensemble des informations contenues dans les métadonnées peut être stocké sous forme de fichier texte, document XML¹², ou comme un registre dans la base de données [Timpf'96]. Cette information doit pouvoir être facilement accessible et partageable, afin de faciliter la découverte, la structuration et l'interopérabilité de l'information.

Dans le domaine géographique, les métadonnées sont décrites par des éléments qui correspondent notamment à la description générale de la donnée, concernant notamment la nature des données, à leur système de projection, l'étendue géographique, la qualité des données, etc. Concrètement, les éléments inclus dans les métadonnées doivent fournir des informations concernant la disponibilité, l'adéquation à l'usage, l'accès et le transfert des données, afin de réduire les risques de mauvaise utilisation de l'information.

Dans les bases de données géographiques, la qualité des données est organisée en utilisant différents niveaux d'agrégation et suivant certaines dimensions associées à un indicateur [Devillers'02]. Pour chacune de ces dimensions, des données qualité et des métadonnées sont associées. Par exemple, les métadonnées orientées dimension géométrique décrivent les primitives géométriques, les objets, les formes, etc. Pour la dimension sémantique, les métadonnées qualité sont liées à la valeur, le domaine, l'attribut et la classe de l'objet. Ainsi, pour la dimension temporelle, les métadonnées qualité dépendent du choix de l'utilisateur mais sont généralement reliées à la temporalité d'un attribut soit à leur existence, leur évolution, etc.

Chaque domaine d'application possède une structure propre de ses données, d'où la nécessité de disposer d'une structure de métadonnées avec un contenu approprié à leur contexte d'utilisation qui permet de partager plus facilement leurs informations.

Face à cette diversité, diverses spécifications et normes concernant les métadonnées ont été proposées, afin de décrire de façon plus uniforme les données spatiales. Parmi elles, le FGDC¹³ (*Content Standard for Digital Geospatial Metadata*), et le standard ISO 19115 – *Geographic Metadata* [ISO'03a], sont les plus notables. Ces travaux de normalisation ont permis la constitution d'annuaires de données géographiques, qui agréent une utilisation plus flexible et optimale de l'information. Par exemple, le NAP¹⁴ se base sur le profil de métadonnées de l'ISO 19115 (lot de

¹²XML - Extensible Markup Language

¹³ Federal Geographic data Committee - <http://www.fgdc.gov/>

¹⁴ NAP - North American Profile

données) et l'ISO 19119 (services), le INSPIRE¹⁵ Metadata, dont un catalogue de métadonnées qui se focalise sur le partage de toute l'information qualité géographique dans un même point d'accès en Europe a été proposé. Cette proposition s'appuie sur l'ISO 19115 et l'ISO 19119. Également, le Dublin Core, résultat du DCMI¹⁶, organisation qui cherche à construire un consensus dans le développement des standards de métadonnées en ligne interopérables et l'ESRI¹⁷-ISO, qui est une interprétation du schéma ISO 19115 avec mes mêmes éléments que l'ISO 19139/19115, mais il utilise différents tags.

D'autres standards de métadonnées se sont plutôt focalisés sur le contenu, comme par exemple, l'ABCD¹⁸ un standard pour l'échange d'observations et collections biologiques, ou encore l'ERESE Standards¹⁹.

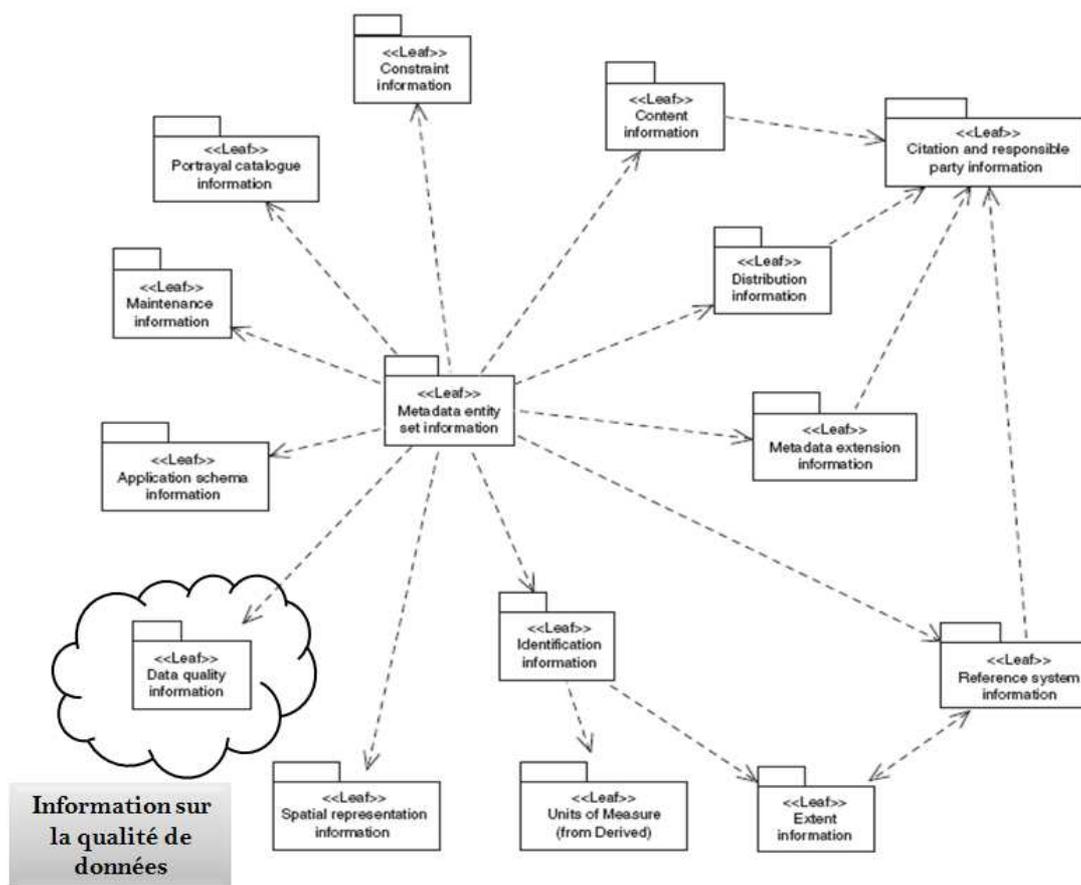


Figure 6 – Diagramme d'entités de métadonnées standardisé par l'ISO 19115 [ISO'03a]

Plus particulièrement, la norme ISO19115 (la plus utilisée de nos jours) fournit une structure pour la description des données géographiques numériques sous la forme

¹⁵ INSPIRE - Infrastructure for Spatial Information in the European Community

¹⁶ DCMI – Dublin Core Metadata Initiative

¹⁷ ESRI - Environmental Systems Research Institute

¹⁸ ABCD - Access to Biological Data Collections

¹⁹ Earth Science Education Metadata Content Standards

d'un modèle conceptuel de métadonnées. Dans cette norme internationale, les métadonnées sont réparties en paquets (packages) (Figure 6), dont chaque paquet contient une ou plusieurs entités (classes UML) qui peuvent être spécifiées en sous-classes ou généralisées (superclasses).

Ce standard est utile car il fournit un langage commun qui permet aux différents acteurs d'échanger des données. Cependant, les métadonnées proposées par ce standard limitent l'utilisation de l'information qualité, vis-à-vis des besoins des utilisateurs. Ces métadonnées sont souvent trop générales et ne permettent pas l'évaluation de la qualité de façon adéquate. Ceci génère souvent des consommations et des applications inutiles des ressources du système.

Parallèlement, ces métadonnées sont typiquement fournies dans des fichiers (i.e. pdf, xml) indépendamment de données, ce qui rend la découverte de l'information plus difficile et la gestion de grands volumes de données restreinte. Cela nuit à la capacité d'exploiter l'information qualité. Dans le meilleur des cas, les métadonnées devraient accompagner les données afin de favoriser la découverte de l'information qualité, et permettre l'évaluation de la qualité de données de façon automatique.

En complément des standards proposés pour les métadonnées dans le domaine géographique, d'autres standards ont été également définis pour la description et la gestion de la qualité de données géographiques. Dans la section suivante, nous présentons les diverses spécificités de ces standards.

2.2.3 Spécification des normes et la qualité de données géographiques

Le début de la normalisation au sein du domaine géographique date des années 80's, fortement encouragé notamment par les secteurs militaires et environnementaux.

Due à la diversité des organisations qui partagent le même type d'information, des standards pour faciliter l'échange d'information géographique entre eux ont été proposés par le comité technique international CEN/TC 287²⁰, en suite par le comité ISO/TC 211²¹ qui a réalisé la concrétisation des standards dans le domaine. L'ISO/TC 211 est actuellement en relation avec diverses organismes comme le FAO²², le ISPRS²³ et l'OGC qui proposent des spécifications (i.e. GML²⁴, devenu ISO 19136). On peut constater qu'après ces années, la collaboration entre ces deux organismes s'est vraiment étendue en donnant comme résultat d'importantes standardisations pour la qualité et les métadonnées, entre autres.

Depuis que les applications sont exposées à l'hétérogénéité de sources, il est devenu impossible de déterminer l'adéquation à l'usage de façon commune. Il a été donc nécessaire d'établir la qualité des données spatiales d'une façon objective et standard. Comme résultat, un ensemble structuré de normes relatives à l'information sur les objets ou les phénomènes qui sont directement ou indirectement associés à une

²⁰ CEN/TC 287 – Geographic Information

²¹ ISO/TC 211 – Geographic Information/Geomatics

²² FAO – Organisation des Nations Unies pour l'alimentation et l'agriculture

²³ ISPRS – International Society for Photogrammetry and Remote Sensing

²⁴ GML – Geographic Markup Language

localisation terrestre ont été établies. Ces normes peuvent spécifier des méthodes, outils et services pour la gestion des données géographiques (y compris leur définition et leur description), l'acquisition, le traitement, l'analyse, etc.



Figure 7 – Critères qualité compris par la norme ISO 19113 [ISO'02]

Nous pouvons citer en particulier la série des normes ISO19000, notamment :

- ISO 19131, concernant la spécification d'un produit,
- ISO 19113 : 2002 sur les principes de la qualité,
- ISO 19114 : 2003/2005 sur les processus d'évaluation de la qualité,
- ISO 19115 : 2003/2006/2009 sur les métadonnées,
- ISO 19139 : 2007 métadonnées (schéma XML d'implémentation),
- ISO 19138 :2006 sur les mesures de la qualité.

La norme **ISO 19113** [ISO'02] fournit les paramètres, définitions et composantes de la qualité des données géographiques. Elle est une norme descriptive et prend en considération principalement *la complétude, la cohérence logique, la précision sémantique et de position*, ainsi que la *précision temporelle* comme critères qualité (Figure 7). Ces critères sont également, décomposés en diverses métriques ou paramètres qui pourront donner une valeur quantifiable à la qualité. Cette norme fournit les principaux éléments pour décrire la qualité des données géographiques, mais elle n'indique pas comment celle-ci doit être évaluée.

Afin de rendre les résultats de l'évaluation plus objectifs et compréhensibles, il a été important d'utiliser des mesures de qualité standardisées. La norme **ISO 19138** [ISO'06], fournit la structure sur comment les mesures de qualité doivent être construites, ainsi que des ensembles de mesures de base qui peuvent être utilisées dans plusieurs cas. Chaque mesure qualité est ainsi définie par 13 composants (Tableau 4).

L'objectif de cette spécification est de guider le producteur de l'information en choisissant les bonnes mesures qualité pour la communication de la qualité des données

et pour guider aussi l'utilisateur à l'évaluation des lots de données utiles par la standardisation des composantes et des structures.

Composantes des mesures qualité	
Nom	Type de la valeur de la qualité de données
Alias	Structure de la valeur de la qualité de données
Élément de la qualité de données	Identificateur
Sous-élément de la qualité de données	Source
Définition (de la mesure)	Exemple
Paramètre	

Tableau 4 – Composantes des mesures qualité [ISO'06]

Nous pouvons remarquer que dans l'ensemble des normalisations définies pour la qualité de données, notamment concernant l'information géographique, le rôle du producteur autant que celui de l'utilisateur a été pris en compte. Cependant, certains aspects orientés utilisateur sont toujours en manque de bonnes spécifications et d'application, comme c'est le cas de l'évaluation et la communication de *l'adéquation à l'usage* [Servigne'06].

L'évolution dans le besoin d'une évaluation de la qualité nécessite fortement une adaptation de ces normes pour pouvoir assurer la qualité de leur information, comme c'est le cas dans les applications exploitant des informations issues de capteurs. Ceci nous montre que les défis en termes de définition de composantes, méthodes et processus pour l'évaluation de la qualité des données sont un des aspects à attaquer dans les futures SIGs.

Nous décrivons, dans la section suivante, les aspects liées à l'évaluation, plus spécifiquement à l'évaluation de la qualité de données géographiques.

2.2.4 Évaluation de la qualité de données géographiques

La qualité des données géographiques peut être abordée selon différentes perspectives. Elle peut être analysée du point de vue du producteur, qui réalise un premier pas vers l'évaluation de la qualité basé sur les spécifications des données, et également selon le point de vue de l'utilisateur, qui analyse les données et détermine si les résultats sont satisfaisants ou pas, pouvant ainsi les rectifier et les ré-analyser selon ses besoins.

L'évaluation de la qualité des jeux de données consiste à comparer les données qui sont produites aux données attendues. Les données utilisées comme référence pour l'évaluation peuvent être celles du « terrain nominal » [David'06] ou un jeu de référence ou de contrôle. Cette évaluation consiste donc, en l'identification des objets qui représentent les phénomènes dans les jeux de données, et les comparer avec un ensemble de critères reliées aux composantes spatiales, sémantiques et temporelles.

Dans ce sens, comme nous l'avons détaillé dans la section précédente, la norme ISO19114 a été proposée afin d'établir un processus adéquat et standard pour

l'évaluation de la qualité des données dans le domaine géographique. L'ensemble de concepts, méthodes et techniques proposés au sein de cette norme sont la référence pour la plus grande partie des travaux liés à l'évaluation de la qualité des données spatiales [Devillers'06].

Selon cette norme, la méthodologie d'évaluation de la qualité de données typiquement utilisée dans ce domaine, est classifiée en une évaluation directe (interne et externe) et indirecte. L'évaluation directe interne utilise seulement les données d'un ensemble de données et l'évaluation directe externe utilise des données de référence externes pour leur évaluation. D'un autre côté, l'évaluation indirecte est basée sur des connaissances externes ou issues de recherches.

L'utilisation de l'évaluation de la qualité se fait pendant et après la collection et la maintenance d'un jeu de données. Mais aussi, les processus d'évaluation de la qualité peuvent être utilisés pendant le développement des spécifications ; ceci veut dire qu'une spécification doit contenir les besoins en termes de qualité des données et que l'évaluation est utilisée pour tester la validité des besoins.

Également, pour déterminer la qualité et éviter que l'utilisateur réalise les mêmes estimations à plusieurs reprises pour analyser si un jeu de données répond aux besoins de son application, l'information sur la qualité doit être stockée et structurée afin d'être plus facilement accessible et partageable. Ainsi, ces informations devront être communiquées à l'utilisateur de façon qu'il puisse les comprendre et les utiliser pour prendre une décision à posteriori.

Dans plusieurs travaux dans le domaine géographique [Hunter'01] [Yong'08], nous trouvons une méthodologie générale pour l'évaluation de la qualité des données géospatiales. Dans cette méthodologie, diverses phases de traitement sont distinguées : le *prétraitement*, le *traitement*, l'*analyse*, la *communication des résultats*. Dans ce cadre, la méthodologie consiste à évaluer la qualité de données brutes, notamment concernant l'exactitude des attributs et de la position. Ensuite, une autre analyse est effectuée selon ces mêmes critères, mais dans la phase de prétraitement, c'est-à-dire, lorsque les données sont corrigées ou éditées. Également, une procédure similaire est appliquée lors de l'analyse des résultats obtenus.

De plus, pour aider les utilisateurs à mieux comprendre et à exploiter leurs informations qualité, diverses techniques de visualisation (visualisation 3D, indicateurs...) sont utilisées pour exprimer les incertitudes dans les données de façon intuitive et effective.

Dans la section suivante, nous allons détailler les aspects et les différentes propositions concernant la visualisation et la communication de la qualité de données géographiques.

2.2.5 Communication et visualisation de la qualité des données géographiques.

La qualité des données est un aspect fondamental pendant tout le processus de production des données géographiques, en débutant par les besoins des utilisateurs et en concluant par la communication de l'information aux utilisateurs finaux.

L'information sur la qualité des données a pour objectif de permettre aux utilisateurs de déterminer dans quelle mesure les données peuvent répondre à leurs besoins [Agumya'97]. Pour cela, depuis plusieurs années différentes manières de communiquer l'information sur la qualité des données géographiques ont été proposées dans la littérature [Devillers'04], [Bastini'02], [Faiz'99], [Laurini'94], [Fisher'94], [MacEachren'92].

La diffusion de métadonnées est, de nos jours, le moyen le plus utilisé pour communiquer l'information sur la qualité, d'ailleurs la norme ISO 19115 est la plus répandue dans le domaine géographique. Les informations contenues dans les métadonnées peuvent être par la suite, communiquées en forme de rapports en fichiers numériques, etc. Cependant, les métadonnées restent limitées, car leur représentation et gestion devient difficile avec les nouveaux systèmes d'information [Gervais'06].

La visualisation est actuellement exploitée comme une méthode pour capturer, interpréter et communiquer l'information qualité aux utilisateurs des systèmes d'information géographiques, notamment en utilisant la visualisation des indicateurs qualité (représentation des mesures qualité) [Huth'07], [Beard'91], [Maceachren'92].

Dans ce cadre, les techniques de visualisation doivent permettre l'affichage de l'information sur la qualité à différents niveaux de détail et d'indiquer clairement les niveaux de données présentées. Ainsi, il est possible d'avoir un accès efficace aux informations sur la qualité et assurer leur association avec les données. Ces travaux utilisent généralement des images juxtaposées, des images composites ou présentées en séquence, la visualisation en 3D, et plus récemment des services de cartographie comme GoogleEarth ou Google Maps, etc. [MacEachren'92].

La qualité de l'information varie de façon spatiale et temporelle, et donc les outils visuels permettent l'affichage de la qualité selon ces variations en facilitant l'utilisation du système. Dans cet esprit, des approches plus récentes visent à identifier les variations spatiales et temporelles [Hunter'01]. D'autres comme celle de [Fisher'99], [Devillers'07], permettent de montrer la qualité des données à différents niveaux de détails et six différentes paramètres qualité. Leur affichage est fait dans un plan principal ou avec un tableau de bord sur la qualité associé au plan. D'ailleurs l'information géospatiale peut être aussi visualisée dans des plans cartographiques ou peut être placée de façon indépendante du plan.

Comme nous l'avons vu dans les sections précédentes, de nombreux aspects accompagnent la notion de qualité autour des données, spécifiquement des données géographiques. Lorsqu'on analyse la qualité, nous ne pouvons pas laisser de côté la qualité en termes de service.

Après une revue des différentes caractéristiques de la qualité des données géographiques, dans la section suivante, nous allons analyser les différentes spécificités

de la qualité de service (QoS). Ceci nous permettra de mettre en évidence l'importance des caractéristiques de la qualité de service par rapport à la qualité des données, notamment dans l'exploitation des systèmes d'information plus applicatifs comme les systèmes d'information environnementaux.

2.3 La qualité de service – QoS

La qualité de service (QoS - Quality of Service) est un ensemble de caractéristiques liées au comportement collectif d'un ou de plusieurs objets, afin de déterminer l'utilité d'un service dans un contexte spécifique d'application [ISO'99], [CISCO'01]. En terme général, la qualité de service est décrite par un certain nombre de concepts applicables aux aspects informationnels et fonctionnels de la qualité de service (Figure 8). Par exemple, les caractéristiques de la qualité de service, les règles, les politiques et fonctions de gestion, les mécanismes qui vont réaliser ces fonctions, ainsi que les catégories de qualité de service qui correspondent aux différentes prescriptions des utilisateurs sont imposées par leur domaine d'application.

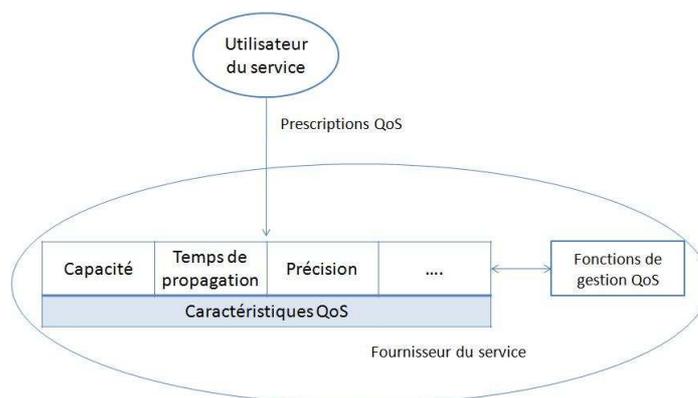


Figure 8 – Concepts de la qualité de service (QoS) [ISO'98]

D'une façon générale, la qualité de service regroupe l'ensemble de fonctions de traitement et de stockage de l'information. Celle-ci est menée par des entités, des objets, des applications, des processus d'application, etc. Elle comprend les interactions entre ces éléments et entre les informations stockées dans le système, les services de communication et l'utilisation d'équipements.

2.3.1 Modélisation de la Qualité de Service

La Qualité de Service (QoS) est un terme assez utilisé mais avec différentes significations et perspectives. Diverses communautés distinguent et interprètent la QoS de différentes façons, comme pour les services web, les réseaux de capteurs, le multimédia, etc., [Ganz'04]. Nous nous intéressons plus particulièrement à la qualité de service dans un environnement de technologies de l'information [ISO'98].

En raison du grand nombre de domaines d'application existants, les utilisateurs doivent appliquer des règles et politiques spécifiques à leur domaine, lors du traitement et de la communication d'information. Ces règles et politiques seront déterminantes sur le choix de l'ensemble de caractéristiques du QoS qui vont permettre de satisfaire les besoins. Une *catégorie* de QoS permet d'analyser les attentes de certains types d'utilisateurs, en organisant au mieux leurs besoins concernant la qualité de service. Certaines des catégories proposées, sont considérées d'intérêt général pour les communications et le traitement : les *catégories liées au temps, à la cohérence, à la capacité, à l'intégrité, à la sûreté, à la sécurité et à la fiabilité*, entre autres. Ces caractéristiques sont fonction notamment du type d'application et concernent principalement les paramètres suivants :

- *Paramètres temporels* : temps de transfert, temps de réponse, temps d'établissement d'une connexion.
- *Paramètres volumiques* : quantité de données qu'une application souhaite transmettre par unité de temps.
- *Paramètres d'erreurs* : pertes de paquets, livraisons tardives ou désordonnées des paquets, erreur d'établissement ou de fermeture de connexion.
- *Paramètres de fiabilité* : la disponibilité et la tolérance aux fautes.
- *Paramètres de sécurité* : degré de protection, de contrôle d'accès, authentification et confidentialité.
- *Paramètres de coûts* : le prix que l'utilisateur est prêt à supporter pour obtenir la qualité de service demandée.

En général, les *caractéristiques* de la qualité de service sont utilisées pour représenter un ou plusieurs aspects, concernant un service ou une ressource d'information identifiable et quantifiable (équipement, communication,...) Ces caractéristiques indiquent l'état réel dans lequel se trouve l'élément par opposition à un paramètre de mesure ou de commande. Les caractéristiques sont dédiées à la modélisation du comportement des systèmes qu'elles caractérisent, certaines d'entre elles sont considérées comme étant des *caractéristiques génériques*, d'autres sont plus *spécialisées* (appliquées à des situations spécifiques), et d'autres sont *des caractéristiques dérivées* (fonction mathématique d'une autre caractéristique). Les valeurs que peuvent représenter les caractéristiques du QoS ne concernent pas seulement les nombres (booléennes, réels, complexes. . .), si non aussi les vecteurs, les matrices, les rangs, etc. [ISO'98].

2.3.2 Gestion de la Qualité de Service

Les besoins des utilisateurs sont explicités par une entité qui demande à utiliser un service et ces besoins sont formulés sous la forme d'une ou plusieurs *prescriptions*. Ces prescriptions peuvent être exprimées en même temps sous forme de *paramètres* (lorsqu'elles sont adressées entre entités) et sous forme d'un *contexte de qualité de service* (lorsqu'elles sont conservées dans une entité).

La gestion de la qualité de service désigne les fonctions dédiées à aider à satisfaire une ou plusieurs prescriptions QoS de l'utilisateur. Cette gestion est réalisée par un certain nombre de *fonctions de gestion de la qualité de service - QMF*²⁵ afin de répondre aux besoins des utilisateurs et des applications. Ces fonctions sont composées par un certain nombre d'éléments appelés *mécanismes*.

Ces mécanismes sont destinés à satisfaire une ou plusieurs spécifications QoS, actionnées généralement par les prescriptions de la QoS, ou qui lui sont communiquées sous la forme du contexte QoS. Les *activités* comprises dans ce contexte sont par exemple, la surveillance, la maintenance, l'alerte, l'établissement, la régulation, l'enquête de la QoS.

La gestion de la qualité de service nécessite différentes fonctions aux différents stades d'une activité quelconque. La gestion peut donc être réalisée aux stades : *avant lancement d'une activité, au moment du lancement de l'activité et pendant l'activité*. Par exemple avant le lancement de l'activité, indiquer les mécanismes qui peuvent être nécessaires pour atteindre l'objectif. Ensuite, négocier les prescriptions entre les utilisateurs et le fournisseur du service, et finalement, considérer la modification des prescriptions au cours de l'activité si besoin.

Les différentes *fonctions* appliquées à la gestion de la qualité de service, peuvent être liées à des prescriptions ou des besoins qui restent constants pendant toute l'activité, comme c'est le cas des fonctions : *la spécification et la négociation de QoS*, ainsi que *le contrôle d'admission et la réservation de ressources* [Shenker'97].

D'un autre côté, il existe des aspects dans la gestion de la qualité de service qui changent dans l'environnement, comme les fonctions de *monitoring, contrôle, la maintenance et la renégociation de la QoS* [Zhou'05].

2.3.3 La qualité de service et la qualité de service spatiale

La Qualité de Service est un terme aussi utilisé dans le domaine spatial. Dans le contexte du service orienté géo-processus, la QoS est une série de « *qualités* » attribuées aux collaborations collectives d'un ou plusieurs géo-services. La QoS représente donc l'ensemble des caractéristiques d'un service donné par la collaboration de géo-services qui renseigne sur l'efficacité à satisfaire les besoins implicites dans le contexte spatial. Plus spécifiquement, la qualité de service spatiale représente un ensemble de qualités apparentées au comportement opérationnel d'un service et à la qualité de l'information géographique que le service génère.

L'infrastructure de géo-services rend capable la découverte, liaison et l'exécution de géo-services en conformité avec les conditions de l'utilisateur [Simonis'05]. Dans ce cadre, la qualité de service spatiale fournit des mesures pour la qualité des éléments qui sont basés sur la performance, la disponibilité, la fiabilité, la réponse à temps, l'exactitude et l'exhaustivité des services (Tableau 5).

²⁵ Quality Management Fonctions

Niveau	Dimension de l'information	Dimension opérationnelle
Utilisateur	<ul style="list-style-type: none"> • Exactitude • Intégrité • Fidélité 	<ul style="list-style-type: none"> • Disponibilité • Interactivité • Prix • Sécurité • Dépendance
Application	<ul style="list-style-type: none"> • Éléments de la qualité des données • Modèle de qualité quantifiable 	<ul style="list-style-type: none"> • Fiabilité • Exécution • Coût
	<ul style="list-style-type: none"> • Incertitude 	<ul style="list-style-type: none"> • Sécurité

Tableau 5 – Dimensions orientées qualité de l'information et QoS [Onchaga'04]

Prenons l'exemple du cadre imprévisible d'un événement catastrophique. L'architecture de services pour la gestion de l'information devient très importante. Les services doivent présenter des niveaux de fiabilité et de résilience et une capacité à s'ajuster à différents environnements. Dans [Onchaga'04], Onchaga propose différentes caractéristiques pour la gestion de sinistres ou phénomènes géographiques. Les caractéristiques sont dépendantes de la dimension de la qualité (information ou opérationnelle) et des niveaux d'architecture. Il fait référence aux niveaux *utilisateur* et *application* (Tableau 5).

Nous pouvons constater après cet aperçu concernant la qualité de service, que pour toute activité, les caractéristiques, les composantes, ainsi que les fonctions et mécanismes les plus appropriés pour la mise en œuvre de la gestion de la qualité de service, dépendent du type de domaine et des activités auxquelles elle doit être appliquée. La structuration de la qualité de service reste, de façon générique et ouverte à toute implémentation au sein des différentes applications dont la gestion des services est présente (web, pervasives...).

2.4 Conclusion

Nous avons présenté dans ce chapitre les concepts liés à notre travail de recherche. Nous avons décrit les principales spécificités par rapport aux méthodologies et activités d'évaluation et d'amélioration de la qualité de données au sein des systèmes d'information en général, et plus spécifiquement dans les systèmes d'information géographiques. Un aperçu sur la qualité de service a été aussi donné.

Nous remarquons dans l'ensemble de travaux existants liées à la qualité de données dans les systèmes d'information [Batini'09], [Wang'98], [English'99], [Kahn'02], [Pipino'02], [Münzenmaier'02], leur intérêt pour spécifier des méthodologies ou processus qui aident à l'évaluation de la qualité de données. Ces travaux se basent principalement sur des données dites traditionnelles ou statiques gérées dans des systèmes de gestion de bases de données ou des entrepôts des données, et sont notamment intéressés à l'amélioration en termes de méthodes et techniques qui peuvent aider à mieux évaluer les données et à économiser les ressources lors de l'évaluation. Ces travaux fournissent d'excellents modèles de qualité de base et enrichissent la perception de la qualité vue autant du côté fournisseur que du côté

utilisateur, et ils prennent de plus en plus en considération le facteur temps dans la sélection des dimensions qualité à évaluer. Cependant, ces travaux ne prennent pas suffisamment en compte la dynamique des données.

Dans le cadre plus particulier de la qualité de données géographiques, nous avons mis en évidence l'importance et les avancements que ce domaine a subis ces vingt dernières années. Dans ce domaine, de plus en plus étudié par les nouvelles technologies (i.e. Internet, GPS...), l'intérêt de l'utilisateur pour la qualité a fait émerger des propositions audacieuses qui ont fait évoluer la perception et l'utilité des systèmes d'information géographiques [Chrisman'83], [Laurini'94], [Bédard'95], [Agumya'97], [Aalders'98], [Goodchild'98], [Veregin'99], [Hunter'99], [Devillers'06], [Akoka'07].

Comme nous l'avons vu dans ce chapitre (Section 2.2), des travaux très importants concernant la perception de la qualité des données vis-à-vis de l'adéquation aux besoins de l'utilisateur a été et continue à être très importante [Guptill'95], [Devillers'02], [Onchaga'04]. Un bon nombre de ces travaux de recherche ont abouti à la spécification des standards pour la gestion de la qualité de données géographiques [ISO'02], ces standards très utilisés dans le domaine géographique ont permis aux différentes organisations d'utiliser un langage en commun pour décrire leurs applications et partager d'une façon plus adéquate l'information, l'exploitation. Également, l'implémentation des techniques de représentation visuelle et des interfaces dédiées et mieux appropriées à l'interaction de l'utilisateur avec l'espace géographique ont fait preuve d'évolution.

L'ensemble de ces concepts et approches existants nous ont apporté de pistes précieuses pour notre recherche. Cependant, nous nous sommes aperçue qu'il existe un manque très important en termes de spécification, modélisation et de gestion de la qualité des données géolocalisées (spatio-temporelles) dynamiques, notamment concernant les métadonnées et l'évaluation de la qualité. Dans ce cadre, nous remarquons que les métadonnées restent avec le stigma « *trop d'information ou manque d'information* » fournies à l'utilisateur, et ne sont pas tout à fait appliquées sur des données avec une variation importante dans le temps. Actuellement, très peu des travaux et aucun standard s'attachent à ce sujet en particulier, cependant, un bon nombre d'associations et de projets dans le domaine environnemental soulèvent le besoin et la priorité de ce travail de recherche (i.e. NOAA, MMI, NASA).

Dans le chapitre suivant, nous allons mettre en contexte ces concepts qualité, en présentant les principales caractéristiques des réseaux de capteurs, les données issues de capteurs et leur application au sein des systèmes de surveillance. L'objectif de cette étude est de soulever les besoins en termes de qualité dans ce domaine particulier d'application.

Chapitre 3

Les réseaux de capteurs : description et applications

L'utilisation des capteurs et des réseaux de capteurs est en train de révolutionner la façon avec laquelle l'information géographique est collectée et analysée. Par exemple, les appareils photos, les capteurs GPS embarqués ou fixes ont la facilité de fournir des données sur de vastes informations géolocalisées. L'ancien modèle de capteurs robustes qui collectent des mesures avec des stratégies de déploiement hautement planifiées, est maintenant remplacé par des réseaux de capteurs sans-fil qui collectent des informations variant considérablement dans le contenu, la résolution et l'exactitude.

Dans ce chapitre, nous présentons les caractéristiques des réseaux de capteurs, ainsi que leur fonctionnement, application, et les spécificités de leurs données, notamment dans le domaine de la surveillance de phénomènes environnementaux.

3.1 Les capteurs et réseaux de capteurs

3.1.1 Les capteurs

Dans notre contexte d'étude, nous entendons par *capteur* un dispositif qui mesure ou détecte les conditions du monde réel, comme le mouvement, la température ou la lumière, etc. Suite aux récentes avancées en accessibilité, adaptabilité de leur technologie, et compte tenu du faible coût de déploiement, les capteurs sont de plus en plus utilisés dans le domaine géographique (i.e. géocapteurs [Nittel'09], [Szewczyk'04]). De tels capteurs produisent des données, ainsi que des flux de données (séquences d'objets générés de façon continue) appelés *mesures* (Figure 9) [Delin'00]. Les capteurs peuvent également travailler en collaboration sous la forme d'un réseau.

Actuellement, des milliers de capteurs sont éparpillés dans les milieux industriels et non-industriels, utilisés de façons très diverses comme par exemple, les thermomètres ou baromètres, les capteurs de sécurité (i.e. les détecteurs de fumée), les capteurs de sûreté (i.e. les capteurs de mouvement), etc. Un grand nombre de ces capteurs fonctionnent au sein d'un système d'acquisition, qui est souvent lui-même intégré au sein d'un système de contrôle encore plus large. Dernièrement, ces systèmes sont composés de capteurs de taille minimale avec une autonomie importante et une capacité à stocker, traiter et transmettre des données, comme c'est le cas des *Smart Sensors* [Frank'00] [Kahn'99] ou capteurs intelligents.



Figure 9 – Capteurs au terrain d'observation [Delin'00]

L'acquisition d'informations est ainsi devenue plus intelligente. Les experts ont de moins en moins un contact direct avec le capteur. Ceci permet donc la récupération des mesures dont les conditions d'acquisition sont souvent périlleuses, comme c'est le cas de la surveillance environnementale. Dans ce cadre, les facilités au niveau de la portabilité et la flexibilité des capteurs sont utiles. Par exemple, grâce à la portabilité, les capteurs peuvent être déplacés d'un environnement d'observation à un autre.

Dans notre étude, nous nous sommes intéressée aux caractéristiques physiques du capteur. Dans ce qui suit, nous détaillons dans les spécificités concernant le capteur, les plus importants dans le domaine géographique.

3.1.1.1 Capteurs intelligents

Les *capteurs intelligents* [Culler'04] ont fait l'objet d'un grand intérêt dans des domaines tels que l'industrie, les systèmes de contrôle et surveillance, les applications biomédicales, etc. Ils possèdent peu de contraintes au niveau de la mobilité, sont flexibles et facilement portables. Les capteurs intelligents ou *Smart Sensors* sont définis comme étant « *des circuits intégrés sans composants externes qui incluent des fonctions de captage, d'interface, de traitement de signal et d'intelligence (auto-test, auto-identification, auto-adaptation)* » [Huijsing'94]. Dans ce cas, les dispositifs capteurs les plus autonomes et intelligents actuellement utilisées sont les « *notes* » ou nœuds capteurs. Ces nœuds capteurs sont des dispositifs (i.e. MICAZ [Fries'07], IRIS, TelosB, Tmote) de taille réduite assez autonome et composés par une unité de mémoire et de calcul à base consommation d'énergie, ainsi que par plusieurs capteurs et un transmetteur sans-fil.

3.1.1.2 Mobilité des capteurs

Un de principaux critères pour déterminer l'utilisation des capteurs au sein du domaine géographique est la mobilité et l'autonomie des capteurs. Comme le propose [Noël'06], une classification des capteurs dans le domaine géographique selon le type de mobilité est convenable. Il s'agit de différencier les capteurs suivant des critères de changement de position à savoir capteurs *fixes*, *agiles* et *mobiles*. D'autres travaux

comme ceux de l'OGC (SensorWeb Enablement), distinguent les : *capteurs statiques* (sur place¹ ou à distance²) ou capteurs dynamiques (sur place ou à distance).

Traditionnellement, les capteurs déployés dans le passé étaient régulièrement fixes, lourds et d'une taille considérable. Ils ne pouvaient pas être déplacés facilement, leur coût d'entretien était assez élevé. Le besoin d'utiliser ces capteurs en différentes positions a conduit au développement de capteurs *agiles*. Ces capteurs sont typiquement analogues et déplaçables entre les prises de mesure. Ils sont forcément plus légers, avec une gestion d'énergie plus importante. L'agilité de ces capteurs est définie comme leur tendance à changer de position entre deux prises de mesures, mais de façon discrète, contrairement aux capteurs *mobiles* dont le changement de position s'effectue en continu. Les capteurs agiles modernes sont principalement appliqués à la navigation, ou à la surveillance de la déformation d'un terrain.

Les *capteurs mobiles* sont considérablement réduits en taille et plus autonomes. Ils sont utilisés dans un grand nombre de contextes dont l'acquisition de l'information par rapport aux changements de position en continu est importante. Pour ce type de capteurs, à chaque changement de position, une mesure est collectée.

Pour augmenter la performance, ces types de capteurs sont traditionnellement organisés en réseaux et munis d'un système de communication qui permet à leurs données de rejoindre un serveur. Dans la section suivante, nous introduisons les spécificités des réseaux de capteurs et leurs secteurs d'application.

3.1.2 Les réseaux de capteurs

Un réseau de capteurs consiste en un ensemble des petits dispositifs de calcul, qui sont capables de se communiquer à travers un canal de communication [Akyildiz'02], [Tubaishat'03], [Khemapech'05]. Chacun de ces dispositifs a la facilité de détecter, traiter et communiquer. Ces dispositifs détectent des valeurs physiques dans l'environnement comme la température, la lumière, etc.

Une contrainte forte des réseaux de capteurs concerne les limitations en termes d'énergie, de complexité et coût de calcul. Comme caractéristiques basiques d'un réseau de capteurs, nous pouvons citer leur capacité d'auto-configuration, leur communication et routage, leur déploiement et leur travail coopératif entre nœuds, etc. Actuellement, les réseaux de capteurs sont plus structurés en architectures adaptés aux objectifs des applications. Mais on peut les différencier, de façon générale, selon leur technologie et leur configuration. Par exemple, divers réseaux de capteurs ont été identifiés, tels que les réseaux de capteurs sans-fil³ (WSN), et autres types de WSN plus spécifiques comme les réseaux de géocapteurs⁴, les réseaux de capteurs environnementaux⁵, les réseaux de capteurs-actionneur⁶, etc.

¹ In-situ

² Remote

³ WSN – Wireless Sensor Network

⁴ GSN – Geosensor Network

⁵ ESN – Environmental Sensor Network

⁶ WSAN – Wireless Sensor-Actuator Network

Actuellement, les applications se focalisent plus particulièrement sur l'utilisation des réseaux de capteurs sans-fil [Aboelaze'05], [Mainwaring'02], [Guitton'07]. Ce type de réseaux, travaillent typiquement de deux façons, soit en *mode continu* (ou *pushing*) soit en *mode requête* ou à la demande (ou *pulling*) [Fries'07]. Dans le mode continu, le nœud détecte l'environnement de façon continue et envoie des données (ou des données traitées) au voisinage ou au nœud central. En revanche, dans le mode requête, le nœud est normalement hors tension en attendant une instruction du nœud central, ou d'un nœud du voisinage. Quand le nœud reçoit les instructions, il rassemble les données des capteurs, les traite et les envoie au nœud ou serveur demandeur.

3.2 Les données et flux issus de réseaux de capteurs

Les données issues de capteurs sont différentes des données exploitées dans les applications traditionnelles [Elnahrawy'03a]. On peut distinguer les *flux de données* ou *streaming*, ainsi que les données acquises à la demande ou par un événement déclencheur, que nous appelons *ensemble de données*. Ces données peuvent également représenter une *seule donnée* ou un *échantillon* d'une population entière selon principalement leur fréquence d'acquisition. Les données capteurs sont aussi plus « *bruyantes* » que les données habituelles. En effet, les données dans un dispositif de captage nécessitent la vérification et la validation de ces données ; l'effet environnemental sur les réseaux de capteurs déployés peut aussi jouer un rôle négatif sur les valeurs détectées.

Les données (ensembles, flux ou échantillons) produites par ces réseaux de capteurs ont souvent besoin d'être traitées en temps réel. Ces données changent généralement au cours du temps d'une façon imprévisible et à la volée, représentant des valeurs de températures, des objets en mouvement, etc. Les données obtenues doivent être contrôlées de façon continue et réagir aux événements intéressants. Dans ce contexte, dû à la difficulté de traitement des grands volumes de données en temps réel, les applications doivent faire un choix des données selon des critères spécifiques à l'usage, par exemple, les plus récentes (dernier jour, heure,...).

3.2.1 Traitement de données de capteurs

Les données capteurs peuvent être traitées par des systèmes de gestion de données concernant principalement, l'interrogation, l'ordonnancement et le stockage de flux de données générés par les réseaux [Gama'07], [Babcok'02]. Dans ce cadre, plusieurs systèmes de gestion de données capteurs ont été dernièrement développés, comme le TinyDB [Madhow'06], Cougar [Yao'02], etc., qui font partie des systèmes de gestion de bases de données des réseaux de capteurs SDBMs - *Sensor Network DataBase Management Systems* [Hellerstein'02]. Ces systèmes ont été inspirés des systèmes de gestion de bases de données traditionnelles, et permettent l'interrogation des données historiques, ainsi que les plus récentes [Bonnet'01] [Bonnet'00] du système, D'autres systèmes, comme le STREAM [Arasu'03], Gigascope [Johnson'03]

et Aurora [Abadi'03] sont plutôt dédiés à la gestion de flux de données, permettant ainsi d'interroger des flux de données à haute fréquence en continu. Cet ensemble de systèmes possède des spécificités par rapport aux attributs temporels et spatiaux qui sont beaucoup plus importants dans les données issues de réseaux de capteurs. Le besoin de savoir quand et où une donnée provenant d'un capteur a été acquise est déterminant.

Un aspect important qui est pris en compte pour le traitement de données capteurs est la gestion de l'énergie lors du traitement. Le besoin de trouver de meilleures stratégies a conduit à des solutions de traitements de données en réseau (*in-network*). Ce traitement en réseau est proposé en raison de la difficulté de transmettre tout le temps au point central toutes les données détectées.

Les données issues de capteurs, notamment issues de réseaux de capteurs, sont aussi exposées au prétraitement comme la fouille de données (*Data mining*), notamment concernant la détection d'anomalies, le regroupement (*clustering*), la classification et la prédiction, les séries du temps, etc. [Lian'07], [Palpanas'03]. Le nettoyage des données en temps réel est également nécessaire pour les applications des réseaux de capteurs afin d'économiser des ressources et faciliter l'accès aux données en temps réel [Gama'07].

Concernant le traitement en temps réel des données capteurs, plusieurs études se sont développées afin d'améliorer le traitement et la localisation des données dans les réseaux de capteurs [Elnahrawy'03b].

Traditionnellement, le temps réel est considéré comme « *le respect de contraintes de temps* », c'est-à-dire, que les transactions et l'accès aux données présentent une limite du temps pour être satisfaits. Dans ce cadre, le temps réel est usuellement considéré comme suit [Lam'01]:

- *temps réel souple*⁷ – la réponse doit être satisfaite avant les délais
- *temps réel ferme*⁸ – la réponse dans les délais est essentielle. Le résultat n'est plus utilisé une fois que le délai est dépassé.
- *temps réel dur*⁹ – La réponse dans les délais est vitale. L'absence de réponse est catastrophique, plus qu'une donnée incorrecte.

Le temps réel doit assurer qu'une tâche soit réalisée selon les contraintes du temps spécifiées.

3.2.2 Stockage des données capteurs

Trois types d'approches ont été proposés en termes de stockage, localisation et interrogation des données capteurs [Elnahrawy'03b], [Govindan'02], [Ratnasamy'03]. Premièrement *les approches locales* qui proposent de conserver le maximum des données au niveau local. Avec cette approche, le coût de la transmission et de mises à jour est minimisé, mais le coût pour l'accès aux données est plus grand car

⁷ Soft real-time

⁸ Firm real-time

⁹ Strong real-time

il est nécessaire de propager les requêtes à travers le réseau de capteurs. D'un autre côté, les *approches externes* proposent de centraliser les données dans un serveur ou base de données à l'extérieur du réseau de capteurs. Dans cette approche le coût de transmission et de la mise à jour est plus élevé, mais l'accès aux données est plus fiable. Si un capteur tombe en panne, les données qu'il a transmises à la base centrale seront toujours accessibles. Cette approche est la plus répandue en matière de réseaux de capteurs, et est hautement utilisée dans les systèmes d'information géographiques. Par contre, nous pouvons remarquer que certaines applications combinent ces deux approches, utilisant par exemple une sorte de système de réplication des données, avec les données réparties entre différents points du réseau (nœuds de réplication) et un stockage centré sur les données. Ceci permet de trouver un compromis entre les coûts de requêtes et de transmission.

Certains travaux ont proposé diverses façons d'interroger les données capteurs [Bonnet'01], [Seshadri'95] en fonction de leur localisation. Ces approches traitent les requêtes et accèdent aux réseaux de capteurs de façons séparée. Les données sont extraites du réseau d'une façon prédéfinie et sont stockées dans une base de données localisée sur un serveur, dont les requêtes seront exécutées. Il existe également, des approches plus distribuées, où la requête détermine les données qui seront extraites des capteurs, par exemple les températures entre 20 et 35°C.

Pour la découverte de ces informations, les interfaces déclaratives du type SQL pour que l'utilisateur puisse spécifier les données qu'il veut interroger, sont de grande utilité. Dans ce cadre, nous pouvons citer quelques langages utilisés dans le domaine comme, par exemple, CQL (Continuous Query Language), StreaQuel [Ma'05], PLACE [Mokbel'05], SoCQL [Gripay'09], etc.

Cependant, même si tous ces systèmes et techniques ont donné de bonnes approches pour la gestion et le traitement des données issues de capteurs, certains défis restent d'actualité. Ces défis concernent essentiellement, la gestion d'énergie, l'efficacité de la communication, la taille de mémoire, etc. De ce fait, la plupart des systèmes de surveillance actuellement utilisés se basent sur les approches externes, dont les requêtes sont centralisées et distribuées. Les données gardent donc une certaine sécurité et fiabilité.

3.2.3 Réseaux de capteurs et domaines d'application

Due à la diversité des applications, les besoins dans l'architecture et l'organisation des capteurs, ainsi que de leurs données se diversifient. Par exemple, dans l'industrie automobile, la préférence s'oriente vers des dispositifs micro et nanoélectroniques s'adaptant fortement à leurs besoins (i.e. l'allumage, la climatisation, l'indicateur de vitesse, etc.). Dans d'autres domaines, comme le médical, la préférence pour des équipements reposant sur des MEMS¹⁰, sont de plus en plus proposés [Sergey'03]. Parallèlement, dans le domaine de la surveillance (environnementale, militaire et industrielle), les dispositifs qui permettent une

¹⁰ *Micro-Electrical-Mechanical Systems*

commande à distance et l'utilisation de systèmes de positionnement (i.e. GPS) sont les plus récurrents (réseaux de géocapteurs).

CATEGORISATION DES APPLICATIONS DES RESEAUX DE CAPTEURS	
Catégorisation orienté-objectif	Catégorisation traditionnelle
1. Militaire	• Militaire
2. Sécurité publique/Alertes	• Observation Environnemental et Prévission • Surveillance Santé • Surveillance des Structures
3. Education	• Observation Environnementale et Prévission • Surveillance Santé • Surveillance des Structures • Surveillance de l'habitat • Salle de cours intelligente
4. Amélioration de la compétitivité des affaires	• Suivie (Système d'inventaire) • Bureau intelligent
5. Amélioration de la qualité de vie QoL (Quality of Life)	• Observation Environnemental et Prévission • Surveillance Routière • Bureau/Maison Intelligent

Tableau 6 – Catégorisation d'application des réseaux de capteurs [Karl'03]

Les réseaux de capteurs représentent un domaine de recherche très intéressant qui entraîne le développement de nouvelles applications grâce aux développements de la micro et nano technologie [Pister'02]. Depuis des années, le domaine qui a poussé le développement des réseaux de capteurs a été le domaine militaire [Gama'07]. Cependant, ces dernières années il y a eu une diversification dans d'autres domaines comme la surveillance environnementale [Vasilescu'05], la surveillance de l'habitat [Mainwaring'02], la surveillance des bâtiments [Tubaishat'03], la surveillance en santé [Milenković'06], etc. Chaque application présente des spécificités et besoins particuliers qui déterminent leur conception et déploiement (Tableau 6).

Dans le paragraphe suivant, nous décrivons plus en détail l'application des réseaux de capteurs au domaine de la surveillance environnementale.

3.2.3.1 Les réseaux de capteurs et les systèmes de surveillance environnementaux

Parmi les divers domaines d'application des réseaux de capteurs [Xu'09], nous nous sommes plus particulièrement intéressée aux systèmes de surveillance des phénomènes environnementaux [Nittel'09], [Worboys'06], [Hart'06].

L'objectif d'un système de surveillance environnemental est de prévenir et gérer les dommages occasionnés par les catastrophes naturelles en surveillant et en analysant les données environnementales en temps réel. L'acquisition et le *monitoring* des données environnementales sont essentiels pour toute gestion ou stratégie de protection. Ceci consiste en une évaluation continue des flux de données capteur [Günther'98].

Les systèmes de surveillance environnementaux sont typiquement larges et distribués couvrant de grandes zones géographiques, surveillent et modélisent des processus physiques, comme la pollution environnementale, les inondations, les éruptions volcaniques, etc. Usuellement, ces systèmes sont composés par trois éléments : *les stations de capteurs, un réseau de communication, et une centrale de traitement.*

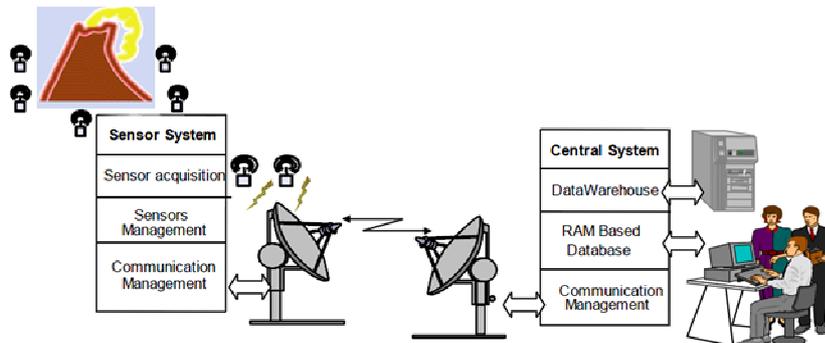


Figure 10 – Système de surveillance de phénomène naturel [Noël'06]

Plus précisément, on retrouve le plus souvent un ensemble de capteurs géolocalisés hétérogènes en réseaux (géocapteurs) mesurant différents paramètres tels que la température, la pression, etc., complétés avec un système de communication permettent que leurs données arrivent au serveur (Figure 10) et soient par la suite, analysées et intégrées avec d'autres ensembles de données environnementales.

Dans les applications environnementales nous pouvons trouver des réseaux contenant des nœuds capteurs statiques ou mobiles, ou attachés aux objets mobiles (i.e. bouées) ou encore utilisés par les humains (i.e. Téléphones mobiles). Dans les systèmes de surveillance environnementaux, les aspects spatiaux de la technologie des réseaux de capteurs en général sont importants.

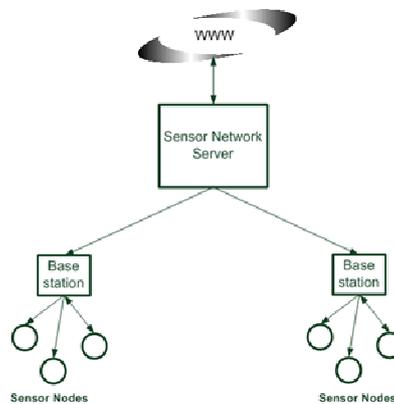


Figure 11 – Diagramme schématique d'un réseau de capteurs environnemental [Hart'06]

Ceci a introduit un nouveau schéma de collection des données, avec des entrées de données en continu issues de capteurs distribués couvrant une zone d'intérêt (Figure 11).

Dans ce contexte, nous pouvons citer les réseaux de géocapteurs [Nittel'09]. Ce type de réseau est un réseau distribué de nœuds sans-fil qui surveille un phénomène dans l'espace. La géolocalisation de ces capteurs est rarement réalisée par l'utilisation de systèmes de positionnement (i.e. GPS¹¹ et DGPS¹²) car leur implémentation peut demander par exemple, des coûts supplémentaires et/ou une augmentation dans la taille du nœud. Dans ce cadre diverses approches s'ont mis en place pour estimer leur position dans l'espace [Xiang'04], [Savvides'01], [Bulusu'02].

Dans ce cadre, les nœuds capteurs sont de bas coût et utilisent peu d'énergie pour leur fonctionnement, ce qui les rend simples à déployer pour surveiller un phénomène naturel. Les réseaux de géocapteurs sont relativement petits, et comprennent des dizaines voire des centaines de capteurs. Par exemple, les réseaux entre 50 et 150 nœuds dédiés à la surveillance de l'habitat pour des espèces en extinction [Szewczyk'04]. De tels capteurs permettent de surveiller des phénomènes géographiques dans des environnements distants, sensibles et périlleux, à une granularité spatiale et temporelle plus haute que les systèmes de surveillance conventionnels [Worboys'06].

Traditionnellement, les données acquises par ce type de réseaux, sont traitées au niveau d'un serveur dans une centrale de traitement, et visualisées et analysées à l'aide d'un Système d'Information Géographique (SIG). Ces données possèdent également la possibilité d'être combinées avec d'autres informations, comme une image satellite, un plan, et publiées via le web pour donner aux utilisateurs un accès en continu à l'information.

3.2.3.2 Les réseaux de capteurs environnementaux et le web

À la fin des années 90 un concept a changé la façon de voir l'application des capteurs et des réseaux capteurs au domaine de la surveillance des phénomènes environnementaux, c'est le *Sensor Web* (Capteur « Réseau »)¹³. Ce terme a été introduit par Kevin Delin de la NASA en 1997, avec une nouveauté en termes d'architecture qui permet de considérer la coordination de divers dispositifs individuels comme un tout.

Cette proposition est un type de réseau de capteurs pour les Systèmes d'Information Géographiques parfois considéré comme un GIS lui même, utilisé notamment dans la surveillance environnementale [Delin'00], [Delin'05]. Le terme « *sensor web* » décrit un type spécifique de réseaux de capteurs distribué dans l'espace qui communiquent entre eux. Ce type de réseau est capable d'interpréter et réagir aux données mesurées, il n'a pas besoin de la présence d'Internet (*world wide web*).

¹¹ GPS – Global Positioning System

¹² WAAS –Wide Area Augmentation System

¹³ Le sens du terme Web se réfère à l'ensemble de capteurs interconnectés en réseau. Il ne s'agit pas de la world wide web.

Cependant, depuis que ce système a évolué, il est maintenant possible de considérer l'Internet comme un moyen d'accès aux données capteurs. Dans la littérature, le concept de *sensor web* a souvent été utilisé pour décrire les capteurs qui sont connectés à Internet. De nos jours, ce lien est palpable. Actuellement, il existe des plateformes proposées avec cette approche : le GSN (Global Sensor Network), le Sensor Web [Sherwood'04], Swiss Experiment¹⁴, entre autres. Dans cet esprit, depuis 2002 l'Open Geospatial Consortium (OGC) s'est concentré sur la spécification et l'application du SWE (Sensor Web Enablement).

La spécification actuelle du SWE¹⁵ est décrite par les interfaces d'interopérabilité et des métadonnées pour permettent l'intégration en temps réel des sensors webs dans une structure d'informations. Le SWE décrit des réseaux de capteurs et des données capteurs accessibles sur le web, qui sont découverts et accessibles par des protocoles standards (décrits en XML). L'OGC a concrètement proposé la gestion des capteurs et des données capteurs via le web, ainsi qu'un ensemble de schémas standards pour la description des capteurs (*SensorML*), l'observation d'un phénomène d'intérêt (*Observation and Measurements*), les services (SOS¹⁶, SAS¹⁷ ...) et les données issues de capteurs (Botts, 2008). Leur objectif est de rendre tous les types de dispositifs (i.e. compteurs de flux d'eau, moniteurs de cœur mobiles, caméras web...) découvrables et accessibles.

D'autres approches reposant sur les spécifications du SWE, ont également étendu leur description en utilisant la sémantique (Semantic Sensor Web). Par exemple, pour la gestion de données issues de capteurs de vidéo, en annotant avec des métadonnées sémantiques les vidéos sur le web [Henson'07], ou pour annoter les données capteurs dans des observations [Thirunarayan'09].

Malgré ces approches innovantes, beaucoup d'autres problèmes de recherche émergent dans le domaine de l'intégration en temps réel des données capteurs au domaine environnemental, comme les nouvelles métadonnées, les droits d'accès, les problèmes de droits de propriété et sécurité, la représentation en échelle uniforme, le passage à l'échelle, la qualité et autres.

3.3 Les réseaux de capteurs et la qualité de données

Parmi les diverses recherches menées jusqu'à aujourd'hui concernant les capteurs et réseaux de capteurs, un bon nombre sont dédiées principalement à l'amélioration de leur technologie et à leur exploitation de plus en plus autonome, comme c'est le cas de l'économie de ressources, l'augmentation de performance de calcul, etc. Un faible niveau d'énergie peut engendrer un retard en termes de transmission des données et par conséquent perturber l'accès et le traitement de ces données. Certaines approches ont été proposées pour améliorer ces éléments [Challen'09], [Hoes'09], [Konstantinidis'07], [Mukhopadhyay'04]. En effet, des protocoles et des techniques

¹⁴ http://www.swiss-experiment.ch/index.php/Main_Page

¹⁵ <http://www.opengeospatial.org/projects/groups/sensorweb>

¹⁶ SOS – Sensor Observation Service

¹⁷ SAS – Sensor Alert Service

d'utilisation ont visé l'augmentation de la performance du capteur et des réseaux de capteur, mais en sacrifiant souvent l'accès et le traitement des données en temps réel.

Concernant cet aspect, diverses améliorations au niveau de la structure du capteur ont été faites, comme dans le cas des motes [Choi'03], [Bauman'05]. Ces dispositifs, possèdent une unité qui leur permet de surveiller l'état du capteur, vérifier leur calibrage, leurs ressources, etc. Cependant, ceci peut donner une meilleure fiabilité de l'équipement mais ne garantit pas que certaines mesures obtenues soient fausses ou impactés par d'autres facteurs. Un rapport entre le mauvais fonctionnement du capteur et la donnée n'est pas toujours prévisible [Lorincz'04], [Stavrou'05], [Bychkovskiy'03], [Haas'96].

Idéalement, les capteurs doivent avoir l'intelligence suffisante pour détecter et corriger les variations de fréquence et les données bruyantes avec une perte de données minimale. Dans ce cadre, la qualité des données peut être évaluée à distance par l'interrogation périodique des données brutes et des données traitées [Klein'07]. Des méthodes typiquement statistiques sont utilisées pour assurer et contrôler la qualité des mesures. Ceci permet aux experts d'identifier et contrôler les variables critiques ou le bruit qui peut dégrader la qualité d'un signal. Par exemple, des paramètres prédéfinis concernant la qualité peuvent être mesurés sur un lot des données spécifiques ou avec des calculs faits au moment où les flux de données arrivent [Holdridge'04].

3.3.1 Les réseaux de capteurs et le contexte environnemental

Dans plusieurs situations, les capteurs en bon état sont déployés en réseau pour opérer dans des conditions parfois difficiles, même périlleuses. Ils arrivent à mesurer le monde réel selon leurs capacités mais en fournissant des informations douteuses. Cet effet est assez important, une donnée incorrecte peut être propagée jusqu'au serveur central et offrir une mauvaise interprétation du monde réel. Néanmoins, une donnée peut ne pas être seulement fausse en raison du fonctionnement défaillant du capteur; les conditions du déploiement et d'utilisation ont également un impact. Par exemple, un capteur placé dans le périmètre d'une montagne peut collecter des mesures de température vraiment basses, hors du seuil spécifié. Au premier regard, la tendance à dire que le capteur est mal calibré est inévitable. Mais cet effet peut être évité si on considère que le capteur est déployé dans une zone dans laquelle une tempête de neige a frappé et que le capteur se trouve deux mètres en-dessous de la neige. Effectivement, le capteur fonctionne correctement, mais les conditions environnementales ne sont pas celles prévues et affectent les mesures. Un capteur de mouvement placé sur les environs d'un volcan peut momentanément expérimenter des vibrations importantes que ne sont pas forcément liées à un phénomène. Ces vibrations peuvent provenir de l'atterrissage d'un hélicoptère qui fait des reconnaissances de terrain. De multiples problèmes peuvent aussi être issus d'attaques malveillantes ou engendrés par des animaux près de la zone de détection.

Dans ce contexte, nous pouvons identifier différents types d'erreurs intervenant dans un système d'acquisition de mesures [Taylor'99] :

- *Erreurs systématiques*

- *Erreurs conditionnelles*
- *Erreurs stochastiques.*

Ces erreurs proviennent parfois du capteur lui-même, mais plus souvent, ces erreurs résultent de l'effet du système de déploiement et leur contexte.

Les erreurs systématiques sont de caractère permanent et dépendent typiquement du système d'acquisition. Toutes les mesures sont exposées à ce type d'erreur. Les sources d'erreurs systématiques peuvent être : un mauvais calibrage, changements dans l'environnement, et souvent des méthodes d'observation imparfaits.

Les erreurs conditionnelles résultent de l'environnement et des conditions d'opération du système. Pour les réduire, des structures plus fiables sont considérées, ainsi que l'utilisation de capteurs du même type pour remplacer les capteurs défectueux, entre autres.

Les erreurs stochastiques résultent de processus fondamentaux se produisant dans les composants du système tels que le bruit, l'échauffement du matériel, la pression, etc.

De ce fait, les systèmes de surveillance utilisant des capteurs ou des réseaux de capteurs exposés à de multiples facteurs qui impactent la fiabilité de l'information collectée. Concrètement, la qualité des données obtenue par ces capteurs est sensible à ces facteurs.

3.3.2 Les réseaux de capteurs et la QoS

Les différentes communautés techniques regardent la qualité de service de façon différente. Cependant, dans le domaine des applications utilisant des réseaux de capteurs, la QoS est définie comme *le nombre optimal de capteurs qui peuvent envoyer l'information à n'importe quel moment* [Perillo'03]. Dans ce sens, d'un point de vue applicatif, comme le montre [Chen'04], la QoS peut être aussi regardée du côté utilisateur ou application dirigée vers les objectifs fixés par les utilisateurs ou l'application. Elle peut également être bien analysée au niveau du réseau physique dont les ressources sont gérées indépendamment de l'utilisateur [Karl'03].

Dans ce contexte, chaque type d'application et de réseau impose ses contraintes sur la QoS selon leurs caractéristiques particulières. La définition des caractéristiques QoS propres à une application spécifique, n'est pas anodine.

Actuellement, ils existent diverses approches qui sont liées à la qualité de service pour les réseaux de capteurs sans-fil, et sont classifiées en trois catégories : *QoS bout-à-bout traditionnelle* ¹⁴, *l'assurance de la fiabilité* et *la QoS spécifique d'application*.

Concernant la qualité de service bout-à-bout traditionnelle, plusieurs protocoles pour le routage ont été proposés (i.e. SPEED, QoS aware, SAR. . .). L'objectif général de ces protocoles est de fournir une assurance du trafic en réseau en temps réel ou en presque-temps réel. D'autres protocoles comme les MAC [Akyildiz'02], ne traitent pas forcément la QoS et sont plus dédiées à la gestion d'énergie.

En relation à l'assurance de la fiabilité, quelques approches ont été proposées afin de déterminer les niveaux de priorités dans les paquets des données à transmettre. En regardant un schéma avec une fiabilité désirée, ce chemin devient prioritaire. Par ailleurs, d'autres approches comme ESRT¹⁸ (Sankarasubramaniam, 2003), prennent en considération d'autres éléments comme le contrôle du trafic, la latence, l'énergie, etc. Cette solution est proposée principalement pour la détection des événements ; le contrôle de caractéristiques est fait dans un nœud dont les capacités de calcul et d'énergie sont plus grandes que dans les capteurs.

Finalement, concernant les solutions proposées pour la QoS applicative, diverses propositions ont fait un compromis entre l'ordonnancement des capteurs et le routage des données [Perillo'03]. Ceci, prolonge la vie du réseau et offre un équilibre entre la fiabilité et la consommation de l'énergie.

D'autres solutions comme [Meguerdichian'01], défissent la QoS selon la couverture et l'exposition des capteurs. D'autres encore, ont été proposées au sein des réseaux mobiles ou des réseaux de surveillance médicale en prenant en compte la *QoS contexte-awareness* orienté service [Wac'07]. Ces travaux sont orientés applications et bout-à-bout en cherchant la correcte transmission de l'information, mais aussi la fiabilité et l'adaptation du service. Ils prennent en compte des caractéristiques comme la vitesse de calcul, l'exactitude, la dépendance, le passage à l'échelle et la gestion de l'information du contexte-QoS.

3.4 Conclusion

Nous avons présenté dans ce troisième chapitre, un état de l'art sur les spécificités des réseaux de capteurs et leur application dans différents domaines, notamment la surveillance. Nous avons identifié les caractéristiques des capteurs et des réseaux de capteurs, ainsi que les caractéristiques des données capteurs. Nous avons également donné un aperçu sur les structures existantes pour la surveillance des phénomènes environnementaux et leurs contraintes en termes de qualité. Nous remarquons suite à cette étude de l'état de l'art que l'utilisation de capteurs et de réseaux de capteurs est un important sujet concernant la qualité des données utilisées pour la prise de décision, notamment sur les systèmes de surveillance des phénomènes environnementaux. En effet, les capteurs sont supposés exécuter des mesures sur n'importe quelle contrainte. Mais, en réalité, dans ce type d'applications, les capteurs sont impactés par l'environnement de déploiement parfois hostile. Dans le monde réel, nous pouvons observer que les conditions physiques et météorologiques comme le vent, la pluie, les orages, séismes, neige, etc., sont souvent la cause des pannes ou dérèglages au niveau de capteurs. Également, d'autres aspects plus restrictifs comme leur capacité d'énergie et de stockage limité, ainsi qu'une mauvaise manipulation du dispositif, peuvent impacter la performance des capteurs.

Dans cet esprit, peu de travaux de recherche se sont intéressés à savoir ce qui se passe avec les données obtenues par les capteurs selon ces circonstances. Une question

¹⁸ Reliable Transport Scheme

fondamentale se pose : *La qualité des données peut-elle être impactée par les caractéristiques des capteurs et leur déploiement ?*

De plus, si on considère des défaillances au moment de l'acquisition et du traitement de ces données, une deuxième question se pose : *comment évaluer ces défaillances et leur impact sur la qualité de données capteurs ?*, et finalement une troisième question se pose : *Comment communiquer cette information à l'utilisateur de façon opportune pour l'aider lors de la prise de décision ?*

Pour contribuer à résoudre ces problématiques, nous proposons dans la deuxième partie de ce manuscrit, notre contribution scientifique qui concerne principalement la formalisation des données issues de capteurs (spatio-temporelles) au sein des systèmes de surveillance environnementaux, ainsi que la proposition d'une méthodologie pour la définition de la qualité des données issues de capteurs qui aboutit à l'évaluation et la communication de la qualité de ces données. Pour sa spécification, cette méthodologie se base sur des méthodologies existantes dans le domaine de la qualité de données dans les systèmes d'information. Elle repose plus spécifiquement sur un modèle de qualité composé de critères, indicateurs et métadonnées qualité associées à des fonctions d'évaluation.

Deuxième partie : Contribution

Chapitre 4

Modélisation des données capteurs dans les systèmes de surveillance

Dans ce chapitre, nous introduisons en premier lieu (Section 4.1) les spécificités des systèmes de surveillance des phénomènes environnementaux sur lesquels nous sommes appuyée pour l'analyse des données géolocalisées (spatio-temporelles) issues de capteurs. Ces spécificités concernent principalement l'architecture du système de surveillance environnemental, le type et l'organisation du réseau (réseau de géocapteurs), ainsi que la gestion de données issues de ce type de système.

En se basant sur ces spécificités, nous proposons en deuxième lieu (Section 4.2), une modélisation des données capteurs, dites spatio-temporelles temps réel. Ainsi, nous effectuons une analyse afin de mettre en évidence les informations utiles pour caractériser les données. Une formalisation de ces éléments est également proposée.

Nous poursuivrons ce chapitre par la description des spécificités concernant le traitement des données capteurs au sein du système de surveillance environnemental (Section 4.3). Nous concluons par une discussion sur les éléments clés de notre domaine d'application (Section 4.4).

4.1 Du phénomène au système de surveillance

Dans notre contexte d'application, les systèmes de surveillance environnementaux, sont composés principalement par des réseaux de capteurs sans-fil (i.e. géocapteurs, réseaux de capteurs environnementaux), qui possèdent un ensemble de capteurs ou nœuds de capteurs, ainsi qu'un système de communication qui permet aux données de rejoindre un serveur central (Figure 12). On notera que dans les systèmes de surveillance environnementaux, les réseaux de capteurs sont composés par des capteurs fixes et agiles qui forment traditionnellement la majeure partie de ceux utilisés dans ce type de systèmes, alors que les capteurs mobiles sont souvent dédiés au suivi ou à la gestion de flottes.

Dans ce type d'architecture, les nœuds capteurs collectent des données de façon automatique; un réseau est utilisé pour transférer les données à la passerelle ou entre plusieurs passerelles, avant d'arriver au serveur central, appelé également centre de données.

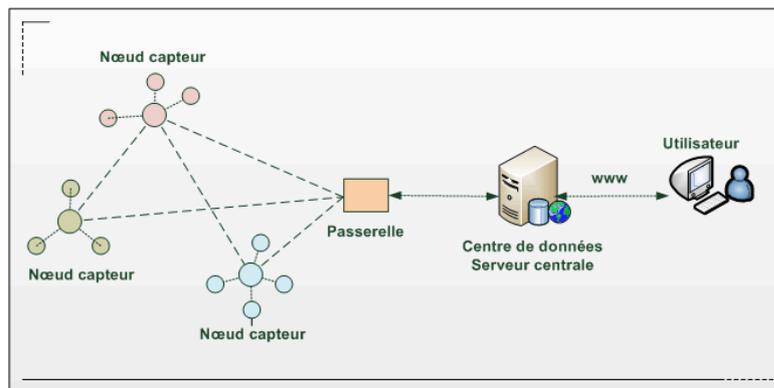


Figure 12 – Architecture technique d'un système de surveillance environnementale

Au niveau du serveur, ces données sont traitées et ensuite exploitées soit par l'utilisateur, soit par un SIG (Système d'Information Géographique) ou par le web.

4.1.1 Couches de traitement dans les systèmes de surveillance environnementaux

Dans la structure d'un système de surveillance environnemental, nous considérons trois couches principales dans lesquelles divers processus de gestion en temps réel sont effectués: la *couche Acquisition*, la *couche Traitement*, et la *couche Exploitation* (Figure 13).

Au niveau de la *couche d'acquisition*, les réseaux sont composés par des nœuds/capteurs fixes ou agiles. Différents types de données peuvent être fournis par ces nœuds/capteurs dépendant du type d'application. Les données acquises, appelées aussi *données observation* correspondent à des éléments comme l'humidité, la température, la pression, etc., et peuvent être représentées sous différents formats, de façon numérique, analogique, spatiale, temporelle, spatio-temporelle, multimédia (image, vidéo), etc.

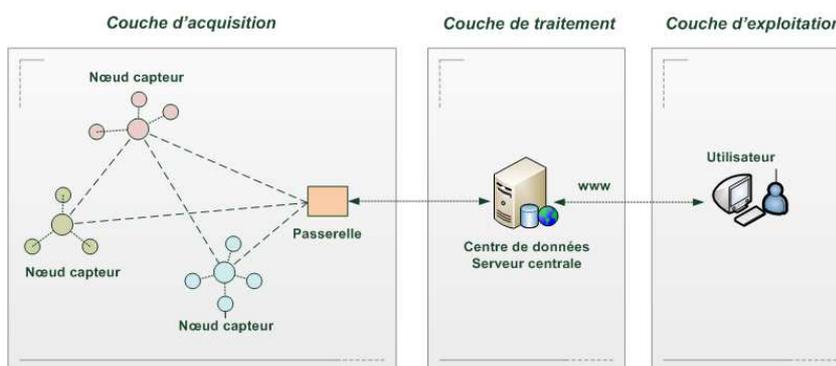


Figure 13 – Couches de traitement dans les systèmes de surveillance environnementaux

C'est au niveau de cette couche que les premières décisions concernant les éléments à mesurer et les données résultantes sont prises. Au sein des nœuds capteurs, les données peuvent être stockées, agrégées ou filtrées afin de déterminer quelles données seront transmises, selon les conditions d'acquisition, et en considérant les capacités physiques et de calcul des nœuds/capteurs. Lors de leur acquisition et prétraitement, ces données sont transférées au centre de données, au sein duquel le processus de traitement se déroulera.

Au niveau de la *couche Traitement*, le traitement des données collectées est réalisé. Leur traitement est effectué sur un serveur, dont les capacités d'énergie, de stockage et du calcul deviennent plus significatives. Même si le prétraitement est réalisé au sein des nœuds, ce processus a normalement pour but de filtrer et réduire la quantité des données à gérer au sein de la mémoire principale, grâce à l'agrégation des données, par exemple. Comme nous l'avons détaillé précédemment, afin d'assurer l'acquisition d'information le plus possible, et éviter la perte d'information à la volée, des approches externes pour la localisation des données, sont utilisées au niveau des systèmes de surveillance environnementale en évitant l'agrégation des données. Ces données seront par la suite mises à disposition aux utilisateurs.

Au niveau de la *couche Exploitation*, les informations sont traitées au centre des données par un SIG, combinées avec d'autres outils comme internet, qui permettent l'accès à distance aux données. Ceci représente un réel avantage pour les utilisateurs, qui peuvent maintenant avoir un accès plus simple à l'information. Ces informations peuvent être exploitées en temps réel ou à posteriori selon les besoins de l'utilisateur et leur application. Dans des applications du type surveillance, le temps réel est le plus recommandé, permettant de suivre l'évolution d'un phénomène et réagir par rapport à son comportement.

4.1.2 Spécificités des systèmes de surveillance environnementaux

Les systèmes de surveillance environnementaux ont fait preuve de notables développements technologiques, qui ont permis l'amélioration de la surveillance de l'environnement naturel. Ces systèmes peuvent être organisés de diverses façons, chaque application et système, présentant des particularités en fonction des besoins des organisations ou des utilisateurs.

De façon générale, nous pouvons signaler que le contexte d'application dans les systèmes de surveillance est assez dynamique et possède des caractéristiques liées au temps réel, qui le distinguent des applications traditionnelles. Les données présentent aussi des caractéristiques dynamiques, voire l'évolution de la donnée au cours du temps. Par conséquent, ce type de données est notamment géré en temps-réel, et ceci n'est pas anodin car l'environnement change aussi au cours du temps. Certaines de ces données varient plus que d'autres et possèdent des priorités différentes, selon le type d'application et le type de phénomène à observer. Parallèlement, les informations obtenues peuvent être exploitées par des SIG ad-hoc à l'application et l'utilisateur peut aussi avoir un accès à distance grâce à l'internet.

Dans ce contexte, en adéquation avec les modèles existants de systèmes de surveillance environnementaux, nous considérons pour notre approche, les caractéristiques suivantes :

a) Spécificités liées au contexte d'acquisition des données capteurs

- Les nœuds capteurs contiennent de l'information géolocalisée (données spatio-temporelles). Les techniques utilisées pour localiser les mesures sont par exemple, le GPS ou la localisation distribuée de services comme la triangulation.
- Les nœuds capteur possèdent des propriétés fonctionnelles liées à la persistance, la gestion d'erreur, la gestion du cycle de vie, et leur déploiement physique. Ces capteurs utilisent des *métadonnées* pour leur identification et leur découverte.
- Le réseau de capteurs doit avoir un support pour la qualité du service afin de garantir la satisfaction des besoins de service pour l'application.
- La redondance de données capteurs doit être gérée de façon pertinente due à leur consommation considérable d'énergie au sein du réseau. Aucune agrégation ou fusion des données est utilisée pour ces fins. Nous estimons qu'il est préférable de transmettre les données en gardant les meilleures conditions au niveau de la qualité de service du réseau.
- La fréquence d'acquisition au sein des capteurs n'est pas toujours la fréquence de transmission. Les conflits entre fréquences d'acquisition et transmission des données capteurs doivent être gérés. Des fréquences de transmission plus basses ou égales à la fréquence de transmission sont conseillées (<100Mhz) afin d'améliorer les caractéristiques de la qualité de service. Les flux de données doivent avoir cette spécificité.

b) Spécificités liées au traitement des données capteurs

- Les systèmes de surveillance sont de façon inhérente des systèmes temps réel, à savoir que les actions de contrôle et surveillance doivent être exécutées dans les systèmes physiques dans des délais de temps impartis. Dans notre contexte, le temps réel ne veut pas dire « rapide », mais que les délais sont limités, prévisibles et gérables (*soft real-time*).
- Les mesures issues de capteurs sont valides (actuelles) pour un temps limité.
- Les mesures issues de capteurs sont enregistrées dans une base de données centralisée ou traitées au niveau du serveur en mémoire vive pendant leur temps de validité. Le risque d'une panne d'un capteur avant la transmission de données limite la possibilité de stockage au niveau du capteur ou du réseau.
- Toutes les mesures collectées n'ont pas la même priorité pour être traitées selon différents critères, plus spécifiquement selon la période d'acquisition. Les mesures acquises en période de crise, sont plus importantes et doivent être prioritaires. Les données les plus récentes sont privilégiées pour être traitées au sein du serveur ou dans la base de données, selon la mémoire disponible. Les plus anciennes sont normalement dirigées vers un entrepôt de données ou sur disque.

Dans la section suivante, nous allons décrire plus en détails, les spécificités des données capteurs dans des systèmes de surveillance et présenter une proposition de structuration selon les spécificités précédemment décrites.

4.2 Données spatio-temporelles issues de réseaux de capteurs : données observation

Dans le domaine de la surveillance de phénomènes environnementaux, l'observation des phénomènes constitue leur principale activité. L'observation d'un phénomène est considérée traditionnellement comme un événement qui a une valeur décrivant un phénomène quelconque. En général, une observation utilise des processus (analytiques, de traitement...) et des observateurs ou capteurs pour déterminer la valeur d'un phénomène.

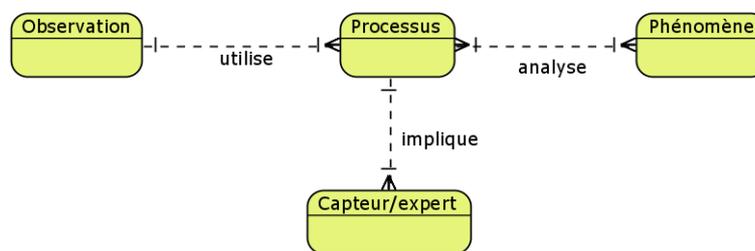


Figure 14 – Modèle conceptuel d'observation

Une observation peut être réalisée en utilisant des capteurs, pendant des périodes de temps pour lesquelles les observations sont demandées, sur un phénomène quelconque ou sur une région qui contient plusieurs capteurs d'intérêt ou des éléments qui sont l'objet d'une analyse (Figure 14).

Nous nous basons sur ce concept (aussi utilisé par l'OGC - SWE¹) dont les données utilisées pour décrire le phénomène sont considérées comme *données observation*. Ces données d'observation sont des données géolocalisées issues de capteurs.

De ce fait, nous considérons que les données observation en temps réel possèdent des propriétés spatiales, temporelles, sémantiques et dynamiques, ainsi que des propriétés liées à des informations complémentaires contenues dans des métadonnées.

Les capteurs liés à une observation, peuvent également être référencés de façon spatiale ; il est donc possible de lier une donnée observation à une position spécifique. En conséquence, une donnée liée à un capteur dans l'espace, peut porter des estampilles de temps pour notifier la date de l'acquisition. Par conséquent, ces données seront aussi définies et ordonnées temporellement.

Par ailleurs, les données observation peuvent varier dans le temps tout au long de l'observation. Nous pouvons avoir une nouvelle valeur de mesure à chaque instant

¹ SWE – Sensor Web Enablement : <http://www.opengeospatial.org/ogc/markets-technologies/swe>

du temps. En conséquence, des variations importantes dans le temps concernant le phénomène peuvent être retrouvées selon l'espace ou le temps.

De plus, si les données sont acquises par différents types de capteurs, ou suite au résultat d'un calcul (i.e. la localisation d'épicentres), des propriétés sémantiques décrivant ces aspects peuvent être ajoutées aux données. Les données peuvent être aussi liées aux différentes informations complémentaires aux données (métadonnées) comme sont par exemple, les propriétés du capteur, ou du système d'acquisition, leur comportement, leur performance, leur qualité de service, les conditions environnementales au moment de la prise, etc. En conséquence, les données peuvent être référencées par ces informations contenues dans les *métadonnées*.

Dans le paragraphe suivant, nous présentons la description d'une modélisation UML² dans un réseau de capteurs environnemental. L'identification des objets statiques et dynamiques au sein de ce modèle de données est mise en évidence, afin de déterminer les informations relatives aux données et aux métadonnées. Nous détaillons ainsi une formalisation des objets issus de ces capteurs.

4.2.1 Gestion de données spatio-temporelles issues de réseaux de capteurs

Les données issues de capteurs présentent, comme nous l'avons détaillé au début de cette section (Section 4.2), des spécificités en partie dues au type de capteurs (fixe, agile ou mobile) et également à la façon dont les données sont acquises. Dans le cas des objets mobiles, il est possible, d'utiliser des langages comme GML³ pour caractériser les objets changeant de position au cours du temps. Cependant, les données décrites par ce langage ne prennent pas en compte les spécificités propres aux capteurs comme la nature du phénomène observé ou les référentiels utilisées par les mesures.

Nous proposons donc un modèle générique UML de données mesures et capteurs (Figure 15) [Gutiérrez'07]. L'objectif de ce modèle est de faciliter la compréhension des caractéristiques des données géolocalisées issues de capteurs (données observation). Ce modèle représente les relations entre les propriétés spatiales, temporelles, spatio-temporelles, ainsi que la dynamique des données dans une observation, voire les données spatio-temporelles temps réel issues de capteurs. Ce modèle nous permet également, de mettre en évidence les possibles informations qui peuvent caractériser la donnée dans un contexte de surveillance ou d'observation (métadonnées).

² UML - Unified Modeling Language

³ GML – Geographic Markup Language (ISO/TC211)

4.2.1.1 Modèle de données capteur

Dans ce modèle de données (Figure 15), un réseau de capteurs est composé d'un ensemble de capteurs. Généralement, dans des applications actuelles (NOAA⁴, SANY⁵, SwissEx⁶, SWE⁷), les réseaux de capteurs sont situés dans un même milieu d'observation, et offrent la possibilité d'avoir des données à positions fixes ou variables (capteurs fixes, agiles, mobiles).

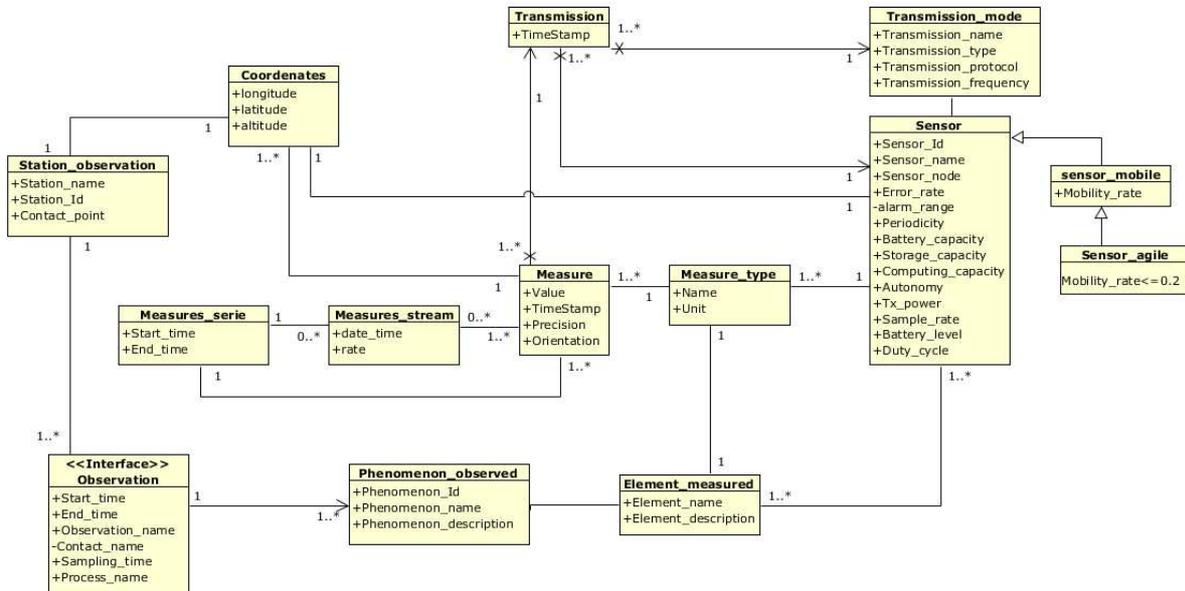


Figure 15 – Modèle de données spatio-temporelles temps réel [Gutiérrez'07]

Plus spécifiquement, dans la surveillance de phénomènes environnementaux, il existe des stations d'observation (stations météorologiques, maritimes, agricoles...) chargées de surveiller, voire observer, différents phénomènes (tsunamis, hurricanes, séismes, éruptions, érosion...) auxquels correspondent un ou plusieurs éléments (température, gazes, mouvement de terrain, humidité...) qui déterminent l'évolution de ces phénomènes. Ces phénomènes sont analysés à partir de l'agrégation de flux de données, ou des ensembles de mesures issues de capteurs géolocalisés (fixes, agiles ou mobiles) en temps réel de façon continue, qui font l'objet de campagnes de mesures pendant des intervalles de temps déterminés. Les capteurs peuvent également effectuer plusieurs mesures de types différents (i.e. Smart dust).

Lors de l'acquisition des mesures en temps réel dans les systèmes de surveillance environnementaux, nous remarquons la dynamique de certains éléments mesurés changeant de valeur, de forme ou de position dans le temps ou/et l'espace. Nous détaillons ces aspects dans la suite.

⁴ National Oceanic And Atmospheric Administration - <http://www.noaa.gov/>

⁵ SANY – Sensors Anywhere - <http://www.sany-ip.eu/>

⁶ Swiss Experiment - http://www.swiss-experiment.ch/index.php/Main_Page

⁷ Sensor Web Enablement

4.2.1.2 Éléments statiques et dynamiques du modèle de données capteurs

Dans notre proposition de modèle de données capteurs présenté dans la Figure 15, nous identifions la relation entre les objets dont la valeur est figée au cours d'une campagne de mesures (*statiques*) et les objets qui évoluent au cours du temps lorsqu'une campagne est déclenchée (*dynamiques*).

a) Éléments statiques du modèle de données capteurs

Dans ce modèle des données capteurs nous pouvons distinguer par exemple, les objets dits statiques qui concernent notamment *la station d'observation, le phénomène observé, l'élément mesuré, les campagnes de mesures et la position fixe du capteur* (Figure 16).

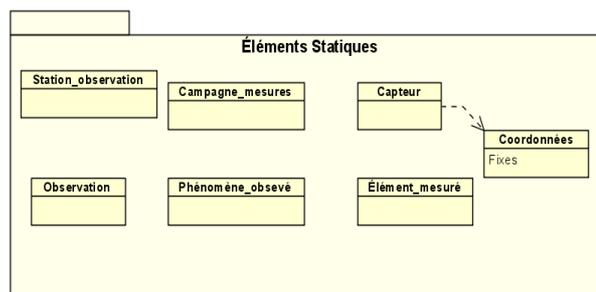


Figure 16 – Éléments statiques du modèle de données spatio-temporelles temps réel

Ces aspects sont généralement valables pour toute la campagne de mesures. Dans ce contexte, lors qu'une observation est déclenchée, les éléments à évaluer sont choisis, ainsi que la période du temps pendant laquelle l'observation sera exécutée. Les capteurs désignés pour une telle observation, sont pratiquement les capteurs qui restent dans le même milieu d'observation, ayant peu de probabilités d'être remplacés ou déplacés au cours de la campagne. Leur position ne change pas, mais leurs caractéristiques physiques sont susceptibles d'évoluer (niveau de ressources) [Gutiérrez'07a].

b) Éléments dynamiques du modèle de données capteurs

Contrairement aux objets statiques, les objets dynamiques possèdent un comportement plus dépendant du phénomène lui-même et du contexte temps réel.

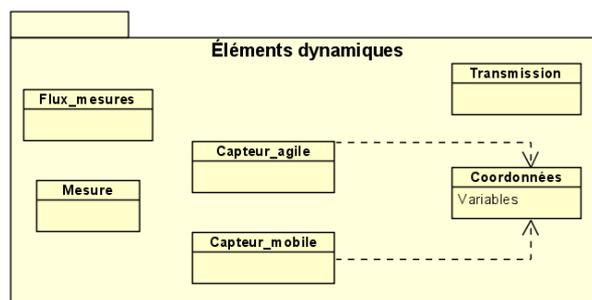


Figure 17 - Éléments dynamiques du modèle de données spatio-temporelles issues de capteurs

Ces aspects peuvent varier à chaque instant du temps : la position variable du capteur, la mesure elle-même, le flux de mesures et les attributs variables des capteurs (i.e. ressources) (Figure 17) [Gutiérrez'07b].

Grâce à notre proposition de modélisation, nous pouvons analyser d'une façon simple les objets et les relations entre eux, notamment le phénomène depuis plusieurs angles, selon leur comportement, leur type, les capteurs utilisés pour l'évaluer, etc.

4.2.2 Formalisation des objets issus de capteurs

Au sein des réseaux de capteurs, les objets principaux concernent les capteurs et les mesures ou information acquises. Nous proposons dans ce paragraphe un formalisme pour les objets correspondant à la relation entre les flux et les ensembles de mesures issues de différents capteurs (fixe, agile et mobile). Dans ce formalisme nous analysons les mesures issues de capteurs, selon les concepts *statique* et *dynamique* par rapport à l'utilisateur final [Gutiérrez'07], [Servigne'09]. Ceci caractérise les objets lors de ces campagnes de mesures, qui concernent les objets statiques ou changeant de valeur, de forme ou de position dans le temps ou/et l'espace.

4.2.2.1 Spécificités de mesures issues de capteurs

Dans la formalisation que nous proposons, le capteur est principalement référencé selon un identifiant que nous appelons *IdS* et sa position *P*. La position du capteur dans notre modèle est spécifiée par des coordonnées (longitude, latitude) mais aussi par la date/heure de localisation à une position donnée. Un capteur possède également des attributs propres (type, taux d'erreur, précision...).

Concernant les capteurs, le cas du capteur mobile représente un type particulier de capteur et donc hérite ses attributs. Le capteur mobile est caractérisé par son taux de mobilité ; par contre, un capteur agile est un capteur mobile mais avec une mobilité restreinte. Les capteurs agiles et mobiles occupent donc diverses positions au cours du temps.

Les mesures effectuées par ces capteurs sont donc réalisées à un instant *t* et une position donnée *P*, cette position étant liée à une période définie par le début et la fin du déplacement, et donc à chaque nouvelle position une nouvelle mesure est datée et enregistrée. Une campagne de mesures peut alors porter sur un flux ou un ensemble de mesures. La campagne de mesures permet donc d'évaluer des flux de données ou des mesures issues de plusieurs capteurs, pouvant eux-mêmes effectuer des mesures différentes à des instants différents, ainsi que des positions différentes.

Selon ces changements de position, nous pouvons trouver des ensembles ou flux de mesures qui correspondent à :

- un capteur à une même position,
- des mesures qui à chaque instant du temps sont référencés à une position différente.

- des mesures qui restent stables pendant une période du temps, appelées mesures sur des phénomènes constants. Dans ce cas, la valeur de la mesure reste constante tout au long d'une campagne de mesure.
- une mesure qui varie sans contraintes, autant en valeur qu'en position, comme dans les flux de données et est liée à un phénomène dynamique.

4.2.2.2 Description de la formalisation des objets issus de capteurs

Les objets résultats sont donc formalisés comme suit : l'objet possède un identifiant id , lequel est unique et ne varie pas dans le temps et peut être issu d'un ensemble de mesures β ou d'un flux de mesures φ (ensemble infini de mesures). Ces objets positionnés P sont associés à des attributs thématiques. Les attributs dont la valeur varie sont nommés A . Pour chaque attribut A du type j , une valeur A_t^j est donnée à un instant t . Les instants t sont ordonnés avec $t_i \mid t_i < t_{i+1}, i > 0$ pour β , et avec $t_i \mid t_i < \infty, i > 0$ pour φ . L'instant du temps (date/heure) pour une position donnée est représenté par t_p [Gutiérrez'07], [Servigne'09].

Les valeurs A_t^j sont fournies par les capteurs à une fréquence f constante ou variable. Les attributs dont la valeur reste constante sont nommés B . Leur valeur ne varie pas dans le temps. Une fois ces notations définies, les objets issus des trois types de capteurs peuvent être formalisés.

a) Objets issus de capteurs fixes (OCF)

Nous présentons la formalisation des objets issus de capteurs fixes (OCF_β et OCF_φ) comme suit (1) et (2) :

$$OCF_\beta : \{id, IdS, P, \{t_i \{A_t^j\}\}, \{B^k\}\} \quad (1)$$

avec $t_p \leq t_i \rightarrow P = fixe$

avec $t_i < t_{i+1}, i > 0$ A une mesure du type j à l'instant i

avec B^k attribut de valeur fixe

$$OCF_\varphi : \{id, IdS, P, \{t_i \{A_t^j\}\}, \{B^k\}\} \quad (2)$$

avec $t_i < t_{\infty}, i > 0$ et A une mesure du type j à l'instant i

avec B^k attribut de valeur fixe

Définition 1. Pour les objets issus de capteurs fixes, la position P du capteur est fixée, seules les valeurs des attributs A varient dans le temps t . Nous pouvons donc définir ces objets OCF_β issus d'ensembles de mesures ou encore des objets issus de flux de mesures OCF_φ à l'aide d'un identifiant id , d'une position P , d'une série temporelle où pour chaque instant t (fini ou infini) est communiquée une valeur pour chaque A ainsi

que des attributs de valeur fixe. L'instant de temps à une position donnée est compris dans la même série temporelle que l'instant pour chaque mesure.

b) Objets issus de capteurs agiles (OCA)

Nous présentons la formalisation des objets issus de capteurs agiles (OCA_{β}) comme suit (3) et (4) :

$$OCA_{\beta} : \{id, IdS, \{P_t \{t_i, \{A_i^j\}\}\}, \{B^k\}\} \quad (3)$$

$$t_i^p \leq t_i^{p+\Delta t} \rightarrow P = agile$$

avec $P_{t+\Delta t}$ la position du capteur dans la période $t + \Delta t$

avec t_i date et heure à l'instant i et $t \leq i \leq t + \Delta t$ et $t_i \leq t_{i+1} \forall i > 0$

avec A une mesure du type j à l'instant i

avec B^k attribut de valeur fixe

$$OCA_{\varphi} : \{id, IdS, \{P_t \{t_i, \{A_i^j\}\}\}, \{B^k\}\} \quad (4)$$

avec $P_{t+\Delta t}$ la position du capteur dans la période $t + \Delta t$

avec t_i date et heure à l'instant i et $t \leq i \leq \infty$ et $t_i < t_{i+1} \forall i > 0$

avec A une mesure du type j à l'instant i

avec B^k attribut de valeur fixe

Définition 2. Pour les objets issus de capteurs agiles OCA_{β} et OCA_{φ} , la position du capteur ne varie pas selon la même dimension que les attributs A^j . Nous pouvons donc formaliser ces objets à partir d'un identifiant Id et d'une première série temporelle des positions prises par ce capteur $P_{t+\Delta t}$ avec un temps fini ou indéfini. A l'intérieur de cette série temporelle, une deuxième série temporelle des valeurs A_i^j pour chaque A est définie. Elle contient la série de mesures réalisées à cette position par le capteur. Des attributs de valeur fixe viennent compléter ce formalisme.

c) Objets issus de capteurs mobiles (OCM)

Nous présentons la formalisation des objets issus de capteurs mobiles (OCM_{β} et OCM_{φ}) comme suit (5) et (6) :

$$OCM_{\beta} : \{id, IdS, \{P_i \{t_i, \{A_i^j\}\}\}, \{B^k\}\} \quad (5)$$

$$t_i^p = t_i \rightarrow P = mobile$$

avec $t_i < t_{i+1} \forall i > 0$ et A une mesure du type j à l'instant i

avec B^k attribut de valeur fixe

$$OCM_{\varphi} : \left\{ id, IdS, \left\{ P_i \left(t_i, \{A_i^j\} \right) \right\}, \{B^k\} \right\} \quad (6)$$

avec $t_i < t_{\infty} \ t_i > 0$ et A une mesure du type j à l'instant i

avec B^k attribut de valeur fixe

Définition 3. Pour les objets issus de capteurs mobiles, la position du capteur est susceptible de varier à chaque envoi d'information. Nous pouvons donc formaliser ces objets comme un identifiant Id , une série temporelle où, pour chaque instant t , est communiquée une position $P_i(t_i^p = t_i)$ et une valeur A_i^j pour chaque A est définie. Des attributs de valeur fixe B^k viennent compléter ce formalisme.

Tout au long de cette formalisation nous avons utilisé les attributs de valeur fixe B^k pour les compléter. Dans un contexte dynamique comme celui de la surveillance, la gestion temps réel des données spatio-temporelles issues de capteurs nécessite une analyse pour déterminer les attributs correspondant aux valeurs variables dans le temps et celles qui sont figées. Dans ce contexte, les éléments de valeur fixe peuvent être considérés comme des *métadonnées*. Ceci pousse le problème de l'identification des données et métadonnées vers un contexte dynamique.

Dans le paragraphe suivant, nous présentons une discussion détaillée concernant cet aspect.

4.2.2.3 Données spatio-temporelles issues de capteurs et métadonnées

A partir de notre proposition de modèle de données capteurs, de son analyse et formalisation (Sections 4.2.1 et 4.2.2), nous poursuivons notre étude afin de définir des règles d'identification permettant de distinguer les données des métadonnées dans un contexte dynamique. Nous considérons, à partir de notre étude, qu'il existe deux aspects à prendre en considération : la grande *dynamacité* des données et la *sémantique* des données.

La sémantique d'une donnée est caractérisée selon son exploitation par un utilisateur final. Selon l'application ou le domaine d'application, une donnée caractérise ou non un phénomène étudié. L'utilisateur final, en décrivant l'objectif et le contexte de son application, peut ainsi permettre d'identifier les informations relevant des données de celles relevant des métadonnées. La difficulté est alors de réussir à différencier les objectifs d'une application et les données nécessaires à cette application, du contexte d'acquisition et d'exploitation de ces données, qui doit être décrit par les métadonnées.

D'un autre côté, la *dynamacité* induit une variation considérable de valeurs dans le temps. La dynamacité et la sémantique sont d'ailleurs très liées dans un contexte temps réel. Prenons comme exemple la mobilité d'un capteur. Si la position d'un capteur varie constamment, ce qui est généralement le cas d'un capteur mobile embarqué, cette information sera inévitablement considérée comme une donnée. En effet comme nous l'apercevons dans les travaux exploitant des données issues de capteurs géolocalisés, la localisation est également une mesure et fait partie intégrante des éléments mesurés au

même titre que les autres éléments mesurés par le capteur. Par contre, dans le cas d'un capteur fixe, la position est une information stable. Elle peut être considérée comme une donnée complémentaire sur les mesures effectuées par le capteur. La position peut alors être étiquetée comme métadonnée.

Nous pouvons donc en déduire qu'une information acquise en temps réel et à forte dynamique (fréquence de modification ou de transmission élevée) est identifiée comme donnée et ne peut pas être considérée comme une métadonnée. La dynamique et également le contexte d'utilisation d'une information sont donc des critères de caractérisation d'une information comme donnée ou métadonnée.

Parallèlement, l'utilisation des réseaux de capteurs dans les systèmes de surveillance implique aussi des contraintes liées au temps, notamment sur la partie de l'accès et le traitement des données capteurs. Nous détaillerons ces aspects par la suite.

4.3 Traitement des données spatio-temporelles issues de réseaux de capteurs.

Dans les systèmes de surveillance de phénomènes environnementaux, le volume de données à traiter est considérable, et donc le traitement s'effectue au niveau du serveur, dont les capacités de calcul et de stockage sont plus élevées.

Ces applications, utilisent généralement une combinaison de données issues de capteurs stockées et organisées de façon relationnelle avec des données issues de flux ou des ensembles de mesures acquises dans une campagne de mesures pendant un période de temps donnée. La relation entre les données stockées et les données issues des campagnes, aussi catalogué comme des *données capteurs*, permet l'interrogation des éléments historiques du système, ainsi que les plus récents.

4.3.1 Données capteurs et temps réel

Dans les systèmes de surveillance environnementaux, l'accès aux données implique dans certains cas des propriétés temps réel. Le temps réel doit assurer qu'une tâche est réalisée selon les contraintes de temps spécifiées ; les requêtes doivent être achevées dans un temps limité. Par exemple, en période de crise, les données doivent être disponibles aux utilisateurs dans un délai imparti. Les données capteurs sont par nature temps réel, et les données issues d'applications dans le domaine de la surveillance suivent ces propriétés.

Dans les systèmes de surveillance du type environnemental, les contraintes de temps souples sont les plus adéquates. Ces contraintes sont les moins exigeantes et gênent moins l'exécution du système, toutefois tout non-respect de ces contraintes est à éviter. En comparaison à un système de surveillance nucléaire par exemple, dont les contraintes du temps sont beaucoup plus strictes et aucun délai n'est admissible, les systèmes de surveillance environnementaux ont une criticité plutôt épisodique et le respect des contraintes n'est pas toujours soutenu, car les phénomènes naturels sont souvent imprévisibles. Dans ce cas, les données repérées dans une situation de crise lors

de la détection d'un événement, doivent être prioritaires et les contraintes de temps respectées.

On considère que pour l'analyse de ce type de systèmes, il est plus intéressant d'offrir des délais confortables à l'expert permettant d'accéder à l'information, de manipuler l'application en cas de retard ou manque d'information. Pour les experts de ce domaine, une donnée même en retard est toujours une information valable. Néanmoins, les contraintes de temps doivent être respectées au maximum.

4.3.2 Spécificités du traitement des données spatio-temporelles issues de capteurs

Comme nous l'avons détaillé précédemment, dans un contexte de surveillance, les données capteurs représentent l'ensemble de données géolocalisées stockées dans une base de données ou de données historiques, conjointement aux données en continu ou les plus récentes. La sauvegarde d'une trace historique des éléments mesurés ainsi que l'accès aux mesures les plus récentes sont nécessaires pour permettre à l'utilisateur d'analyser les phénomènes observés.

Ainsi, nous nous sommes intéressée à l'exploitation de ces données pour répondre à des interrogations de type spatio-temporelles en temps réel ou en temps différé, de façon continue ou ponctuelle. En plus, nous nous intéressons aussi à l'historique de ces données, car dans le domaine de la surveillance, ces données permettent l'analyse d'un phénomène au cours du temps. De cette façon, l'analyse du phénomène est beaucoup plus efficace. Dans notre cas, nous considérons que le phénomène à observer est décrit par des données analysées en temps réel et ensuite stockées pour permettre à l'utilisateur leur correcte utilisation, comme c'est le cas des systèmes de surveillance environnementaux (i.e. volcaniques, maritimes...). Par exemple, les experts ont parfois besoin d'analyser l'évolution d'un phénomène grâce à la mesure de plusieurs éléments pendant une période de temps donnée, d'autres ont juste besoin d'exécuter une mesure à un moment de la journée, ou dans une zone en particulier. Certains experts préfèrent avoir la possibilité de choisir un seul capteur pour leurs analyses, et de comparer les données collectées avec les données des journées précédentes, etc.

Nous pouvons décrire ce processus de traitement des données capteurs de façon générale comme suit :

- Les données issues de capteurs (ensemble de mesures ou flux de mesures) sont estampillées selon l'horloge interne du capteur ou du nœud capteur en gardant leur position comme attribut.
- Selon ces estampilles du temps, les données sont sélectionnées en fonction des critères temporels, voire des périodes du temps. Ces données sont donc évaluées et stockées dans des relations temporelles.
- Ces données peuvent être interrogées selon les relations temporelles ou/et les relations spatiales (i.e. région, zone, position, période, instant).
- Le résultat peut être stocké de façon permanente ou selon une temporalité, et notifiés quand une nouvelle donnée est ajoutée à cet ensemble.

- La fréquence d'acquisition au sein du serveur peut être limitée afin d'éviter des surcharges qui peuvent générer des délais. Dans ce cas, l'échantillonnage ou un seuil (min-max) peut réduire la fréquence d'incidence.
- Nous pouvons également ajouter une temporalité à la vie des données gardées en mémoire au sein du serveur afin d'économiser des ressources. Lorsque ces données aboutissent à leur temporalité désignée, elles pourront être stockées de façon définitive dans une base de données, dans un entrepôt ou encore sur disque sous forme de résumés ou d'agrégations.

La spécification du traitement des données spatio-temporelles issues de capteurs doit utiliser un langage pour le décrire. Actuellement, un bon nombre d'approches existent vers le traitement des flux de données comme nous l'avons précédemment mentionné (Section 3.2). L'avantage d'utiliser un langage de type SQL⁸, est qu'il est bien connu et standard, l'optimisation des requêtes et les techniques d'ordonnancement pouvant être directement appliquées. Nous considérons des langages comme PLACE (Pervasive Location-Aware Computing Environments) [Mokbel'05] et SoCQL (Service-oriented Continuous Queries Language) [Gripay'09] adéquats pour la description, afin de permettre unions et jointures entre le temps et l'espace, ainsi que l'interrogation en continu ou ponctuelle des données.

Ce processus de traitement générique, nous permet d'exploiter les données issues des réseaux capteurs et de les gérer de façon adéquate afin d'éviter de possibles divergences au sein du serveur et de la base de données. Le fait de combiner des données entrantes en temps réel avec des données stockées sur une base de données, nous permet de satisfaire plus largement les besoins de l'utilisateur en terme de compréhension du phénomène au monde réel. L'importance dans ce processus est le traitement et l'analyse en temps réel effectué sur les données voire dynamiques conjointement aux données historiques.

4.4 Conclusion

Nous avons présenté dans ce chapitre les spécificités considérées et définies pour la modélisation et la gestion des données spatio-temporelles issues de capteurs dans le domaine de la surveillance environnementale.

Concernant les systèmes de surveillance, nous avons décrit leurs différentes caractéristiques notamment dans le domaine environnemental. Nous avons également défini des couches de traitement dans ce type de systèmes qui sont : les couches *Acquisition*, *Traitement* et *Exploitation*. Ces couches nous permettent de faire un suivi des données et de découper les processus qui les « impactent » à travers ce type de systèmes. Ensuite, nous avons décrit les aspects que nous avons pris en considération dans notre étude, à savoir : le type de réseau utilisé, leurs conditions, etc.

Nous avons détaillé dans ce chapitre (Section 4.2) notre perception des données capteurs et leur gestion en temps réel. Nous soulevons d'ailleurs dans cette partie, la

⁸ SQL – Structured Query Language

relation entre le contexte d'observation et les éléments qui interviennent dans l'observation afin de fournir des informations pour décrire un phénomène. Ceci nous a conduit à la découverte des différentes propriétés spatiales, temporelles et temps réel que les données issues de capteurs expérimentent.

Pour gérer ce type spécifique de données, nous avons proposé tout d'abord un modèle de données capteurs intégrant les éléments concernant le contexte d'observation, ainsi que le réseau de capteurs et les mesures obtenues. Dans ce modèle de données (Figure 15) à la différence des modèles existants, nous avons présenté la structuration des données spatio-temporelles issues de capteurs ainsi que l'interaction des caractéristiques statiques et dynamiques dans leurs informations. Les objets dynamiques concernent les objets changeant de valeur, de forme ou de position dans le temps ou l'espace, et les objets statiques sont ceux qui restent figées tout au long de l'observation. Également, pour compléter ce modèle de données capteurs, nous avons proposé une formalisation des objets relatifs au type de capteurs et leurs mesures. Nous montrons grâce à cette formalisation la relation entre la position des capteurs et la dynamique dans leurs attributs.

Vis-à-vis de la dynamique des attributs, nous avons exposé dans le paragraphe 4.2.2.3, la problématique sur l'identification des données et métadonnées dans un contexte dynamique pour les systèmes de surveillance. Nous avons analysé cette problématique depuis deux points de vue concernant notamment la sémantique et la dynamique de données. Nous partons de l'hypothèse que l'utilisateur peut désigner les informations qu'il considère complémentaires à ses données, ou que les données de forte dynamique seront considérées comme des données et non comme des métadonnées si leur fréquence de variation est vraiment haute.

Finalement, dans la Section 4.3 nous avons abordé différents aspects concernant le traitement des données spatio-temporelles issues de capteurs. Ces aspects sont liés notamment à l'impact du temps réel sur leur traitement. Dans un contexte temps réel, la grande quantité et la dynamique dans les informations issues de capteurs rend la tâche de gestion de la qualité de ces données plus difficile. Une telle gestion est limitée en raison de la capacité de stockage et de l'énergie des capteurs. Dans de telles situations, la panne de capteur ou la présence de mauvaises situations environnementales peut mettre en danger l'intégrité de l'information. Il est ainsi nécessaire de conserver les données dans des bases de données centralisées pour satisfaire toutes les conditions. Des spécificités concernant ces aspects ont été détaillées.

Selon les spécificités décrites dans ce chapitre, nous proposons dans la section suivante une méthodologie pour la définition de la qualité de données issues de capteurs. Cette méthodologie a pour objectif de fournir à l'utilisateur un moyen d'évaluation et de communication de la qualité des données issues de capteurs en temps réel. Cette méthodologie entame une étude plus approfondie sur les facteurs qui influencent la qualité des données issues de capteurs dans le domaine de la surveillance environnementale. Elle propose également un modèle de qualité basé sur des critères et indicateurs adaptés aux spécificités des données issues de capteurs dans les systèmes de surveillance. Nous proposons d'exploiter les *métadonnées* afin d'évaluer la qualité des données. Finalement, nous abordons les différentes fonctions adaptées pour l'évaluation

de la qualité des données spatio-temporelles issues de capteurs, ainsi qu'une proposition de communication de ces informations.

Chapitre 5

Définition et évaluation de la qualité des données capteurs

Nous avons identifié lors de notre étude de l'état de l'art les différents concepts liés à la qualité des données, ainsi que les différentes méthodologies existantes pour l'évaluation de la qualité des données au sein des systèmes d'information. Nous avons également, analysé les caractéristiques des réseaux de capteurs, les données issues de capteurs et leur application dans les systèmes de surveillance. Cette étude nous a permis de distinguer les aspects potentiels qui pourraient « impacter » la qualité des données capteurs lors de leur acquisition et ainsi aboutir à une évaluation de la qualité.

Suite à cette étude, nous avons conclu qu'il n'existe a priori aucune méthodologie qui permette d'évaluer en temps réel la qualité des données spatio-temporelles issues de capteurs. De ce fait, nous avons été amenée à proposer une méthodologie plus spécifique qui permette l'analyse, la modélisation, l'évaluation et la communication de la qualité des données géolocalisées issues de capteurs.

L'idée que nous défendons dans cette troisième partie de ce manuscrit, est que la qualité des données issues de capteurs dépend principalement de la qualité dans les sources de données et du processus de traitement que ces données subissent. Pour évaluer la qualité des données issues de capteurs dans un système de surveillance, nous devons prendre en considération la qualité du contexte d'acquisition, ainsi que les processus utilisés pour extraire, intégrer et transmettre les données aux utilisateurs finaux. Afin d'aider l'utilisateur lors de la prise de décision, nous proposons une méthodologie pour la définition de la qualité des données dans ce contexte et qui fournis à l'utilisateur un moyen d'évaluation et communication de la qualité des données issues de capteurs exploitées. Notre proposition de méthodologie au regard des méthodologies existantes, ne s'attache pas seulement à la détection d'erreurs de la mesure. Cette méthodologie fait l'étude des spécificités du contexte d'acquisition et du traitement des données qui peuvent avoir un impact sur la qualité de ce type de données.

Nous considérons que la qualité des données d'un point de vue intrinsèque (ou interne), n'est pas suffisante pour décrire la qualité d'un système comme celui de la surveillance environnementale. Ceci nous amène à la prise en considération et à l'adaptation de certains critères et mesures (standardisés), notamment dans le domaine de la qualité des données géographiques et de la qualité de service (QoS). Notre proposition d'évaluation est également basée sur la gestion des informations qualité dans des métadonnées exploitées en temps réel. En considérant les phases classiques des méthodologies existantes (cf. Section 2.1.3), cette méthodologie se décline en trois phases (Figure 18) présentées ci-dessous :

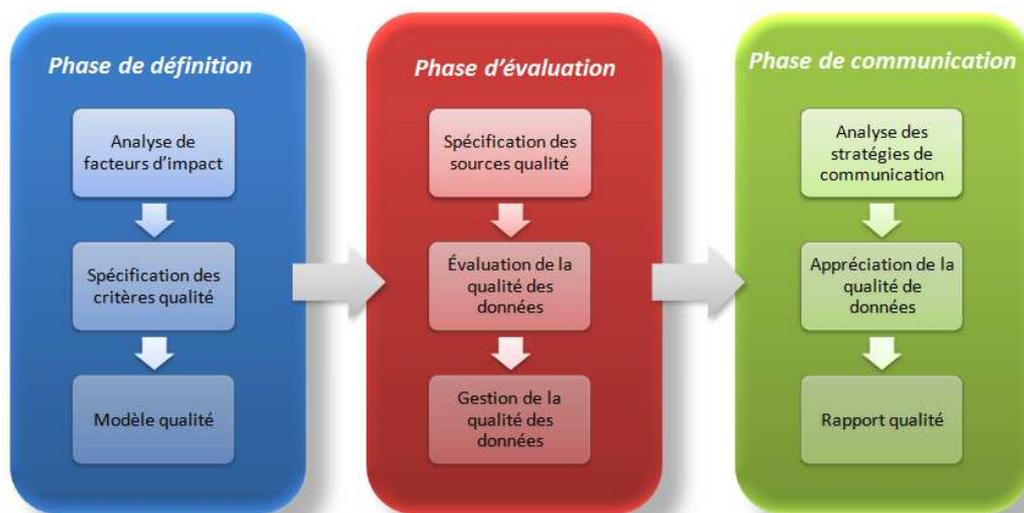


Figure 18 – Méthodologie d'évaluation de la qualité des données spatio-temporelles issues de capteurs

- **Phase 1 : la définition.** Cette phase représente l'étude de la qualité des données issues de capteurs dans le contexte de systèmes de surveillance. Elle est caractérisée par trois étapes importantes : l'analyse des facteurs d'impact de la qualité de données capteurs, la spécification des critères et des mesures qualité et la modélisation de la qualité des données issues capteurs.
- **Phase 2 : l'évaluation.** Nous obtenons suite à la phase précédente un ensemble de critères qualité à évaluer. Dans cette phase, nous effectuons la sélection et la spécification des sources d'information qualité (métadonnées qualité). Ensuite, nous fournissons les fonctions d'évaluation pour chaque critère qualité proposé, ainsi que leur définition et technique utilisée pour leur calcul au regard de notre contexte d'application. Cette étape consiste en 3 étapes principales : la spécification des sources qualité, la spécification et l'évaluation des critères qualité et la gestion de l'information qualité.
- **Phase 3 : la communication.** Cette phase a pour objectif de communiquer à l'utilisateur les informations résultantes de la phase précédente. Pour cela, nous nous basons sur la visualisation en temps réel des indicateurs qualité et l'exploitation des rapports qualité. Les indicateurs qualité correspondent à des mesures relatives aux critères qualité sélectionnés et le rapport qualité implique l'ensemble d'informations contextuelles et résultantes lors de l'évaluation de la qualité de données à un instant ou période donné.

Dans ce cinquième chapitre nous commençons dans la Section 5.1 par la phase de définition de la qualité de données spatio-temporelles issues de capteurs. Ensuite, la Section 5.2 est destinée à présenter la description de la phase d'évaluation de la qualité des données capteurs et ses composantes. Pour finaliser la description de cette méthodologie, la Section 5.3 spécifie les éléments pris en considération pour la

communication de la qualité de ces données en temps réel. Nous concluons les trois phases proposées au sein de cette méthodologie dans la Section 5.4.

5.1 Phase 1 : définition de la qualité des données issues de capteurs

Dans cette phase, nous étudions de façon plus approfondie les facteurs « impactant » la qualité des données issues de capteurs et leur relation avec les critères de qualité existants. Plus spécifiquement, nous décrivons les facteurs qualité selon les processus génériques qui interviennent dans le traitement des données au sein d'un système de surveillance. Ces éléments sont associés aux données dans un modèle de qualité.

La sélection des facteurs qualité appropriés implique également la sélection des métriques qui estiment de tels facteurs qualité. Dans notre approche, ces critères et mesures sont basés sur une approche, résultant de notre étude au sein des réseaux de capteurs et les besoins en termes de qualité dans les systèmes de surveillance, notamment environnementaux, exprimés par les recherches existantes (NOAA, TAO, SWE...).

Cette phase de définition est composée principalement par les étapes suivantes (Figure 19) :

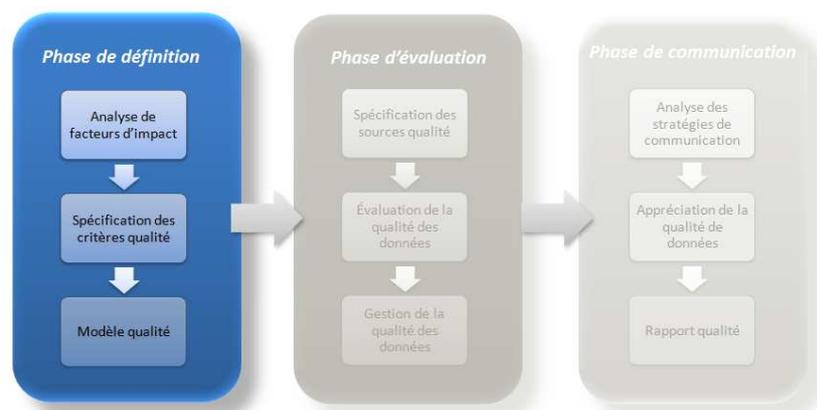


Figure 19 – Phase 1 : Définition de la qualité des données issues de capteurs

- a) L'analyse des facteurs qui impactent la qualité des données selon le système d'acquisition, de traitement et d'exploitation.
- b) La spécification des critères qualité associés aux données issues de capteurs.
- c) La spécification du modèle qualité adéquat aux données issues de capteurs.

Nous allons décrire, dans ce qui suit, chacune de ces étapes.

5.1.1 Analyse des facteurs d'impact de la qualité des données issues de capteurs

L'objectif de cette étape est d'analyser les facteurs d'impact sur la qualité des données issues de capteurs (Figure 20). Nous présentons plus en détail, les caractéristiques de chaque facteur (contextuel, du traitement ou d'exploitation) et leur possible impact sur la qualité des données capteurs.

Dans les capteurs et les réseaux de capteurs utilisés pour la surveillance environnementale (i.e. géocapteurs), principalement les réseaux de capteurs sans-fil, les données sont exposées à diverses sources d'erreurs pendant les phases de traitement au sein du système : d'acquisition, de transformation et de communication.

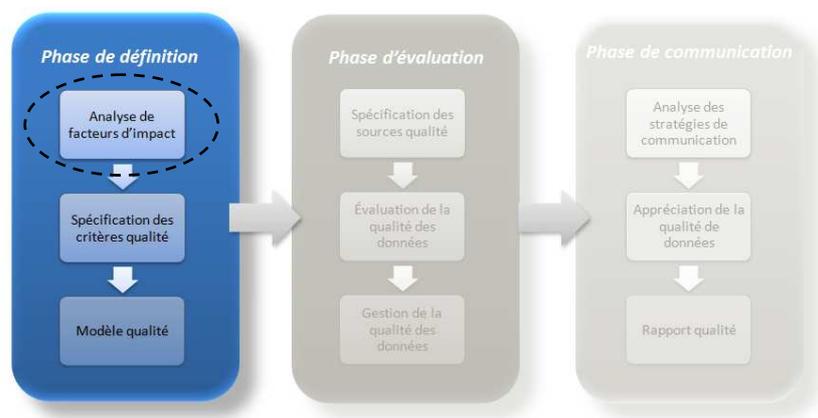


Figure 20 – Phase 1 – Analyse de facteurs d'impact de la qualité des données capteurs

On part du constat qu'il est difficile de garantir la qualité intrinsèque de la donnée, mais on peut déterminer a priori les sources d'erreur et les facteurs qui l'impactent.

Pour analyser les possibles facteurs qualité, nous prenons les trois couches de traitement proposées dans la Section 4.1.1 : *la couche d'Acquisition, la couche Transformation et la couche Exploitation* (Figure 21).

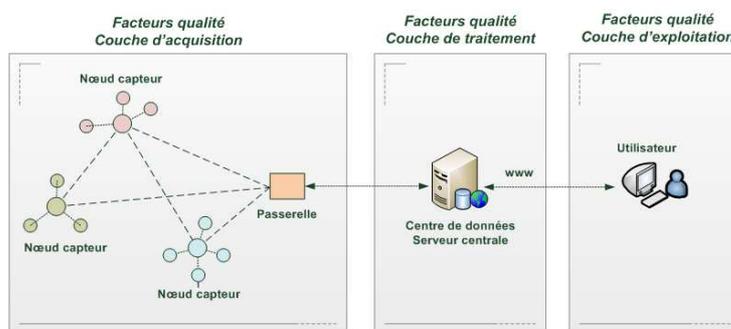


Figure 21 - Étapes de sélection des facteurs qualité dans le système de surveillance

Ces couches sont composées par différents processus qui ont une influence directe ou indirecte sur l'état des données et représentent des sources potentielles d'erreurs. Nous décrivons ce processus ensuite.

5.1.1.1 Processus génériques et facteurs qualité

Nous définissons un *processus générique*, comme étant tout processus qui intervient dans la construction des données dans un système exploitant des données capteurs, spécialement dans les systèmes de surveillance environnementaux. Dans ce contexte, nous distinguons trois types de processus (Figure 22) :

- *Processus d'acquisition des données,*
- *Processus de transformation des données,*
- *Processus d'exploitation des données.*

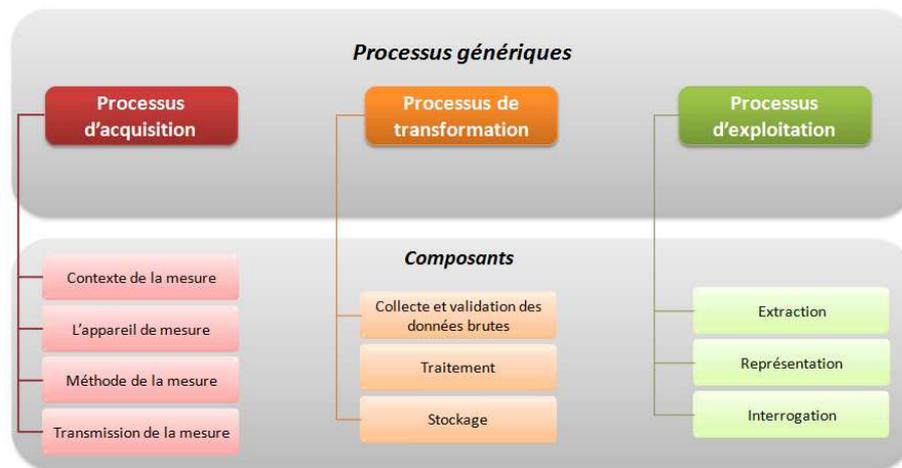


Figure 22 – Analyse des facteurs qualité selon les processus de traitement de données capteurs

Chacun de ces *processus* est constitué par différents *composants* qui nous permettront de faire une détection plus spécifique d'une possible erreur au sein du système. Celles-ci aboutissent à l'identification des facteurs qualité. Nous appelons toute condition ou aspect qui diminue ou « impacte » la qualité des données issues de capteurs, *facteur qualité*. Chaque facteur est lié à des caractéristiques spécifiques qui décrivent la qualité des données capteurs. Notre schéma d'analyse est présenté dans la Figure 22.

Pour mieux comprendre les facteurs qualité attachés à chaque processus générique, nous présentons ensuite les caractéristiques de leurs composants et les facteurs qualité associés. L'objectif principal est de sélectionner les critères et les mesures qualité adaptés à notre contexte d'étude.

5.1.1.1.1 Processus d'acquisition de données et facteurs qualité

Dans ce processus les données du monde réel sont collectées. L'acquisition dans notre contexte, se fait grâce à des capteurs individuels ou en réseau qui nous permettent d'obtenir les données nécessaires pour la description du monde réel. Les données obtenues dans ce processus sont désignées comme brutes et dépendent notamment du type d'application.

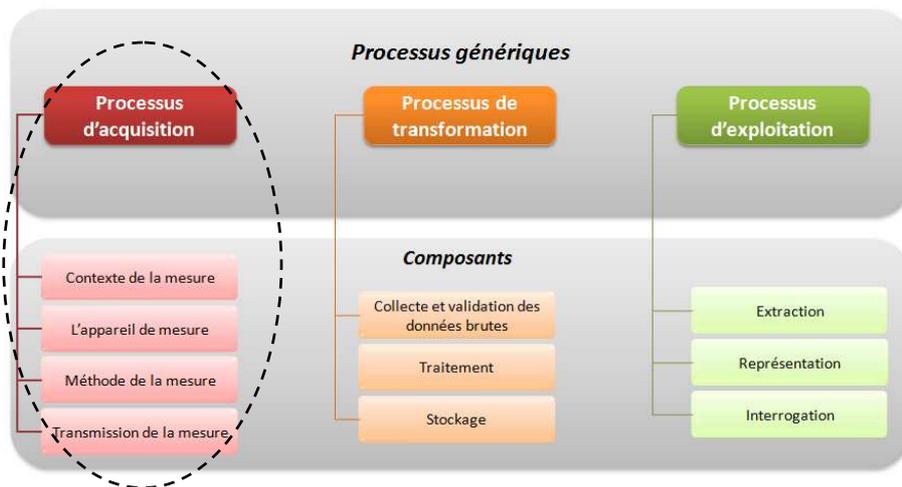


Figure 23 - Processus d'acquisition et composants

Nous considérons que le processus d'acquisition des données est caractérisé par quatre aspects principaux agissant sur ces données et sur leur qualité : *le contexte de récupération des mesures, le capteur, les méthodes de mesure et la communication* (Figure 23).

a) Contexte de récupération des mesures

Dans la littérature [Dey'99], le terme *contexte* a été utilisé pour définir toute information qui peut être utilisée pour caractériser la situation d'une entité. Dans ce cadre, une entité peut être une personne, un endroit, un objet (ou élément) qui est considéré comme relevant pour l'itération entre l'utilisateur et l'application [Dey'99]. Selon cette définition, l'information contextuelle peut inclure une identité, l'information spatiale, l'information temporelle, les mesures physiques, situations critiques ou dangereuses, etc. Typiquement, les applications *contexte-awareness* utilisent ce type d'information pour réguler le comportement des éléments dans un système.

Dans notre contexte d'étude, le contexte de récupération des mesures représente l'environnement dans lequel le dispositif de mesure est déployé. Plus spécifiquement, ce contexte évoque les conditions environnementales et d'opération dans lesquelles le capteur ou le réseau de capteurs effectue la prise de mesure.

Les réseaux de capteurs peuvent être déployés à distance dans des environnements hostiles tels que les volcans. Dans ces cas, les nœuds ne sont souvent pas protégés contre les mauvaises conditions environnementales (tempêtes de neige, inondations, hautes températures, etc.) ou d'opération du capteur (hautes températures,

excès d'humidité...), ou encore se protéger contre les attaques malveillantes puisque n'importe qui pourrait avoir accès à l'endroit où ils sont déployés (animaux ou personnes).

b) Le capteur

Le capteur est un dispositif qui transforme les signaux physiques du monde réel en données. Des nos jours, ces appareils (i.e. *motes*) ont des capacités pour mesurer, calculer et communiquer les mesures acquises. La détection exécutée par ces capteurs et la transmission des données acquises sont un processus automatique.

Selon les capacités et l'autonomie des appareils de mesure, nous avons deux aspects importants à considérer : l'état physique du nœud/capteur et la limitation des ressources (d'énergie, puissance de calcul et la capacité de stockage).

Concernant l'état physique du nœud capteur nous considérons que le temps de vie du capteur, son calibrage, la dégradation de l'appareil, etc., sont des aspects qui peuvent impacter la qualité des données lors de la prise de mesure. Si un capteur présente des inconsistances causées par un mauvais réglage ou par la dégradation résultat de l'usage ou des conditions environnementales, leur effet sera visible au moment de l'exploitation des données. Par exemple, des fausses données ou données manquantes peuvent être détectées lorsqu'un capteur est abimé.

Concernant les ressources du capteur, nous considérons dans notre approche, que tous les capteurs ont la capacité de stocker, traiter et communiquer les données qu'ils produisent (considérés comme des capteurs intelligents). Cependant ces capacités présentent des limitations notamment sur leur énergie, la capacité de calcul et de stockage, leur fiabilité, etc. Dans ce cas, le fait que les capteurs soient typiquement alimentés par des batteries (souvent non renouvelables), la probabilité qu'ils manquent à des tâches prévues n'est pas négligeable. De plus, les capteurs ne possédant pas une capacité de calcul et de stockage adaptable pour traiter de grandes quantités de données, ceci peut générer des pertes des données ou des transmissions forcées au serveur central.

c) Méthode de mesure

Les mesures et les méthodes de mesure des données consistent à quantifier les objets aperçus dans le monde réel, soit par l'énumération ou par l'attribution d'une valeur numérique. Les mesures représentent la valeur particulière d'une donnée dans une échelle de mesure et les sources potentielles d'erreur sont issues des méthodes utilisées pour la mesurer.

Généralement, les capteurs mesurent un seul élément (i.e. humidité, température...) à la fois, en considérant certaines exceptions (i.e. capteur RDI – pression, champ magnétique, température). Ceci est dû au fait que la taille, l'énergie et la capacité de traitement au sein du capteur sont limitées. Ces aspects incitent à utiliser des méthodes de mesure peu coûteuses en termes de capacités (activation par périodes de temps, l'échantillonnage, l'activation par détection des seuils...) en dépend de la fiabilité de la donnée en termes de précision et exactitude de résultats.

d) Communication (Transmission de la mesure)

Une fois les données collectées (acquises et validées), la communication représente le moyen de transmission de ces données vers le traitement, afin que les données soient préparées pour leur exploitation.

La communication englobe différents types de communications possibles (i.e. sans-fil, radio,...). Dans le contexte de surveillance, les capteurs/nœuds capteur utilisent notamment, une communication sans-fil pour communiquer les données acquises à un nœud central ou à un serveur. Cependant, cette communication peut être endommagée à cause de mauvaises conditions environnementales, les faibles ressources du capteur ou le partage d'une même chaîne de communication avec des autres capteurs au sein du réseau. Une mauvaise communication peut générer des données manquantes, des données en retard ou l'indisponibilité des sources. Ces aspects impacteront de façon importante la qualité de l'analyse des données en temps réel.

Après la description de chaque composant dans le processus générique d'acquisition et ses caractéristiques, nous considérons important de définir les facteurs dans chaque composant, qui sont concernés au détriment de la qualité des données capteurs. L'ensemble de facteurs sont décrits comme suit.

- Facteurs d'impact qualité

Le processus d'acquisition implique les facteurs qualité suivants (Tableau 7) :

	Composants	Facteurs
Processus d'acquisition	Contexte de récupération des données	<ul style="list-style-type: none"> • Contexte environnementale • Contexte d'opération • Attaques malveillantes
	L'appareil de mesure	<ul style="list-style-type: none"> • État de l'appareil • Autonomie • Ressources limités (énergie, calcul, stockage)
	Méthode de mesure	<ul style="list-style-type: none"> • Types de mesure • Méthode de mesure
	Communication (transmission)	<ul style="list-style-type: none"> • Type de communication et protocoles • Conditions environnementales • Fréquence de transmission

Tableau 7 – Facteurs d'impact qualité : Processus d'acquisition

i. Conditions environnementales ou d'opération

Les réseaux de capteurs sont souvent déployés dans des zones à risque, dont les conditions de fonctionnement d'un capteur ou d'un réseau de capteur pourront être affectées par les conditions environnementales (i.e. inondations, neige, orage, séismes, etc.), ou dont les conditions d'opération du capteur peuvent ne pas être les optimales (i.e. températures élevées, excès d'humidité, obstacles, etc.). Ces facteurs ont un impact important sur la qualité du fonctionnement des capteurs, et par conséquent sur les mesures qu'ils réalisent. De données erronées (❄️-30°C → capteur sous la neige) ou fausses assumptions (i.e. capteur en panne) peuvent être produites.

ii. *Attaques malveillantes*

Dans certains contextes de surveillance, par exemple routes ou champs de culture, les capteurs peuvent être attaqués par des agents externes qui peuvent endommager leur équipement ou perturber leur fonctionnement. Les animaux sauvages sont un bon exemple. Souvent, ils agressent les capteurs pendant leurs parcours, ou changent la position, l'orientation ou affectent la sensibilité du capteur. Évidemment, ces aspects vont effectivement modifier la perception du monde réel.

iii. *L'état du capteur*

Les défauts dans les appareils de mesure peuvent se présenter comme des actions incorrectes. Malheureusement, ces défauts sont communs et peuvent être dus à différentes causes comme : le mauvais fonctionnement, le haut étalonnage ou les erreurs d'installation, ou simplement la dégradation de l'appareil.

iv. *Ressources limitées*

Comme nous l'avons soutenu tout au long de ce manuscrit, les capteurs possèdent de plus en plus une flexibilité et une mobilité importantes. Cependant, nous estimons que les facteurs suivants peuvent impacter la qualité des données fournies par les capteurs :

L'énergie. Les capteurs sont généralement fournis avec des batteries non-renouvelables et le changement de ces batteries est souvent très délicat, voire impossible selon la localisation du capteur, le milieu d'observation et le type de réseau utilisé. D'un autre côté, es capteurs utilisant des batteries renouvelables sont souvent plus limités en termes de calcul et stockage.

La capacité de calcul et stockage. Typiquement, les capteurs possèdent des micro-processeurs souvent peu performants (~ 4 Mhz) et ne possèdent pas une capacité de stockage persistante (~ 512 bytes de mémoire SRAM). Ces aspects réduisent la capacité et la quantité de traitement exécuté au sein des capteurs. Ceci peut générer une perte d'information, de mauvais calculs, l'indisponibilité du capteur, l'envoi de données au serveur central sans demande directe.

v. *Types de mesure*

Les capteurs actuels ont la capacité de mesurer différents éléments, en générant des mesures de différents types. Chaque mesure doit être paramétrée à chaque fois selon son type (échelles, tolérance, fréquence de mesure...). Ceci peut représenter une contrainte en termes de ressources et de performance du capteur, car les ressources s'épuisent plus rapidement en mesurant plusieurs éléments au même temps. Dans le cas où le capteur réalise un seul type de mesure, une surveillance est également nécessaire en termes des seuils concernant la tolérance, les fréquences. Ces éléments sont facteurs qui peuvent impacter la qualité des données acquises.

vi. *Méthode de mesure*

Par méthode de mesure, nous entendons la façon dont les mesures sont prises. Par exemple, de façon manuelle (par l'expert et de façon ponctuelle), automatique

(prédéfinie pour un période de temps défini ou indéfini ou déclenché par un événement) ou semi-automatique (prédéfini partiellement et avec une intervention manuelle).

Les capteurs de nos jours, dits intelligents, ont une autonomie importante, donc la prise de mesure peut être réalisée de façon automatique ou semi-automatique (selon le type d'application) et prédéfinie en avance. Nous considérons que ces types de méthodes économisent des efforts, mais aussi impliquent un investissement plus fort des ressources.

vii. Type de communication

Au sein des réseaux de capteurs la communication peut être de divers types comme sans-fil (la plus réponde de nos jours), microondes, etc. Dans le cas particulier du sans-fil, cette communication peut être du type synchrone et asynchrone. Dans ce cas, des problèmes concernant les protocoles de communication, l'indisponibilité du nœud ou serveur (selon le cas) peuvent générer des congestions et latence au niveau de la livraison des paquets des données. Dans ce cadre, nous pouvons signaler que les conditions environnementales comme, le brouillard, les tempêtes, etc. peuvent affaiblir également la puissance de la communication et engendrer la perdre d'information.

viii. Fréquence de transmission

Pareillement aux aspects précédents, nous considérons que la périodicité de transmission des données est aussi un facteur à considérer. Si elle est trop assez fréquente, les ressources au sein du capteur risquent de s'épuiser.

5.1.1.1.2 Processus de transformation et facteurs qualité

Le processus de transformation des données capteurs implique différents aspects qui constituent les étapes intermédiaires entre les données à l'état brut, telles qu'elles ont été collectées, et la représentation des objets tels qu'ils sont dans le modèle conceptuel du monde réel.

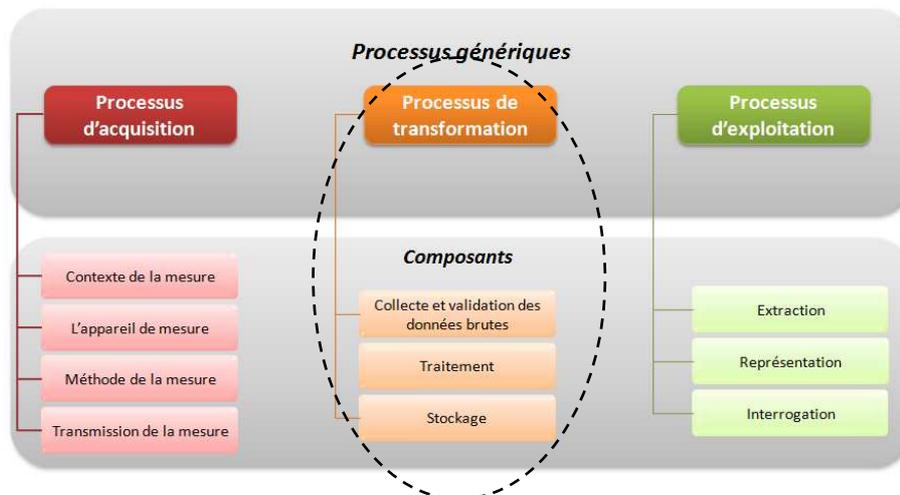


Figure 24 - Composants du processus de transformation

Dans notre cas, nous considérons la collecte et la validation des données, le traitement et le stockage comme composants principales du processus de transformation (Figure 24).

La collecte et la validation des données peuvent être effectués au sein des nœuds capteurs et ensuite transmises au serveur central. D'autres étapes, comme l'agrégation ou le stockage des données sont complémentaires. Il est important de noter que la dynamique dans les données dans ce contexte demande aussi un traitement dynamique exploitable notamment en temps réel. Dans ce cas il n'y a pas que les données qui peuvent être considérées comme dynamiques, mais aussi leur analyse et transformation.

Afin d'éviter des surcharges ou des traitements inutiles au sein du serveur, les données produites par le capteur ou le réseau de capteurs peuvent être prétraitées au sein du réseau avant d'être envoyées à un centre de données. Cependant, ce filtrage doit être réalisé avec un seuil adapté, car des informations réelles et d'importance pour l'observation peuvent être supprimées. Toutefois, dans les applications de surveillance, le plus grand nombre de données est envoyé au centre de données pour effectuer plus de traitement. Ce traitement peut être exécuté principalement au niveau du serveur en mémoire vive. Ceci produit des données exploitables, qui sont ensuite stockées dans une base de données ou dans un entrepôt de données pour de futures transformations.

Le résultat final du processus de transformation des données est une représentation des objets qui sera finalement exploitée en mémoire vive au sein du serveur ou stockées dans une base de données ou un entrepôt de données. Les divers processus de transformation jouent un rôle très important dans la détermination de la qualité des données issues de capteurs et c'est dans ce processus, plus spécifiquement, que les erreurs ont tendance à se propager.

Nous étudions ensuite ces aspects, afin de déterminer les divers facteurs qui peuvent influencer une mauvaise qualité dans les données.

a) Collecte et validation des données capteurs

La collecte des données capteurs fournit la source d'information principale pour la représentation du monde réel. La collecte est une phase générale qui réceptionne toutes les données émises lors de l'acquisition.

La validation des données est le complément de la collecte des données, et elle est représentée de façon générale, comme une phase de sélection des données à exploiter. La validation des données capteurs représente la vérification et le filtrage de données acquises. Cette validation est faite en regardant les spécifications de l'application (i.e. min, max); cet aspect est indispensable afin de limiter le traitement des données bruyantes. Cependant, dans notre approche nous préférons accepter toute information étant validée ou non validée et les étiqueter selon leur comparaison avec les valeurs de référence, ceci afin d'éviter une perte d'information.

b) Traitement des données capteurs

Ce processus permet de réaliser les traitements possibles sur les données brutes ayant un format (valeur et attributs). Les traitements effectués constituent les étapes

intermédiaires entre les données à l'état brut et les données telles qu'elles vont être exploitées.

Les traitements dépendent en premier lieu des spécifications des besoins, de la nature des données, de leur distribution et des facteurs annexes comme la fréquence d'arrivée au sein du traitement. Par exemple, un ou plusieurs capteurs peuvent produire des données sous forme de flux en périodes définies ou indéfinies du temps (fenêtres du temps) ou en réponse au déclenchement d'une campagne de mesures (i.e. trigger). Dans ce cas, il est nécessaire d'effectuer des transformations multiples sur des sources de données diverses, qui ont une grande variation de valeur, ou qui ont besoin d'être analysées le plus vite possible (i.e. en situation de crise). Le grand volume des données peut alors causer des collisions et contribuer à une surcharge et une dépense excessive des ressources au sein du serveur. Dans ces conditions, certaines contraintes sur les tâches ou le temps, peuvent manquer et engendrer des faux résultats.

Nous considérons que les sources possibles d'erreurs sont liées à la quantité de sources de données, aux ressources disponibles au sein du serveur, à la synchronisation temporelle et les heuristiques.

c) Stockage des données capteurs

La grande quantité de données capteurs produites dans un réseau nécessite de préférence, un traitement et un stockage centralisés (au serveur central), afin d'éviter des pertes d'information suite à des pannes de capteurs lors de l'acquisition et pour économiser des ressources au sein de la mémoire vive. Dans ce processus, les données issues des capteurs possèdent une période de validité qui détermine leur actualité et ensuite détermine leur stockage.

Dû à la dynamique d'acquisition et du traitement des données capteurs, le stockage peut être aussi impacté par cette dynamique. Par exemple, des données peuvent être destinées à être stockées dans la base de données, sans avoir vérifié leur format et validité. Également, ces données peuvent demander le stockage avec une fréquence relativement rapide qui peut générer des inconsistances dans la base de données. Ces aspects peuvent rendre l'information non-disponible ou manquer d'attributs pour leur exploitation (i.e. métadonnées), ce qui nuit à la qualité des données traitées.

Suite à la description des composants dans le processus de transformation de données au sein d'un système de surveillance et leur implication sur la qualité des données exploitées, nous décrivons les facteurs qualité que nous considérons importants dans l'évaluation de la qualité des données issues de capteurs.

- Facteurs d'impact qualité

En conclusion dans ce processus de transformation, nous distinguons les facteurs qualité suivants (Tableau 8) :

	Composants	Facteurs
Processus de transformation	Collecte et validation des données capteurs	<ul style="list-style-type: none"> • Prétraitement • Validation
	Transformation des données capteurs	<ul style="list-style-type: none"> • Nature des données (hétérogénéité) • Volume de données • Ressources physiques • Temps de transformation
	Méthode de mesure	<ul style="list-style-type: none"> • Types de mesure • Méthode de mesure
	Stockage des données capteurs	<ul style="list-style-type: none"> • Format de données • Fréquence de stockage • Volume de données • Métadonnées • Mise à jour

Tableau 8 – Facteurs d’impact qualité : Processus de transformation

i. *Prétraitement*

Dans la collecte des données au niveau du serveur, un prétraitement est effectué afin de structurer les données selon un format et les faire valider. Dans ce cadre, le taux de collection de ces données et le volume de données à prétraiter est important. Un taux de collection trop rapide peut amener à une perte d’information, et un grand volume de données peut demander un temps considérable pour l’agréger. Des pertes de données et des délais de temps sont possibles. Des formats non standardisés ou mal spécifiés peuvent créer des conflits.

ii. *Validation*

Typiquement, les problèmes au sein de la validation des données concernent la prise en considération de méthodes de validation inadéquates ou manquantes, ainsi que la validation sur un grand volume de données. Par exemple, quelques conséquences de cet aspect sont la validation des fausses données, ou l’absence d’une vérification. Aussi, la vérification des grands volumes de données peut générer une prise du temps considérable pour effectuer cette validation.

iii. *Nature des données (hétérogénéité)*

L’hétérogénéité dans les sources des données génère également une hétérogénéité dans les données à transformer. Chaque type de données possède des spécificités dans son type de traitement. Par exemple, les flux des données issues de capteurs sont traités de façon différente des données traditionnelles. Le facteur temps, leur formalisme, les affectations de priorités, etc., sont particuliers pour chaque type de donnée. Des erreurs peuvent surgir si le système n’est pas adapté à ce type d’aspects.

iv. *Volume de données*

Le grand volume de données représente implicitement une problématique dans le système et les processus impliqués. Si le volume est considérable, ceci peut contribuer à une surcharge dans le système et perturber les processus de traitement utilisés. Des défaillances du système et des pertes d’information sont donc probables.

v. *Ressources*

Les ressources sont typiquement liées à la robustesse du système, mais sont également liées au manque d'entretien et d'amélioration du système. Ceci est notamment présent dans les systèmes de surveillance à longue durée. Si les tâches programmes pour le système ne sont pas mises en adéquation avec le type d'application, les ressources au niveau du système sont typiquement réduites.

vi. *Temps de transformation*

Le traitement au sein des systèmes de surveillance est fait en temps réel (soft), c'est-à-dire, les sources d'exploitation des données doivent être prêtes pour satisfaire les requêtes futures. Ceci implique que chaque donnée ou flux de données doit être traité le plus rapidement possible, selon leur priorité. Un temps de transformation trop lent peut générer des absences de données ou des délais dans la mise à disposition des informations. Un temps inconsistant (trop variable) peut générer une mauvaise synchronisation dans le système.

vii. *Format des données*

Au niveau du stockage des données capteurs, il est probable de trouver des problèmes liés notamment aux modèles et structures de données inappropriés. Ceci peut générer des inconsistances au niveau de la base de données et la perte d'information.

viii. *Type de stockage*

Le stockage des données au sein d'un système de surveillance peut s'effectuer de façon temporelle au sein d'un serveur en mémoire vive et ensuite les données seront stockées au sein d'une base de données ou sur disque. Dans ce cadre, les données (flux ou ensembles) entrantes peuvent arriver avec une fréquence assez rapide, et une fréquence non adaptée à leur stockage (trop rapide ou trop lente). Cet aspect peut générer des files d'attente longues et des problèmes de synchronisation.

Également, selon le type de stockage, des contraintes matérielles peuvent arriver. Si ce type de stockage n'est pas adéquat à la quantité des données à stocker ainsi qu'à leur dynamique, des limites au sein de la base de données et du serveur peuvent engendrer la saturation de la mémoire ou de la base de données et générer la perte des données ou l'arrêt du système.

ix. *Métadonnées*

Les métadonnées au sein de n'importe quel système sont une ressource importante pour la compréhension des données utilisées. L'absence de métadonnées, ainsi que leur excès peut provoquer une dégradation dans le partage ou l'exploitation des données.

x. *Mise à jour*

La mise à jour des données stockées garantit la validité de l'information. Toute absence ou inconsistance dans la mise à jour peut affecter la compréhension et

l'exploitation des données. Également, d'un point de vue plus stochastique, des duplications des données, la suppression des données et le coût excessif de ressources peuvent être produits.

5.1.1.1.3 Processus d'exploitation et facteurs qualité

Pendant le processus d'exploitation, les données sont extraites de la base de données ou de la mémoire vive du serveur et mises à disposition à l'utilisateur. L'utilisateur pourra ainsi prendre la décision la plus adéquate selon ses besoins et le type d'application qu'il est en train d'exploiter.

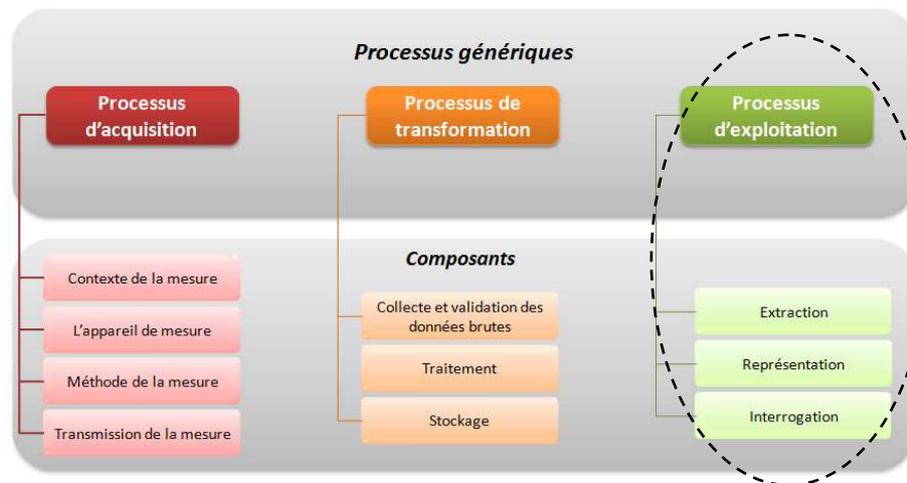


Figure 25 - Composants du processus de transformation

La propagation des incertitudes tout au long des processus de traitement rend au moment de l'exploitation la prise de décision de l'utilisateur plus difficile. Plus spécifiquement, dans la partie exploitation, certaines erreurs peuvent survenir de la part des techniques d'extraction de l'information et des outils utilisés pour représenter l'information. L'exploitation des données présente donc, l'extraction de l'information et l'exploitation de l'information comme principaux composants (Figure 25).

a) L'extraction et représentation de l'information

Le processus d'extraction de l'information consiste en la découverte de l'information implicite non connue a priori, mais utile à partir des données obtenues. Ce processus a pour objectif, l'extraction, notamment automatique, d'informations qui peuvent être catégorisées de façon contextuelle ou sémantique. L'extraction des données permet la découverte de données non-structurées dans une forme exploitable. Par exemple nous pouvons extraire des données d'une base de données spatiale et les représenter de façon graphique dans un plan cartographique construit à partir de ces données. Également, nous pouvons extraire des données issues d'un flux de température et les représenter sous forme de graph, laquelle est mise à jour tous les seconds. Diverses techniques sont utilisées à ces fins, comme la fouille de données ou

les simples interpolations. Ce processus est notamment utilisé dans les outils d'aide à la décision.

Les principales problématiques en termes de qualité sont liées aux modèles de représentation de l'information (mauvaise reconnaissance, mauvaise temporalité, mauvaises relations...), les méthodes d'analyse des données, l'adaptation aux besoins des utilisateurs et les interfaces visuelles. Concernant ces dernières, la représentation de l'information extraite dans des interfaces visuelles, représente un défi car les besoins des utilisateurs et de l'application en termes de visualisation peuvent varier d'un utilisateur à l'autre. Par exemple, dans certains cas, l'utilisateur est plus confortable avec l'utilisation du langage naturel sous la forme de textes ou des documents électroniques. D'autres cas considèrent que les objets visuels peuvent communiquer et déclencher plus rapidement des actions à court terme (i.e. indicateurs d'alerte). Egalement, nous considérons que le manque d'annotations dans les objets, l'utilisation des mauvaises gammes de couleurs, les descriptions trop exhaustives peuvent nuire à la compréhension de l'utilisateur.

b) Interrogation de l'information

L'interrogation des données capteurs est la clé pour tout utilisateur désirant fonder ses décisions sur une information validée. Cette exploitation et interrogation est généralement faite par l'utilisateur final. Nous pouvons ainsi considérer que des erreurs en termes d'interrogation peuvent survenir du facteur humain. Ce type d'erreurs est souvent expérimenté dans ce processus, notamment en temps réel. Le manque d'expérience de l'utilisateur et sa faible familiarisation avec les données et les outils, peut engendrer une mauvaise interprétation et par conséquent une mauvaise exploitation de l'information.

Après cette étude des processus et composantes impliquées dans l'exploitation des données, nous concluons sur différentes aspects qui peuvent être représentés comme des facteurs qualité ; c'est-à-dire, des caractéristiques spécifiques qui peuvent impacter la qualité des données exploitées.

Dans le paragraphe suivant, nous détaillons les facteurs qualité définis lors du processus d'exploitation, ainsi que leur impact dans la qualité des données capteurs.

- Facteurs d'impact qualité

L'exploitation de l'information implique l'accès et l'interrogation de l'information nécessaire et utile pour la prise de décision de l'utilisateur.

	Composants	Facteurs
Processus d'exploitation	Découverte de l'information	<ul style="list-style-type: none"> • Type d'extraction • Méthode d'analyse • Modèle de représentation
	Exploitation de l'information	<ul style="list-style-type: none"> • Facteur humain • Interprétation de l'information

Tableau 9 – Facteurs d'impact qualité : Processus d'exploitation

Dans ce composant, nous détectons les facteurs suivants comme étant impliqués dans la qualité des données exploitées (Tableau 9).

i. Extraction automatique

L'extraction automatique d'information implique la sélection des données à exploiter en temps réel. Dans notre cas d'étude, ceci implique souvent la perte d'informations due au manque d'adéquation à l'usage. Des informations considérées pertinentes dans une situation ne peuvent pas l'être dans un autre contexte, cela demande une extraction adaptative ou l'assistance de l'utilisateur.

ii. Méthode d'analyse

La méthode d'analyse des données est souvent problématique car cela demande une complexité de calcul particulière pour chaque type de données. Si cette méthode n'est pas adéquate, des problèmes de passage à l'échelle ou de confiance dans les résultats peuvent se manifester.

iii. Modèle de représentation

Un modèle de représentation non adéquat implique directement le manque d'adéquation aux besoins de l'utilisateur. Nous trouvons souvent des modèles trop complexes ou saturés qui empêchent l'utilisateur de comprendre l'information. Par exemple, des textes non structurés ou des images sans annotation qui rendent difficile la prise de décision de l'utilisateur dans un moment critique.

iv. Facteur humain

Le facteur humain représente une source d'erreur potentielle vis-à-vis de l'exploitation des données. Le manque d'expertise de l'utilisateur peut générer des problèmes lors de l'interrogation de l'information et les conclusions qu'il extrait. Il peut également, faire de mauvaises manipulations du système amenant à des pertes d'information ou inconsistances dans le système.

v. Interprétation de l'information

L'interprétation de l'information peut être aussi considérée comme un facteur conséquent au facteur humain. Mais celui-ci peut être une conséquence de la mauvaise représentation de l'information. Pour cela, nous l'avons considéré comme un facteur problématique lors de l'exploitation de l'information.

Lors de notre étude menée dans ce paragraphe, nous avons distingué et décrit les possibles facteurs qui peuvent « impacter » la qualité des données issues de capteurs en temps réel. Pour aboutir à ces facteurs, nous avons analysé chaque composant des processus génériques de traitement et distingué des facteurs potentiels impactant la qualité des données capteurs.

En considérant l'étude réalisée, nous présentons dans le paragraphe suivant, une mise en relation entre les facteurs sélectionnés et divers critères qualité utilisés dans les travaux actuels. Pour cette comparaison, nous avons choisi les critères qualité selon

deux axes principaux liés à notre problématique : la qualité des données spatiale et la qualité de service dans le domaine de traitement de l'information. Conjointement à ces aspects, nous considérons également important de comparer les différents critères qualité liés au facteur temps proposés dans la littérature.

Nous décrirons dans ce qui suit leur relation et les critères qui s'adaptent le mieux à l'évaluation des facteurs précédemment définies.

5.1.1.2 Les facteurs et leur impact sur la qualité de données

Les facteurs qualité impliqués dans les processus de traitements génériques présentés précédemment, ont un fort impact sur la qualité des données et la gestion de la qualité des données au sein des systèmes de surveillance, notamment les systèmes de surveillance environnementaux. Dans ce paragraphe nous essayons d'identifier les critères qualité adéquats pour l'évaluation de la qualité des données issues de capteurs. Nous commençons pour définir les sources potentielles d'erreur liées aux facteurs et sur quels critères nous pourrions les lier.

5.1.1.2.1 Sources d'erreurs et critères qualité

Comme nous l'avons décrit dans notre état de l'art, un critère qualité (ou dimension) se réfère à un type d'erreur spécifique de la donnée. Quand la qualité des données est évaluée, elle est évaluée par un ou plusieurs critères et à chaque critère une ou plusieurs mesures sont associées.

Dans notre cas d'application, les erreurs dans la donnée peuvent surgir de différentes étapes au niveau du système de surveillance. Ces erreurs comme le montre la Figure 26, sont directement liées aux processus d'acquisition et de traitement, représentant des défaillances plus spécifiques des facteurs qualité identifiés précédemment (le facteur environnemental, le capteur, la communication, etc.).

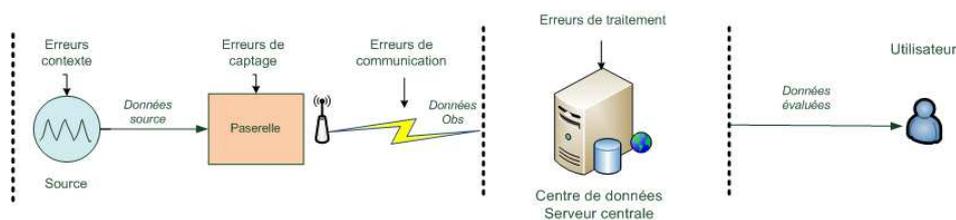


Figure 26 - Modèle conceptuel des erreurs dans un système de surveillance

Dans un cadre général, nous partons de l'hypothèse que toutes les erreurs dans un système de surveillance sont considérées comme *épisodiques*. Nous soutenons cette hypothèse par le fait que dans un contexte dynamique les conditions du système varient au cours du temps et les erreurs peuvent apparaître et disparaître. Cependant, nous remarquons également, que dans ce même contexte, certaines erreurs considérées comme épisodiques peuvent évoluer en *persistantes (constantes)* et affecter la qualité des

données perçue par l'utilisateur final. Donc, nous faisons la distinction dans ce contexte des erreurs du type *épisode ou non-persistantes* et des *erreurs persistantes*.

Nous considérons formellement comme des erreurs *épisode*, celles qui sont dans le contexte d'acquisition, car les facteurs impliqués ne sont pas forcément prévisibles et arrivent de manière aléatoire. Nous pouvons citer concrètement, les conditions environnementales, l'état du capteur et du réseau, le bruit d'interconnexion, l'interférence électromagnétique, les décharges électrostatiques. Les perturbations à court terme causées par l'un de ces effets peuvent conduire à des erreurs persistantes dans les circuits logiques ou de mémoire qui affectera les données traitées ou stockées.

Également, à propos de la communication, les effets comme le bruit, l'interférence, et l'affaiblissement (*fading*) sont des problèmes considérables. Actuellement, les sources d'erreur dominantes dans les réseaux de capteurs sont les erreurs de captage et du canal de communication.

Concernant les erreurs persistantes, nous remarquons que dans la phase de traitement, dans un même esprit que les erreurs dans le processus d'acquisition, des erreurs épisode comme les délais, la saturation du système, le manque de données, etc. sont présentes. Mais, dans ce cadre, des erreurs de type persistant sont plus importantes comme le manque de structure dans les données, le manque de métadonnées et de mise à jour.

Nous considérons suite à cette distinction entre les types d'erreurs, que certaines des erreurs décrites au niveau du système de surveillance sont fortement liées à la qualité de service au sein du réseau et de la communication. Également, d'autres sont plus liées à la qualité de la donnée elle-même (intrinsèque). Dans notre cadre d'étude, la qualité sur les données issues de capteurs implique certaines caractéristiques du type géospatial discutée en termes de la qualité intrinsèque (ou interne) des données, mais dans un environnement dynamique, la qualité interne ne définit pas complètement le système. Des aspects sur l'acquisition, l'accès et le traitement en temps réel des données capteurs impactent aussi l'utilité de l'information et contribuent à la qualité finale que l'utilisateur perçoit. Si bien que la qualité de données concerne les propriétés tangibles des données, la qualité de service concerne les propriétés non tangibles de la qualité associées à l'accès aux données, comme la performance, la sécurité, la fiabilité, etc.

En conséquence nous estimons qu'il est important d'analyser l'impact de ces facteurs et de ces erreurs vis-à-vis des critères dans ces deux axes de la qualité, afin de donner une meilleure spécification des critères et mesures à utiliser pour évaluer la qualité des données issues de capteurs. Également, due à la dynamique de l'environnement, il est important d'apprécier l'impact du temps réel dans la qualité des données.

Dans le paragraphe suivant, afin de définir les critères adéquats pour l'évaluation de la qualité des données issues des capteurs, nous effectuons une analyse qui consiste à croiser les facteurs qualité définis dans notre étude précédente (Section 5.1.1) avec les critères qualité standardisés et utilisés dans la plus grande partie des systèmes d'information exploitant l'information géographique (Section 2.2): la *généalogie*, l'*exactitude*, la *cohérence*, la *précision*, ainsi que les caractéristiques génériques de la qualité de service, notamment sur la *performance*, la *fiabilité*, la *disponibilité* et la *punctualité*.

5.1.1.2.2 Tableaux de relations : les facteurs qualité vs les critères qualité existants

Ce paragraphe met en relation les facteurs qualité distingués auparavant et les critères qualité existants notamment dans deux axes de la qualité : la qualité de données géographiques et la qualité de service, en prenant en considération les critères liés au temps. Cette étude a pour objectif principal évaluer une possible réutilisation ou adaptation des critères qualité existants dans notre domaine d'application. Également, cette étude nous permet d'évaluer le manque en termes de mesures adéquates pour la qualité de données issues de capteur et d'effectuer une proposition des critères qualité potentiels à évaluer.

Plus spécifiquement, nous avons pris en considération les critères qualité concernant : *la qualité dans le domaine géographique, les caractéristiques génériques de la qualité de service (technologies de l'information), et les critères qualité liés au temps.*

Nous avons sélectionné ces critères dans cette mise en relation, car les spécificités des données capteurs montrent le lien avec certaines dimensions, comme la dimension spatiale et temporelle, ainsi que leur comportement dynamique.

Dans ce but, nous utilisons une série de tableaux pour montrer les relations possibles entre ces éléments.

Nous commençons par décrire plus en détails les critères pris en compte lors de cette relation.

a) Critères de qualité des données géographiques

Concernant la qualité des données géographiques, comme nous l'avons décrit dans la Section 2.2, il existe deux types de critères, les critères quantitatifs et les critères qualitatifs. Les critères quantitatifs offrent une expression quantitative de la qualité et sont également appelés « paramètres qualité ». Les critères qualitatifs donnent une expression qualitative de la qualité et requièrent des techniques plus complexes pour leur évaluation.

Les critères que nous analysons, concernant la qualité des données géographiques, sont les critères (ou dimensions) suivants (cf. Section 2.2.1):

- *Généalogie,*
- *Précision géométrique*
- *Précision Sémantique*
- *Exhaustivité ou complétude*
- *Cohérence Logique*
- *Cohérence Sémantique*
- *Précision Temporelle*

b) Critères de qualité de service

La qualité de service est régie de diverses manières par des dimensions et caractéristiques qui permettent d'arriver à l'obtention de la fiabilité quelque soit le secteur d'activité. Concernant la qualité de service, les caractéristiques génériques au sein de notre cas d'application, peuvent être exprimées en termes de :

- *Délais* - le temps utilisé pour transporter les données d'une source à une destination.
- *L'exactitude* - le niveau de conformité avec lequel le transport de données est exécuté.
- *Fiabilité* – degré de confiance de source
- *Disponibilité* – temps de service de la source.

c) Critères de qualité liés au temps

Un aspect important des données est leur variation et la mise à jour, et certains critères sont liés aux aspects temporels en termes de stabilité, de changement à long terme et de la fréquence de variation de la donnée. Dans ce cadre, nous considérons les dimensions suivantes dans notre analyse :

- *Actualité (currency)* - fréquence avec laquelle la donnée est mise à jour.
- *Volatilité* - la fréquence à laquelle chaque donnée varie en temps
- *Ponctualité (timeliness or freshness)*- décrit le degré auquel les données sont utilisables pour une tâche quelconque.

Ensuite, nous présentons un ensemble de tableaux comparatifs entre les facteurs qualité sélectionnés et les critères qualité sélectionnés. Nous structurons ces tableaux en trois sous-sections de façon à représenter les relations entre les facteurs et les critères qualité par type de processus générique (processus d'acquisition, processus de transformation, processus d'exploitation). Nous décrivons plus en détail ces aspects comme suit.

I. Facteurs qualité dans le processus d'acquisition vs critères

Dans le premier tableau (Tableau 10), nous effectuons la mise en relation des facteurs qualité détectés au niveau du processus d'acquisition avec les critères qualité existant dans le domaine géographique.

Composants	Facteurs qualité	Critères qualité						
		Qualité des données géographiques						
		Généalogie	Précision géométrique	Précision sémantique	Exhaustivité	Cohérence logique	Cohérence sémantique	Précision temporelle
Processus d'acquisition	Conditions environnementales et d'opération	x	x			x		
	Attaques malveillantes	x	x			x		
	L'état de l'appareil	x	x		x	x		x
	Ressources limitées	x	x		x	x		x
	Type de mesure	x	x	x		x	x	x
	Méthode de mesure	x	x		x			x
	Type de communication	x	x		x	x		x
	Fréquence de transmission	x			x	x		x

Tableau 10 - Facteurs qualité du processus d'acquisition vs critères qualité au domaine géographique

Bien que les critères qualité du domaine géographique ne prennent pas en compte le contexte dynamique d'acquisition, de traitement et d'exploitation, il semble que certains d'entre eux auraient néanmoins une pertinence dans un contexte de surveillance temps réel. Ce tableau nous révèle qu'il existe une forte relation entre les processus d'acquisition et la généalogie, la précision, notamment géométrique, l'exhaustivité et la précision temporelle. Ceci étant, l'acquisition impacte l'origine des données, la précision des données acquises, ainsi que leur temporalité.

Plus spécifiquement, dans un contexte temps réel, les éléments d'acquisition comme les capteurs ou les réseaux de capteurs correspondent à la *généalogie*. Comme indicateurs de ce critère en temps réel, on peut citer : les ressources du capteur ou du réseau, son état (dégradation), les types de mesures réalisées, les fréquences de ces mesures, etc. Dans une observation donnée, la majorité de ces facteurs évoluent au cours du temps. Nous considérons que la prise en compte de la qualité de service (QoS) dans le critère généalogie semble indispensable.

Également, une relation existe entre le critère *précision géométrique* et la position du capteur. Ce critère correspond aux écarts de valeurs de position et dépend directement des moyens d'acquisition et du traitement de mesures de position. La détermination de la position d'un capteur est fondamentale dans l'utilisation des capteurs, car pour l'exploitation des données issues des capteurs est souvent nécessaire avoir leur information estampillée selon le temps et le espace. Lorsque nous analysons les caractéristiques des capteurs mobiles et agiles par exemple, la variation de la valeur de position est prépondérante, entraînant donc le besoin d'une précision géométrique ou spatiale dynamique.

De manière similaire, nous considérons que dans le contexte d'acquisition les contraintes de la taille mémoire et la grande quantité des mesures acquises au sein des capteurs sont importantes par rapport au niveau de *cohérence*. Usuellement les données issues de capteurs ne sont pas stockées au sein des capteurs pour une longue période, sinon envoyées au centre de données sous forme de flux ou d'ensembles finis de données capteurs. Comme résultat, la cohérence des données au sein des réseaux ne se focalise pas seulement sur la cohérence traditionnelle de lecture et d'écriture. En effet, la cohérence des données capteurs est plus focalisée sur la cohérence spatiale et temporelle de ces données. Par exemple, la cohérence entre plusieurs aperçus de la même donnée à différentes positions et à différents instants de temps.

Les applications utilisant des capteurs ont plus d'intérêt pour l'ensemble des données qui proviennent d'un élément surveillé ou suite à un événement déclenché. Les modèles de cohérence pour les données capteurs ne prennent pas en compte les informations du contexte d'acquisition, comme par exemple, le risque d'une perte d'information générée par des attaques malveillantes ou des facteurs environnementales.

Au niveau du processus d'acquisition, nous considérons que à cause de la dynamique des données et l'importance des temporalités, la cohérence peut être considérée d'un point de vue interne (la donnée elle-même), d'un point de vue temporel, et du point de vue dynamique (changement de la valeur de la donnée). En effet, la cohérence interne nécessite que les données acquises soient exactes, la cohérence

temporelle implique que les données doivent être livrées en temps, et la cohérence dynamique concerne les changements de valeur dans les données, des fois dramatiques. Cette cohérence dynamique peut être liée aux changements temporels et spatiaux de la donnée.

Composants	Facteurs qualité	Critères qualité						
		Qualité de Service				Critères liés au temps		
		Délais	Exactitude	Fiabilité	Disponibilité	Actualité	Volatilité	Ponctualité
Processus d'acquisition	Conditions environnementales et d'opération	x	x	x	x			x
	Attaques malveillantes	x	x	x	x			x
	L'état de l'appareil	x	x	x	x			x
	Ressources limités	x		x	x			x
	Type de mesure	x	x		x	x	x	
	Méthode de mesure	x	x	x	x	x	x	x
	Type de communication	x	x	x	x	x	x	x
	Fréquence de transmission	x	x	x	x	x	x	x

Tableau 11 - Facteurs qualité du processus d'acquisition vs critères QoS et temps

Dans ce deuxième tableau (Tableau 11), nous mettons en évidence la relation entre les facteurs qualité choisis, les caractéristiques proposées au sein de la qualité de service, et les critères qualité liés au temps réel. Comme le montre ce tableau, la relation entre ces facteurs qualité et les critères est beaucoup plus consistante. La plus grande partie des facteurs sont concernés par ces critères.

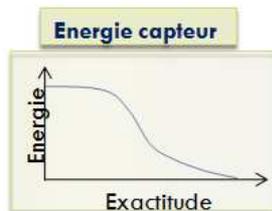


Figure 27 – Énergie du capteur vs Exactitude

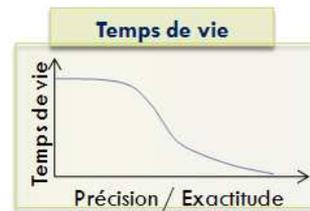


Figure 28 – Temps de vie vs Précision

On peut conclure donc, qu'au niveau du processus d'acquisition, les facteurs qualité sont fortement liés aux délais, à la fiabilité et à la disponibilité.

Dans ce cadre, la *fiabilité* représente un fort appui pour l'évaluation des sources de données. Les capteurs sont par nature typiquement non fiables, et ceci conduit à une dégradation au niveau de la qualité des données à exploiter. Également, le critère *disponibilité* fournit un moyen de qualifier la source des données en considérant leur implication dans le processus d'acquisition.

À propos de *l'exactitude* et *la précision*, nous constatons que l'état du capteur représente un aspect à considérer. Nous mettons en évidence, comme illustré à la Figure 27, que l'énergie du capteur est un facteur qui peut limiter l'exactitude et la précision des mesures effectuées par le capteur. Comme le cite Karl H. [Karl'03], si

moins d'énergie est disponible au sein du capteur, moins d'exactitude et de précision pourrait être offerte. Par ailleurs, nous montrons sur la Figure 28 le lien existant également entre les deux critères exactitude et précision et le temps de vie du capteur. Un capteur avec un temps de vie réduit ne peut pas offrir autant de précision et d'exactitude qu'un capteur débutant son activité. Typiquement, les capteurs avec un temps de vie d'activité assez considérable sont plus sensibles à des fautes dans leur structure (i.e. calibration, réactivité), et à l'impact de l'environnement de déploiement.

II. Facteurs qualité dans le processus de transformation vs critères qualité existants

Dans cette sous-section, nous présentons la mise en relation des facteurs qualité sélectionnés lors du processus de transformation et les critères qualité du domaine géographique.

Nous remarquons suite à cette représentation (Tableau 12) que les facteurs qualité proposés ont une relation plus consistante avec la plupart des critères dans le domaine géographique, notamment avec la validation, la nature des données, et les métadonnées. Ceci est dû à la nature et à la quantité des données à traiter en temps réel. Nous distinguons également que la précision temporelle représente un dénominateur commun dans tous les facteurs, car les changements du monde réel au cours du temps peuvent générer des mauvaises appréciations de la réalité s'ils ne sont pas correctement gérées.

Composants	Facteurs qualité	Critères qualité						
		Qualité des données géographiques						
		Généalogie	Précision géométrique	Précision sémantique	Exhaustivité	Cohérence logique	Cohérence sémantique	Précision temporelle
Processus de transformation	Prétraitement		x	x	x	x		x
	Validation		x	x	x	x	x	x
	Nature de données (hétérogén)	x		x		x	x	x
	Volume de données				x	x		x
	Ressources	x			x			x
	Temps de transformation				x			x
	Format des données			x	x	x	x	
	Type de stockage	x			x			x
	Métadonnées	x	x	x	x	x	x	x
	Mise à jour		x				x	x

Tableau 12 - Facteurs qualité du processus de traitement vs critères qualité au domaine géographique

Lors de la comparaison avec les critères de la qualité de service et ceux liées au temps (Tableau 13), nous remarquons que concernant la qualité de service, les critères plus adéquats concernent les délais dans le traitement des données et la difficulté d'accès et disposition d'informations que cela implique.

Composants	Facteurs qualité	Critères qualité						
		Qualité de Service				Critères liés au temps		
		Délais	Exactitude	Fiabilité	Disponibilité	Actualité	Volatilité	Ponctualité
Processus de transformation	Prétraitement	x	x		x	x	x	x
	Validation	x	x		x	x	x	x
	Nature de données (hétérogén)	x	x			x	x	
	Volume de données	x	x		x	x	x	x
	Ressources	x		x	x	x	x	x
	Temps de transformation	x			x	x	x	x
	Format des données					x	x	x
	Type de stockage	x	x		x	x	x	x
	Métadonnées			x	x	x	x	x
	Mise à jour	x	x		x	x	x	x

Tableau 13 - Facteurs qualité du processus de transformation vs critères QoS et temps

Nous remarquons également que la disponibilité et la capacité des ressources dédiées au traitement et stockage des données peuvent être impliquées dans la dégradation de la qualité, notamment sur la disponibilité des informations à fournir à l'utilisateur. Les facteurs qualité plus orientés sur la donnée elle-même montrent plus de relations avec les critères dédiés à l'exactitude et la fiabilité des sources.

Dans ce cadre également, les facteurs dans le processus de transformation marquent leur corrélation avec les critères agissant sur les mesures du temps de traitement, la mise à jour des données et la ponctualité. La nature du traitement dans un contexte de surveillance pose des contraintes temporelles qui sont plus délicates en temps réel, donc, les critères *actualité*, *volatilité* et *ponctualité* sont pertinents.

III. Facteurs qualité dans le processus d'exploitation vs critères qualité existants

Concernant les facteurs sélectionnés dans le processus d'exploitation et les critères qualité du domaine géographique, nous remarquons une correspondance très faible. En effet, les facteurs dans ce processus, sont plus focalisés sur les besoins de l'utilisateur et de l'application et plus difficiles à mettre en relation avec des critères qui caractérisent plus la qualité interne des données.

Malgré cette difficulté, nous remarquons que les méthodes d'extraction et de représentation peuvent à un moment donné impacter l'exploitation des données en termes de cohérence et précision (Tableau 14).

Une mauvaise extraction d'information, peut générer une information erronée et en conséquence des problèmes de cohérence sémantique ou logique qui peuvent impacter la qualité des données perçue par l'utilisateur.

		Critères qualité						
		Qualité des données géographiques						
Composants	Facteurs qualité	Généalogie	Précision géométrique	Précision sémantique	Exhaustivité	Cohérence logique	Cohérence sémantique	Précision temporelle
Processus d'exploitation	Extraction automatique		x	x	x	x		x
	Méthode d'analyse		x	x	x	x	x	x
	Modèle de représentation		x	x	x	x	x	
	Facteur humain					x	x	x
	Intérpretation de l'information					x	x	x

Tableau 14 – Facteurs qualité du processus d'exploitation vs critères qualité au domaine géographique

Dans cette relation le facteur humain est un aspect particulier car l'utilisateur intervient plutôt au moment de la prise de décision et non dans le traitement des données. En revanche, d'un point de vue plus subjectif, nous pourrions dire qu'il peut mal interpréter les informations qui lui sont communiquées et ainsi considérer certaines données caractérisées par une mauvaise qualité.

		Critères qualité						
		Qualité de Service				Critères liés au temps		
Composants	Facteurs qualité	Délais	Exactitude	Fiabilité	Disponibilité	Actualité	Volatilité	Ponctualité
Processus d'exploitation	Extraction automatique	x	x	x	x	x	x	x
	Méthode d'analyse	x	x	x				
	Modèle de représentation		x	x	x			
	Facteur humain	x	x	x	x	x		x
	Intérpretation de l'information		x	x				x

Tableau 15 – Facteurs qualité du processus d'exploitation vs critères qualité de service et liées au temps

Finalement, nous présentons (Tableau 15) la relation entre le processus d'exploitation et les divers critères liés à la qualité de service et à la temporalité. Nous remarquons dans ce tableau, que la relation entre ces aspects est plus concrète car les critères s'adaptent mieux aux caractéristiques décrites par les facteurs, notamment concernant l'extraction et la représentation de l'information, ainsi que l'intervention de l'utilisateur. Nous considérons qu'avec ce type de critères les facteurs qualité sélectionnés seront mesurables en termes d'accès, ou de disponibilité de l'information à leur usage.

Lors de cette étude, nous constatons que les trois axes des critères qualité existants (cf. Section 5.1.1.2), fournissent une importante source de sélection de critères de la qualité selon différentes perspectives (qualité interne dans le domaine géographique, qualité de service, et l'impact du temps réel). Nous considérons que ces

critères sont le résultat d'un ensemble de critères dits génériques (ou de base) souvent cités dans les travaux existants (cf. Section 2.2.1) et qui sont ensuite appropriés à chaque domaine ou type d'application. C'est aspect est plus évident lorsque nous essayons de comparer avec les mêmes spécificités, les facteurs qualité sur des données non géographiques. Toutefois, nous considérons que parmi ces critères, certains montrent une perspective intéressante et applicable aux données issues de capteurs. En conséquence, une spécification des critères adéquats pour l'évaluation des données spatio-temporelles issues de capteurs est nécessaire. Nous détaillons cet aspect dans la section suivante.

5.1.2 Spécification des critères d'évaluation de la qualité de données capteurs en temps réel

Suite à l'analyse des facteurs et probables critères qualité à appliquer dans l'évaluation des données issues de capteurs, ce paragraphe est focalisé sur la définition des critères qualité nécessaires pour être associés à l'exploitation en temps réel des données issues de capteurs (Figure 29).

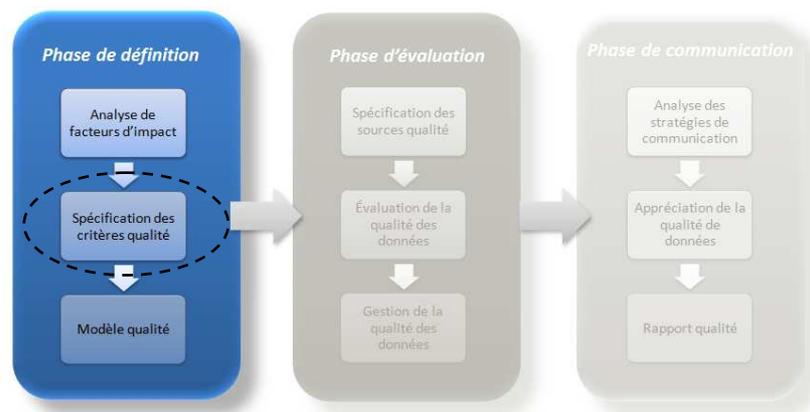


Figure 29 – Phase 1 – Spécification des critères qualité

Pour cette définition des critères qualité, nous nous appuyons sur notre analyse des facteurs d'impact et la mise en relation avec les critères qualité existants. Nous concluons qu'en effet, dans le contexte de la surveillance, le processus de traitement de la donnée commence lors de son acquisition au sein du capteur. Ces mesures sont souvent exposées à des erreurs lors de la prise de mesures, comme la calibration de l'appareil, la méthode de mesure, etc. Ensuite, ces données sont transmises au centre de données ou à un nœud capteur, donc les données sont encore exposées à des erreurs issues du transfert. Évidemment, les effets issus de ces erreurs seront perceptibles lors de la gestion d'informations au centre des données. De fausses données, de données manquantes, etc., sont probables.

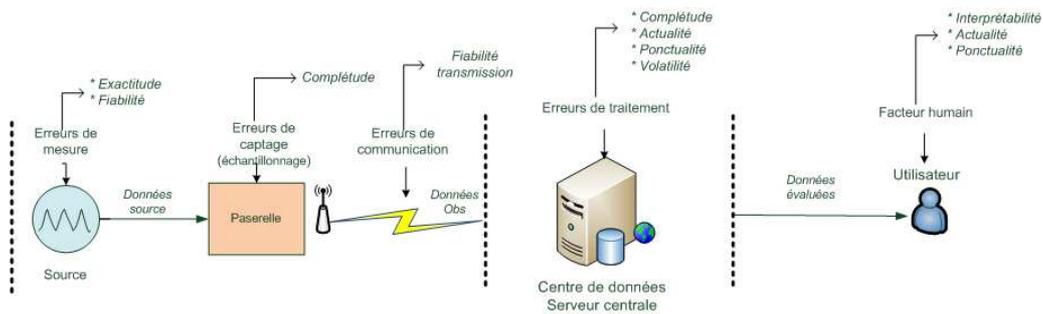


Figure 30 – Erreurs et critères qualité au sein du système de surveillance

Suite à cette conclusion, dans la Figure 30 nous montrons les différentes sources d'erreurs distinguées lors de notre étude, et la mise en relation avec les critères de qualité sélectionnés pour l'évaluation de la qualité.

Nous essayons de fournir une sélection assez générique des critères, pour permettre l'adéquation de n'importe quel type de système de surveillance à cette analyse. Nous nous basons donc sur les critères proposés lors des travaux existants au sein des systèmes d'information et la qualité de service (fiabilité, exactitude, précision, et ponctualité). Les critères considérés dans notre approche, sont des critères qui ont été catégorisés de différentes façons et avec différentes définitions selon leur domaine d'application (exactitude, complétude, ponctualité...). Par exemple, en *qualité interne* et *qualité externe*, dans le domaine géographique [Devillers'02], ou en *qualité intrinsèque*, *qualité d'accessibilité*, *qualité contextuelle* et *qualité de représentation* [Wang'01].

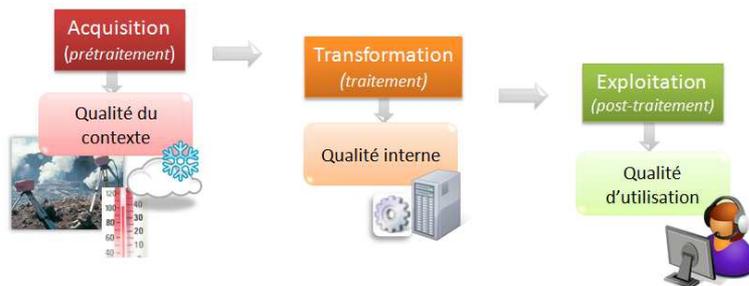


Figure 31 - Catégories de la qualité des données issues de capteurs

Dans cet esprit, nous proposons également des catégories caractérisant la qualité des données issues de capteurs au sein des systèmes de surveillance : *la qualité du contexte*, *qualité interne* et *qualité d'utilisation* (Figure 31).

Premièrement, la catégorie de *la qualité du contexte* implique les critères qualité liés à la source de données. Ces critères nous permettent évaluer la qualité de la source et les données brutes produites. La catégorie de *la qualité interne*, concerne les critères qualité liés à la donnée et à la validité des données, afin de prendre en considération les caractéristiques statiques et dynamiques des critères qualité. Finalement, *la catégorie de la qualité d'utilisation* est composée par les critères concernant la qualité lors d'exploitation, en termes d'interprétabilité, d'actualité et de disponibilité. La Figure 32

montre la structuration de l'ensemble de ces critères qualité proposés dans notre approche.

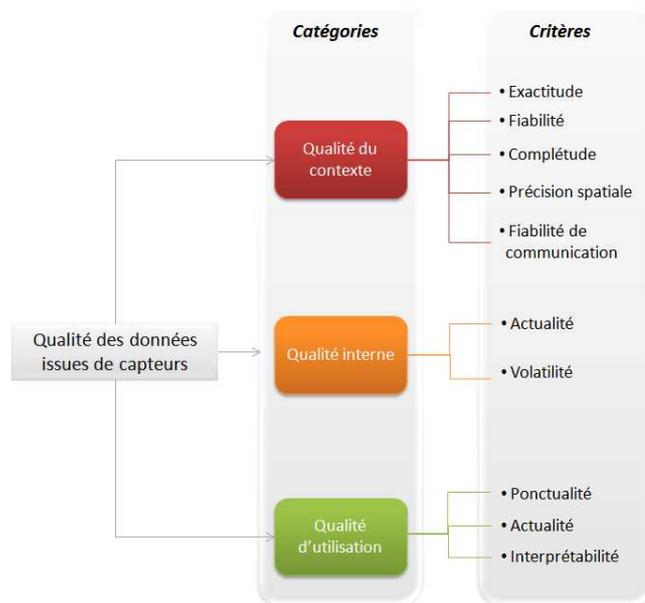


Figure 32 – Structuration des critères qualité des données capteurs

Dans ce qui suit, nous définissons l'ensemble de critères qualité sélectionnés pour chaque catégorie et comment ils sont considérés dans notre contexte d'application. Ces critères sont inspirés des critères existants et adaptés à notre contexte d'application.

5.1.2.1 Définition des critères qualité des données issues de capteurs

Dans ce paragraphe, nous définissons l'ensemble de critères pris en compte dans notre approche pour l'évaluation de la qualité des données issues de capteur. Dans notre contexte, ces critères sont décrits comme suit :

i Critères qualité du contexte

La catégorie sur la qualité du contexte inclut les dimensions suivantes :

- *L'exactitude* décrit la précision numérique d'une donnée, représentée comme l'erreur absolue d'une valeur physique. Dans notre cas d'application, l'exactitude d'un capteur est comparée avec la précision fournie dans la spécification technique du capteur. En effet, nous considérons qu'à un instant donné une mesure peut être très précise mais pas exacte [Servigné'06]. Dans ce document, quand nous exprimons l'exactitude d'une donnée, nous faisons référence à la mesure selon laquelle la donnée est correcte par rapport une valeur de référence en considérant sa précision.

- La *fiabilité* d'une donnée issue de capteurs est perçue comme les erreurs de mesures statistiques dues aux interférences aléatoires ou sporadiques (i.e. vibrations, chocs, attaques, etc.) Avec l'exactitude, la fiabilité est aussi perçue comme une erreur maximale absolue.
- La *complétude* s'adresse au problème du manque de valeurs à cause des problèmes au sein des capteurs et leurs défaillances d'opération. Dans ce critère nous prenons en compte combien des valeurs mesurées sont disponibles pour le traitement au sein du centre de données.
- La *précision spatiale* dans notre contexte d'application s'adresse à la mesure avec laquelle les coordonnées de la position d'un capteur sont proches de leur emplacement réel.
- *Fiabilité de la communication*. Mesure avec laquelle les données transmises sont réceptionnées au centre de données. Cette mesure implique la puissance de transmission du capteur, ainsi que la puissance du récepteur en prenant en compte les probabilités d'une perte d'information à cause du signal de bruit.

ii Critères qualité interne

Critères qualité concernant la catégorie sur la qualité interne :

- *Actualité*. Distance entre l'instant où une valeur du monde réel est modifiée et l'instant où la valeur elle-même est modifiée dans le système. Ce critère se réfère seulement aux données qui changent au cours du temps.
- *Volatilité*. Mesure dans laquelle les mesures changent. Ce critère implique la rapidité avec laquelle les données changent au cours du temps.

iii Critères qualité d'utilisation

Critères qualité concernant la catégorie sur la qualité d'utilisation :

- *Ponctualité*. Mesure dans laquelle la validité des données est adéquate pour leur découverte.
- *Actualité*. Mesure dans laquelle les données sont à jour pour une tâche quelconque.
- *Interprétabilité*. Mesure dans laquelle les données sont décrites dans une langue ou une représentation appropriée pour la compréhension de l'utilisateur.

Après cet aperçu sur les critères qualité liées à l'évaluation en temps réel des données issues de capteurs, nous remarquons que certains de ces critères se focalisent sur la qualité de la mesure et d'autres sur la temporalité de cette mesure. Nous considérons donc que la qualité des données issues de capteurs est définie notamment, par l'exactitude et la fiabilité au sein du capteur et par la temporalité des données exploitées.

5.1.2.2 Définition des mesures pour l'évaluation des critères qualité

Pour permettre l'évaluation de la qualité des données issues de capteurs nous nous basons sur l'ensemble de critères qualité proposées précédemment. Nous avons sélectionné *l'exactitude*, qui décrit l'erreur maximale d'une mesure issue de capteur, *la fiabilité* qui représente l'erreur de mesure statistique en raison des interférences environnementales aléatoires, *la complétude* qui caractérise la manque des données dans un ensemble, *la précision spatiale* qui représente le niveau d'exactitude entre la position informé par le capteur et leur position réelle, *la fiabilité de la communication* qui décrit la probabilité que la moyenne communication utilisé pour livrer les données, *et l'actualité, la ponctualité et la volatilité* pour évaluer le contexte temporelle des données issues de capteurs.

Dans la suite, nous nous attachons seulement à la description de ces critères fortement liée à la dynamique de l'environnement d'acquisition et de traitement au sein du système de surveillance. Les critères qualité d'utilisation restent toujours aux titres informatifs, et optionnels pour leur évaluation.

❖ L'exactitude

Dans notre approche, l'exactitude des données issues d'un capteur décrit la mesure d'erreur systématique résultante des erreurs dans le processus de mesure, comme le non-calibrage du capteur, l'influence environnementale sur les valeurs obtenues, les problèmes dans la méthode de mesure utilisée, etc. Nous considérons l'exactitude comme une valeur de mesure qui définit l'erreur maximale et absolue « e_x », telle que la valeur réel v se trouve dans l'intervalle entre la valeur mesuré v sans l'erreur ou avec l'erreur. Pour le calcul initial de cette valeur d'erreur, nous prenons en considération la valeur de précision fournie par le fabricant et qui est contenue dans les métadonnées.

Donc, cette exactitude à un instant t_j est calculée par une fonction $e_x(j) = f(e_x(j))$. Dans notre approche, nous considérons que chaque application en particulier peut appliquer une fonction adéquate à ses besoins en termes de transformation de l'erreur détectée. On ne s'attache pas à la déclaration des fonctions idéales pour un contexte donné.

❖ Fiabilité

La fiabilité des mesures issues d'une source donnée est impactée par les erreurs aléatoires présentes lors de l'acquisition, comme les interférences environnementales, les attaques, les vibrations, changements de position soudains, entre autres. Due à son type aléatoire et pas prévisible, cette erreur s'éparpille autour d'une valeur donnée. La fiabilité définit donc les bordures de cette distribution aléatoire et est décrite par l'intervalle de confiance [Haas'96].

De la même façon que nous considérons l'estimation de l'exactitude, la fiabilité est aussi donnée comme l'erreur maximale absolue. Cela dit, la fiabilité est une valeur

issue d'erreur aléatoire ϕ qui est définie dans l'intervalle de la valeur mesurée m avec cette erreur et sans cette erreur en gardant une probabilité de fiabilité p .

En adaptant [Haas'96], nous considérons la fiabilité à un instant t comme :

$$\phi(f) = \frac{d \cdot \sigma(f)}{\sqrt{r_s}} \cdot \sqrt{1 - r_s} \quad (9)$$

Avec d qui représente la probabilité de fiabilité comme étant $(1-p/2)$, r_s qui représente la fréquence d'échantillonnage et σ qui fournit la variation de la donnée.

❖ Précision spatiale ou de position

La détermination de la précision de la position d'un ou plusieurs capteurs est fondamentale pour l'exploitation des données issues des capteurs. Par exemple, afin de suivre ou détecter des objets avec un réseau de capteurs, les positions des capteurs doivent être préalablement connues pour détecter les positions des objets ou être estimées à partir des positions connues. Dans notre contexte, nous considérons ce critère comme le niveau de conformité entre la position estimée du capteur par un système de positionnement et d'autres informations acceptées comme précises ou exactes.

Actuellement, le DGPS¹ est un système qui a été souvent utilisé dans le domaine géospatial, car il offre une meilleure exactitude en termes d'estimation de position par rapport à un système GPS classique. Ce système peut déterminer la position d'un objet avec une précision en centimètres, voire en millimètres par rapport à une station de référence fixe. Cependant, ce niveau d'exactitude ou de précision ne peut pas être atteint pour des objets mobiles, surtout à cause de la dynamique et des conditions environnementales. Dans ce cas, la distance entre l'objet et la station de référence peut varier et engendrer des erreurs (i.e. manque de visibilité aux satellites, problèmes de communication...). En général, le DGPS comprend une station de contrôle et une station de référence, dont le calcul des corrections différentielles pour une position exacte à une période ou un instant donné sont effectuées, ainsi que des moniteurs d'intégrité qui leur sont associés pour vérifier l'état de la communication de l'information position aux utilisateurs. Les corrections effectuées peuvent souffrir d'une dégradation en fonction de la distance de séparation entre l'objet et la station de référence ou de certains facteurs environnementales impactent. Avec ces conditions d'affaiblissement dans la précision de positionnement, le DGPS peut fournir une précision autour d'un mètre, ce qui est pertinent dans notre cas.

Nous nous basons dans notre approche sur l'utilisation du DGPS comme système de positionnement du capteur. Dans notre cas, nous supposons que les capteurs utilisés lors d'une observation sont notamment du type statique ou agile. Ainsi, nous connaissons la position initiale ou de référence du capteur ainsi que la position de la station de référence. Afin de vérifier leur précision, nous prenons la valeur de position

¹ Differential Global Positioning System

de référence du capteur enregistrée (position initiale ou connue), ainsi que la valeur d'exactitude considérée par le système de positionnement selon les conditions environnementales actuelles. Ensuite, à chaque nouvelle mesure estampillée en temps et en espace et à chaque nouveau changement de position (i.e. capteurs agiles) la position est corrigée de façon différentielle et dynamique par le système de positionnement. Dans ce cadre, nous faisons un rapport de précision entre la valeur de position estimée et la valeur de position enregistrée comme référence.

On peut dire donc, qu'une erreur dans la position du capteur est présente quand la position initiale du capteur p_0 n'est pas la même que celle estimée p_i , $p_0 \neq p_i$, en respectant l'exactitude de la position estimée par le système de positionnement selon leurs spécifications initiales.

Rappelons que des algorithmes d'amélioration de l'exactitude de positionnement peuvent être utilisés pour aider à une meilleure approximation de la valeur position estimée, comme dans les réseaux de géocapteurs (Section 3.2.3.1).

❖ Complétude

La complétude est un des critères les plus importants concernant l'information qualité. Cette mesure s'adresse aux problèmes de manque des valeurs causé par les défaillances ou le mauvais fonctionnement du capteur. Entre plus complète soient les données, plus es la qualité des conclusions sur un ensemble de données [Biswas'06].

Il existe actuellement un bon nombre de stratégies pour gérer les valeurs manquantes dont l'interpolation est la plus utilisée. Cependant, avec la complétude nous essayons de distinguer la quantité de données qui sont arrivées au point de traitement en prenant en compte également leur temporalité.

Comme le montre [Biswas'06], tout au long du traitement des données capteurs, diverses fréquences au sein du système et du capteur interviennent :

- *La fréquence des données du système* qui représente le taux maximal des données qui prévaut pour toutes les sources et des utilisateurs finaux dans un système.
- *La fréquence de données de mesure* est la fréquence à laquelle le capteur effectue des échantillons dans son environnement.
- *La fréquence de données capteur* est la fréquence à laquelle le capteur communique les données au centre de données ou à un autre nœud.
- *La fréquence de données opérateur* est la fréquence à laquelle un opérateur logique dans le réseau de capteurs produit des données.
- *La fréquence de données requête*, qui représente la fréquence à laquelle l'utilisateur final souhaite que les données soient livrées dans une réponse à une requête.

Dans notre contexte d'application nous prenons en considération les fréquences relatives aux données capteurs, les mesures de capteurs et la fréquence des données système. Ceci se traduit par :

$$\mathbf{CompSys}(s_i) = \frac{r_o(s_i)}{r_{sys}} \quad (10)$$

Avec la fréquence des données $r_o(s_i)$ du capteur s_i , et la complétude de sortie $CompSys$.

Également, au même niveau que les fréquences de données, nous détectons aussi divers types de temps auxquels les données sont traitées.

- t_{max} : est le temps maximum dans lequel une agrégation d'objets sera complète
- t_{input} : instant du temps dans lequel la donnée entre au système

Dans ce point de vue, nous essayons également, de distinguer si en termes de temporalité une donnée est ou non prise en compte dans l'intervalle du temps. Dans ce cas, nous considérons que chaque valeur obtenue respecte une temporalité existante dans l'agrégation des données, la complétude dans ce contexte est définie par les valeurs suivantes :

En considérant une fonction $C(t)$ qui estime la valeur de complétude à l'instant t , dont $t \in [t_{input}, t_{max}]$, avec cette fonction nous pouvons calculer la complétude des données à l'entrée du système. On définit la complétude comme :

$$C = \int_{t_{act}}^{t_{max}} C(t) \quad (11)$$

dont t_{act} est l'instant de temps dont la complétude est évalué et $t_{act} < t_{max}$.

Également, nous considérons que cette complétude peut être évaluée soit sur un ensemble de données dans l'intervalle du temps $[t_{input}, t_{max}]$ ou dans une fenêtre de temps d'une taille x , définie par le $[t_{start}, t_{end}]$. Dans ce cadre, la probabilité de la complétude est représentée par :

$$C(t) = \frac{\text{sum}(v^i(j) | t_{start} \leq j \leq t_{end})}{x} \quad (12)$$

❖ Fiabilité de la communication

Pour calculer la fiabilité de la communication, nous reprenons l'étude réalisée par divers travaux de recherche au sein des réseaux de capteurs [Hoes'09]. Dans cette approche en particulier, la fiabilité de la communication est étudiée du point de vue de la qualité de service. Quand un paquet de données est transmis, un autre nœud ou le serveur ont la possibilité de recevoir ce paquet en dépendant de l'indice SNR (rapport signal bruit) et la taille du paquet envoyé. Selon le modèle, un affaiblissement peut être présent α , et la puissance du signal de réception P_{rx} dépend du signal de transmission P_{tx} et de la distance à parcourir d .

Selon un facteur de gain constant $k_r P_{tx} \left(\frac{d_0}{d}\right)^\alpha$ le coefficient de perte doit être $\alpha \geq 2$ et référencie à la distance d_0 . Pour simplifier le calcul, nous assumons qu'il existe

une probabilité de perte prédéfinie L . Donc, la fiabilité de la communication $Fiabilité_{com}$ représente la probabilité que les données soient correctement envoyées.

$$Fiabilité_{com} = P_{tx}(L) \quad (13)$$

❖ Actualité

Avec une dynamicité assez forte, comme dans les flux de mesures, il n'existe pas a priori une formule spécifique pour déterminer l'actualité des données. Cependant, divers travaux optent pour prendre en considération la temporalité des mesures pour l'estimer conjointement en évaluant sa pertinence en valeur (i.e. exactitude). Dans ce cas, l'actualité peut être perçue de deux points de vue : l'âge d'une donnée en considérant le temps de son enregistrement et le temps qui correspond au système. Nous pouvons percevoir l'actualité, comme la ponctualité de la donnée, qui est plutôt concernée par la perception que l'utilisateur a sur les données. Nous aborderons la ponctualité dans les critères utilisés pour évaluer la qualité d'utilisation.

Dans ce cadre, nous analysons l'actualité d'une donnée m à un instant t comme la différence entre le temps de l'introduction de la donnée au système t_{input} et le temps de sa livraison t_{liv} .

$$Actualité = t_{val} + (t_{liv} - t_{input}) \quad (14)$$

Nous ajoutons à cette différence l'intervalle de temps t_{val} pendant lequel la donnée reste valide avant d'être stockée. Cet intervalle du temps est fixé au départ.

❖ Volatilité

La volatilité concerne la variation de la donnée au cours du temps, plus spécifiquement, la durée de la validité de la donnée. Pour le calcul de la volatilité au sein du centre de données, nous considérons les valeurs suivantes :

- t_{input} : instant du temps d'entrée de la donnée au système
- t_{val} : temps de validité

avec $t \in [t_{input}, t_{val}]$

Donc la volatilité est défini comme :

$$V = (t_{val} - t_{act}) - \int_{t_{act}}^{t_{val}} P_{val}(t) \quad (15)$$

dont t_{act} est le moment de l'évaluation de la volatilité, $t_{act} < t_{val}$ et $P_{val}(t)$ est la probabilité qu'à l'instant t , le t_{val} soit le temps réel d'expiration.

Pour des données qui sont stables dans le temps, nous assumons que : $\forall t \in [t_{\text{input}}, +\infty], Pval(t) = 1 \text{ et } V=0.$

❖ Ponctualité

La ponctualité représente la mesure de l'adéquation des données et leur disponibilité pour une tâche, notamment leur exploitation par l'utilisateur. Elle est définie par :

$$Ponctualité = \left\{ \max \left[\left(1 - \frac{Actualité}{volatilité} \right), 0 \right] \right\}^{\beta} \quad (16)$$

Après avoir défini les critères qualité pertinents pour l'évaluation des données issues de capteurs et leurs mesures, nous pouvons considérer que la qualité de ces données *QDC (Qualité de Données Capteurs)* est caractérisée par les critères de qualité : *exactitude (e_s), la fiabilité (ϕ), la complétude \mathcal{C} , l'actualité, la volatilité et la ponctualité.*

Suite à la sélection des critères spécifiques à mesurer et notre perception de leur application, dans nous présentons, dans le paragraphe suivant, notre proposition de la caractérisation de la qualité de données capteurs sous la forme d'un modèle de qualité. Nous précisons également les différentes composantes qui le décrivent.

5.1.3 Définition du modèle de qualité pour les données issues de capteurs

Nous allons proposer dans cette section un modèle adapté pour la structuration et l'évaluation de la qualité des données capteurs (Figure 33).

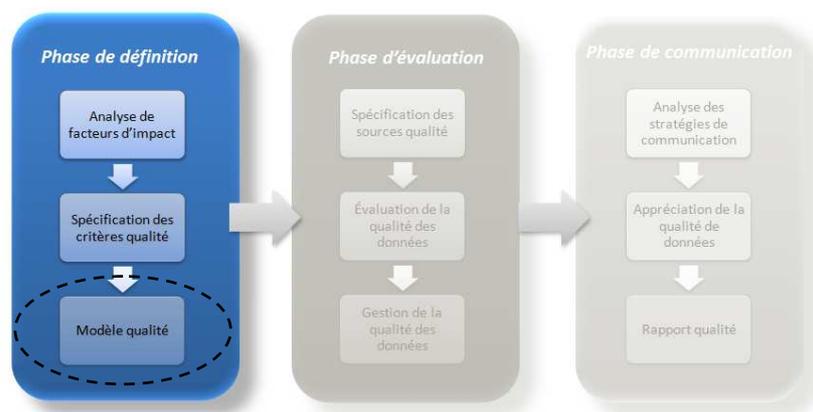


Figure 33 – Phase 1 – Définition du modèle de qualité

Nous nous sommes inspirée pour l'élaboration de ce modèle de modèles de qualité existants et pertinents dans notre contexte [Berti-Equille'04], [Devillers'04], [ISO'02].

Actuellement, il n'existe pas de modèles ou de structures génériques qui permettent la description et la définition de la qualité des données indépendamment du domaine d'application. Si un modèle générique semble à l'heure actuelle difficile à concevoir, nous proposons un modèle de qualité offrant une meilleure généricité que les modèles existants grâce à une abstraction des critères et indicateurs. Comme nous l'avons cité dans la Section 2.2.1, la qualité définie par les modèles existants se caractérise par différents éléments appelées : dimensions, facteurs, ou bien critères [Berti-Equille'04], [Devillers'04]. Chaque modèle possède sa propre vision du niveau de description de ces éléments en s'appuyant sur une diversité de techniques et méthodes pour leur évaluation.

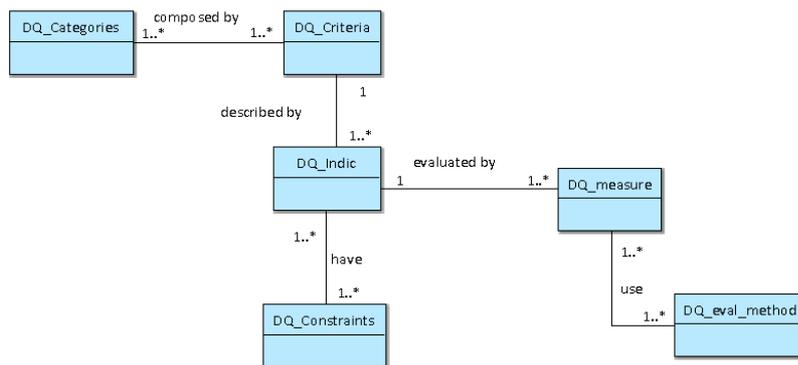


Figure 34 – Modèle de qualité de données capteurs

Ainsi, afin de définir notre méthode d'évaluation de la qualité de données capteurs, nous proposons le modèle de qualité représenté dans la Figure 34.

Suite à l'analyse des facteurs d'impact, la qualité est définie à l'aide de catégories et de critères. Chaque critère peut être associé à un ensemble d'indicateurs, et à un indicateur donné peuvent correspondre différentes mesures. La définition des critères et leur composition pour leur évaluation a été définie dans la Section 5.1.2. Nous définissons chacun de ces éléments comme suit :

- *Catégorie qualité* : représente la catégorisation de la qualité de données au sein d'un système de surveillance (qualité du contexte, qualité interne, qualité d'utilisation). (cf. Section 5.1.2).
- *Critère qualité* : représente une caractéristique de la qualité à évaluer ; il est considéré comme une extension de la donnée. Un critère peut être qualitatif ou quantitatif. (cf. Section 5.1.2).
- *Indicateur qualité* : l'indicateur est une valeur issue d'une mesure ou d'un ensemble de mesures qualité.
- *Mesure qualité* : représente la méthode d'évaluation appliquée sur les données qui contribue à valuer l'indicateur qualité.

Dans les paragraphes suivants, nous nous attachons à la mise en correspondance du modèle des données capteurs défini lors de notre étude initiale (Section 4.2.1.1) avec

notre modèle de qualité (Figure 34). L'objectif de cette mise en correspondance est de représenter l'intégration de la caractérisation de la qualité vis-à-vis des spécificités des données capteurs. Une formalisation des objets issus de cette relation est aussi détaillée.

5.1.3.1 Modèle de qualité et les données issues de capteurs

Suite à la description du modèle de qualité proposé, nous présentons dans ce paragraphe l'interaction de ce modèle avec les spécificités des données capteurs.

Comme nous l'avons précisé dans le chapitre 4, nous supposons que l'exploitation des données issues de capteurs, se réfère au traitement de campagnes de mesures générant des flux de mesures ou des mesures à basses fréquences d'acquisition. Selon notre modèle de qualité, ces données exploitées sont liées à un ou plusieurs critères qualité qui sont représentés par des indicateurs et évalués grâce aux mesures qualités (Figure 35).

Dans la Figure 35, nous montrons que l'information qualité est liée à des mesures et des flux de mesures issus de campagnes de mesures lancées lors de l'observation d'un phénomène. Dans ce contexte, un ou plusieurs critères qualité peuvent être attachés à une mesure en particulier, à un ensemble de mesures issues d'une campagne ou à des mesures acquises de façon continue par un flux d'un capteur.

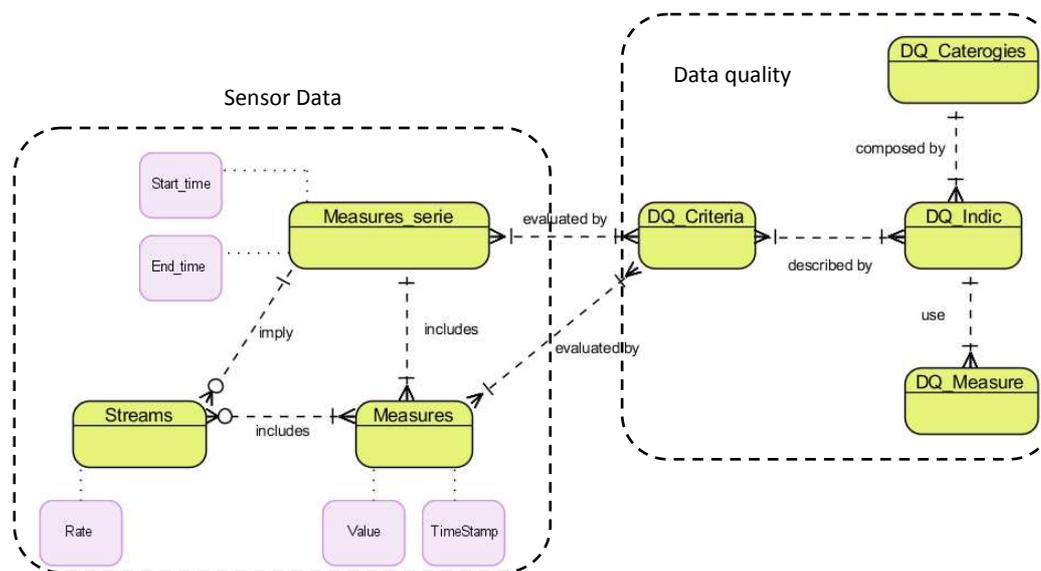


Figure 35 – Interaction du modèle de données capteurs et modèle de qualité

Dans le paragraphe suivant nous montrons comment nous pouvons associer les valeurs représentatives de la qualité aux données capteurs analysées. Une formalisation de ces éléments est également proposée.

5.1.3.2 Formalisation et gestion des valeurs qualité

Pour caractériser les valeurs qualité dans notre approche, il est nécessaire de tenir compte des mesures acquises de façon ponctuelle, en continue ou dans un intervalle de temps. Afin d'associer l'information qualité à ce type de données, il faut y prendre en considération la granularité de telles données. Pour clarifier cet aspect, nous assumons qu'une mesure peut être associée à l'information qualité résultante de l'évaluation des critères chaque fois que la donnée arrive, ou dans un intervalle de temps spécifié ou calculé dans le cas des campagnes et flux de mesures.

Dans ce cadre, nous partons de notre formalisation des objets issus de capteur proposée dans la Section 4.2.2.2, pour spécifier la relation entre les valeurs qualité et les données capteurs. Dans cette formalisation, nous considérons l'instant de temps indépendant de la position du capteur. Donc, les objets sont formalisés selon l'évaluation de la qualité des données acquises et non en regardant la mobilité du capteur.

Les objets résultants continuent à se distinguer par un identifiant Id , qui est unique et constant dans le temps, issue d'une mesure m , d'un ensemble de mesures β ou d'un flux de mesures φ . La qualité est modélisée par des critères q spécifiques pour un type particulier de données j . Ici, les valeurs des propriétés d'une donnée, dites attributs A , possèdent des valeurs des critères qualité vm associées aux valeurs de la donnée à un instant donné t . Les instants t sont ordonnées avec t_i pour m_j , $t_i \mid t_i < t_{i+1}, i > 0$ pour β et $t_i \mid t_i < \infty, i > 0$ pour φ . Ainsi, chaque donnée et valeur qualité sont associées à des critères qualité. Ayant définie ces notations, les objets complétés avec l'information qualité peuvent être formalisés.

a) Objets issus de mesures et de campagnes de mesures

Premièrement, nous formalisons les objets issus de mesures ponctuelles à une basse fréquence d'acquisition. Ces mesures ne font pas partie d'un flux, mais sont comprises dans des campagnes déclenchées dans l'observation d'un phénomène.

Définition 4. Considérons $M_j\{m_1 \dots m_n\}$ comme un ensemble de mesures issues de capteurs et une partie des attributs variables dans le temps A . Ces mesures sont décrites par leur valeur v^j , mais aussi par leur information qualité q à un instant de temps t_i donné. La qualité associée est représentée par :

$$m_t^j = (v_t^j, q_t^j) \tag{7}$$

avec $q_t^j \in \{\text{fiabilité, précision, actualité...}\}$

Selon ce formalisme, chaque valeur de mesure v_t^j est associée à un ensemble de critères qualité. Dans le cadre d'un ensemble de mesures comprises dans une campagne, l'instant t_i sera pris en considération comme un période compris dans $t_i \mid t_i < t_{i+1}, i > 0$.

Définition 5. Les critères qualité q sont définies par :

$$q = (\text{name}, \text{vm}), \text{ avec :} \tag{8}$$

- name qui identifie le critère,
- vm qui est la valeur de la mesure et ses unités.

b) Objets issus de flux de mesures

Deuxièmement, nous formalisons les objets issus de flux de mesures et leur information qualité associée.

Définition 6. Considérons φ comme des flux de mesures de taille n et à une fréquence r , avec un ensemble d'attributs A_i estampillés $t_i (0 \leq j \leq m)$ avec m valeurs de mesure, dont $t_i \mid t_i < \infty, i > 0$.

De façon similaire à la modélisation des mesures, chaque élément du flux de données peut être associé à des informations qualité. Mais ceux-ci représentent une masse importante d'informations à traiter qui peut générer des dégâts au sein du système et d'exploitation des données. Ainsi, nous avons choisi de déterminer un intervalle du temps parallèle à celui de la campagne de mesures et du flux de mesures, afin d'assigner une temporalité aux évaluations des informations qualité.

Dans ce cadre, chaque attribut de mesure A dans le flux de mesures est composé par des intervalles de temps indéfinis d'une taille z qui contient les mesures et l'information qualité $w (Z) \rightarrow (1 \leq Z \leq z)$. Chaque intervalle du temps est identifiée par le temps de début t_{start} , le temps de fin t_{end} , la taille de l'intervalle du temps w avec leurs attributs A_i . Les mesures issues de capteurs $v_i^j (t_{start} \leq i \leq t_{end})$ contiennent un ensemble d'information qualité IQ , telle que chaque information décrit un critère qualité $q \in IQ$. Donc, l'intervalle contient une valeur pour chaque critère q_w^j . Par exemple, l'exactitude des valeurs sera comprise dans les attributs et dans l'intervalle de temps $w : A_w^j$. L'intervalle du temps w porte les valeurs de mesures comprises depuis le temps de début de la prise de mesure, jusqu'à la fin des mesures établies dans l'intervalle, ainsi que les critères qualité associés.

Pour évaluer les critères qualité dans ce cadre, il est nécessaire d'appliquer des fonctions d'agrégation pour tout critère évalué sur un ensemble de données définies. Ces fonctions d'agrégation ne sont pas fixes, elles dépendent de chaque application et des besoins (i.e. exactitude de la moyenne de température du mois). Dans cette thèse, on ne s'attache pas à définir des fonctions d'agrégation optimales pour un cas particulier. On définit notre modèle de façon générique afin d'appliquer n'importe quelle fonction selon les besoins.

Sensor_Id	3056X7T					
Timestamp	550	551	552	553	554	555
Temperature	23,5	23,5	23,7	24,02	23,8	23,8
exactitude	2,8	2,8	2,5	2,2	2,6	2,6
complétude	0,98	0,98	0,86	0,8	0,93	0,93

Tableau 16 – Valeurs de température caractérisées par deux critères qualité

Suite à cette formalisation, nous concluons en que les valeurs qualité peuvent être gérées selon deux façons : La première considère que les valeurs qualité peuvent caractériser les mesures au moment où les données sont analysées (Tableau 16), pour chaque donnée analysée une valeur de qualité est enregistrée. Cette possibilité est fortement utilisée lorsque l'analyse des données est réalisée de façon irrégulière et à une fréquence de calcul très raisonnable. Le volume de données et la fréquence d'acquisition des données ne sont pas un facteur déterminant pour le calcul et le stockage des valeurs qualité.

La deuxième possibilité consiste à ce que les valeurs qualité soient générées selon des intervalles du temps (prédéfinies ou calculées) afin d'économiser des ressources (Tableau 17), comme nous l'avons cité lors de notre formalisation, notamment dans la gestion des flux de mesures. Cependant, dans ce type de données, les besoins en termes de qualité entraînent une déclaration d'intervalles de temps bien courts qui peuvent potentiellement surcharger le traitement. L'utilité des métadonnées dans ce sens est importante car elles fournissent l'information qualité nécessaire pour évaluer la qualité en parallèle de l'exploitation des données acquises. Nous détaillerons l'application des métadonnées dans un contexte dynamique dans la section suivante (Section 5.2).

Sensor_id	305X6Z											
Timestamp	550	560	570	580	590	600	610	620	630	640	650	670
Temperature	23,5	23,5	23,7	23,7	23,8	23,7	23,7	23,9	23,7	24	24	23,8
exactitude				2,4				2,7				2,2
complétude				0,87				0,9				0,82

Tableau 17 – Valeurs d'un flux de température et informations qualité

Dans la suite, représentant la seconde phase de notre méthodologie, nous introduisons le principe de notre approche concernant l'évaluation de la qualité en temps réel. Après notre analyse des données et du modèle de qualité proposé, on constate que, dans le contexte temps réel, la dynamique des données pose le problème de la dynamique des informations qualité et donc des métadonnées. Si on considère qu'à un moment donné, les intervalles de temps nécessitent une évaluation de la qualité des données à une fréquence plus rapide, la masse d'informations à gérer en temps réel est également problématique. Ainsi, dans la section suivante, nous présentons d'abord une analyse et une structuration des métadonnées considérées comme source d'information qualité et ensuite nous présentons les différents éléments pris en compte pour l'évaluation de la qualité des données issues de capteurs.

5.2 Phase 2 : l'évaluation de la qualité des données issues de capteurs

L'objectif de cette phase d'évaluation consiste en la spécification des métadonnées nécessaires pour exprimer et gérer l'information qualité, l'évaluation et la gestion de l'information qualité selon les critères définis dans la phase 1 de notre méthodologie, afin d'aider l'utilisateur lors de la prise de décision en temps réel.

Dans notre approche, nous nous sommes attachée à la découverte de l'information qualité de données capteur en temps réel. Selon notre étude, le principal intérêt de la part des experts dans le domaine de la surveillance environnementale, est de comprendre ce qui se passe dans le monde réel, et d'avoir des moyens de prendre de décisions de la façon la plus fiable possible. En conséquence, dans notre approche, nous défendons l'importance de la qualification et la communication de l'information qualité au regard de la correction ou l'amélioration des données acquises.

Nous proposons donc, dans cette **Phase 2 : l'évaluation**, de cette méthodologie, la spécification et gestion des informations qualité contenues dans des métadonnées, ainsi que la spécification des critères qualité sélectionnés lors de notre étude et leur évaluation. Cette phase est composée de trois étapes importantes (Figure 36) :

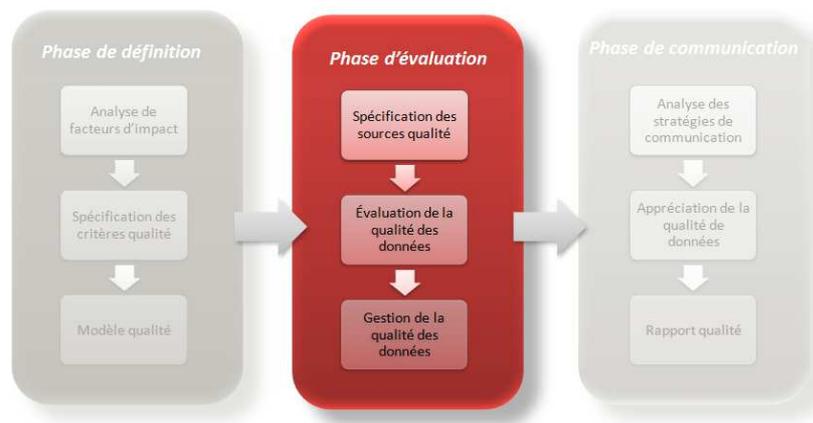


Figure 36 – Phase 2 : l'évaluation de la qualité des données issues de capteurs

- 1) La spécification des sources d'information qualité
- 2) La spécification et l'évaluation des critères qualité définis
- 3) La gestion de la qualité de données.

Dans cette section nous présentons tout d'abord l'analyse menée pour la spécification des sources d'information qualité : métadonnées spatio-temporelles. Cette structuration nous permettra de gérer l'information contextuelle et sur la qualité (critères qualité et mesures associées).

Ensuite, nous présentons notre proposition d'évaluation de la qualité. Dans ce cadre, notre approche s'adapte à l'évaluation de la qualité des données statiques autant que dynamiques. Nous évaluons donc, les critères qualité précédemment définis. Ces

critères prenant en compte les aspects lors de l'acquisition des mesures, afin de qualifier les données selon les sources, en regardant les possibles erreurs dues à l'appareil de mesure et leur contexte de mesure, statistiques, etc., ainsi que sur l'évaluation de critères qualité qui considèrent le contexte temporel lors de la phase de traitement.

La phase d'évaluation implique également l'analyse de résultats de l'évaluation de critères qualité. Les informations qualité obtenues lors de cette évaluation seront donc comparées avec une valeur de référence. Les valeurs de référence peuvent être spécifiées par les experts lorsqu'une observation est programmée, ou être le résultat de divers calculs exécutés en temps réel.

Ensuite, ce processus d'analyse sera exécuté de façon continue (si besoin) afin de contrôler des résultats en temps réel, qui permettra de vérifier les niveaux de qualité pertinents pour l'exploitation des données. Ce processus, s'appuie sur les dérivations statistiques de base (min, max, ratio...etc.) dont l'utilisateur a prédéfini les valeurs de référence.

Le paragraphe suivant décrit plus en détail les divers éléments pris en considération pour l'évaluation de la qualité des données selon ces termes.

5.2.1 Métadonnées spatiotemporelles en temps réel : sources d'information qualité

L'objectif de ce paragraphe est de définir les possibles sources d'information qualité dans un contexte dynamique comme celui de la surveillance (Figure 37).

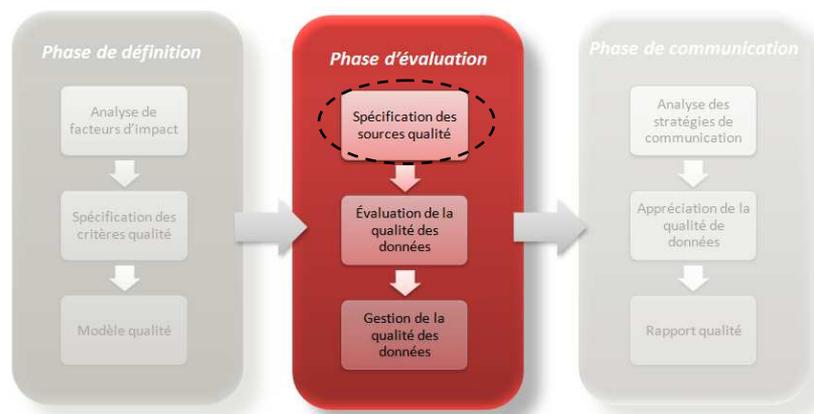


Figure 37 – Phase 2 – Spécification des sources qualité

Pour ceci, nous nous attachons à l'étude des métadonnées comme source d'information qualité. L'utilisation des métadonnées dans un contexte temps réel, requiert une classification de métadonnées spécifiques pour chaque domaine d'application, afin de se limiter aux métadonnées indispensables selon le contexte d'utilisation. Ces métadonnées doivent contenir des informations pertinentes qui permettront d'identifier et de mieux exploiter le système.

Dans notre activité de recherche, nous avons identifié quelques travaux posant aussi le problème des métadonnées dans un contexte temps réel (i.e. [MMI'07], [TAO'06]). Ces études ont montré un intérêt concret à la gestion des données océaniques et les données concernant des phénomènes naturels. Leur principale problématique concerne la gestion des grandes quantités d'informations en temps réel.

Par contre, il existe peu de travaux et aucun standard proposé pour la modélisation et structuration des métadonnées spatiotemporelles issues de capteurs, notamment sur les métadonnées qualité. Toutefois, les spécifications de l'Open Geospatial Consortium [OGC'09] proposent des langages de description pour les données capteurs, orientées à la découverte, l'échange (métadonnées) et le traitement des observations issues des capteurs et de tout type de systèmes capteurs. Bien que ces spécifications soient riches en contenu, et que nous nous inspirions de leur structure pour notre approche de modélisation, elles n'abordent pas la gestion de la qualité des données issues de capteurs de façon importante.

Dans ce contexte, l'OGC [OGC'09] s'appuie sur les spécifications géographiques, notamment l'ISO 19115 [ISO'03a]. Cette norme offre des réponses pour la structuration de métadonnées spécialement concernant la qualité des données géographiques. Toutefois ces normes ne prennent pas en compte la dynamique des données voire des métadonnées. Une spécification des métadonnées issues de données capteurs intervenant dans l'évaluation de la qualité des données capteurs est donc indispensable.

Dans le contexte temps réel, en raison du dynamisme des données et de leur exploitation, une analyse des métadonnées est pertinente. Nous proposons donc, une analyse des métadonnées suivant quatre axes d'étude [Gutiérrez'09] : la généralité, l'espace et le temps, les caractéristiques des processus temps réel et la granularité des données. Cette analyse nous permettra de définir les types et spécificités des métadonnées dans un contexte temps réel, notamment sur la qualité des données. Nous présenterons également, l'ensemble de critères à prendre en considération pour leur extraction et leur mise à jour.

5.2.1.1 Analyse et structuration des métadonnées qualité

Les métadonnées sont sélectionnées pour fournir des informations sur la disponibilité, l'adéquation à l'usage des données, etc. Nous effectuons dans ce paragraphe, une analyse des métadonnées suivant quatre axes d'étude (Figure 38) : la généralité (métadonnées génériques et applicatives), l'espace et le temps (métadonnées statiques et dynamiques), les processus de traitement et de transformation (métadonnées d'acquisition, de transformation et d'exploitation) et la granularité des données (mesures, ensemble, flux).

Ensuite, nous décrivons plus en détail chacun de ces axes [Gutiérrez'09].

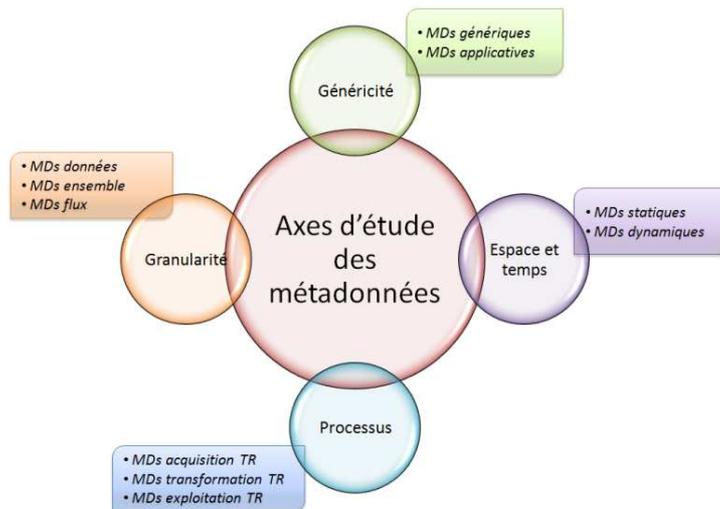


Figure 38 – Axes d'étude des métadonnées spatio-temporelles temps réel [Gutiérrez'09]

a) Axe 1 : Généricité

Une des difficultés précédemment soulevée concerne la masse d'informations à gérer dans les métadonnées, difficulté qui devient cruciale dans un contexte temps réel. Dans les SIG, nous pouvons apprécier deux niveaux d'information [Taoussi'03]: le niveau générique, qui décrit les aspects de l'information commune aux applications et le niveau spécifique, qui décrit les aspects de l'information spécialisée selon le type d'application.

Il apparaît donc nécessaire de définir une classification des métadonnées afin d'établir les métadonnées génériques qui se révèlent indispensables et constitueront ainsi un socle commun à l'ensemble des applications temps réel. Les autres métadonnées applicatives, plus spécifiques ou épisodiques, pourront alors être traitées au fil de l'eau.

Nous distinguons donc :

- Les *métadonnées génériques* qui décrivent l'information minimum requise pour l'identification et la définition d'un ensemble de données. Ces métadonnées représentent le socle des éléments indispensables et communs à la documentation d'une application en temps réel.
- Les *métadonnées applicatives* qui décrivent l'information spécifique à chaque application, par rapport au type ou domaine d'application et à leurs caractéristiques. Ces métadonnées sont optionnelles et peuvent être ajoutées selon les besoins de l'utilisateur.

b) Axe 2 : L'espace et le temps

Nous considérons dans notre analyse l'importance de la constante évolution des valeurs de certaines informations (position du capteur mobile, mesures continues, etc.). Certaines données peuvent faire évoluer le contenu des métadonnées et ainsi leur

concéder une caractéristique dynamique. Nous proposons d'abord deux types de métadonnées selon l'évolution des données dans l'espace et dans le temps qui peuvent aider à mieux traiter les métadonnées spatio-temporelles temps réel : *les métadonnées statiques* et *les métadonnées dynamiques*.

Les métadonnées statiques sont habituellement exploitées dans des applications traditionnelles. Ces métadonnées décrivent des lots de données et leur valeur reste stable durant le développement de l'application. Dans le contexte temps réel, une métadonnée peut alors être considérée comme statique lorsque sa valeur n'évolue pas au cours d'une campagne de mesures.

Les métadonnées peuvent être caractérisées comme *dynamiques* en raison du comportement dynamique des données. Ces métadonnées représentent une information variable dans le temps. Nous considérons que l'information qui contribue à renseigner une métadonnée dynamique ne fait pas partie de la campagne de mesures, puisque ce n'est pas une donnée, mais la valeur de cette métadonnée est susceptible d'évoluer ponctuellement au cours du temps et notamment durant une campagne de mesures (fréquences d'acquisition, capacité de mémoire, etc.). Par exemple, la qualité des données peut être affectée au cours de la transmission temps réel car elle dépend particulièrement de la qualité de transmission, du contexte de l'acquisition... etc. qui sont des éléments stockés dans les métadonnées mais néanmoins susceptibles d'évoluer au cours du temps durant la période de campagne de mesures. Cette information doit permettre au système de s'adapter au contexte évolutif de l'application, et analyser de près les aspects dynamiques qui peuvent servir à l'optimisation du système.

Prenons un autre exemple : la position du capteur. Si ce capteur est fixe, notamment dans le domaine de la surveillance de phénomènes naturels, la position du capteur peut être considérée comme une métadonnée statique. Si le capteur est mobile, la position est une donnée et non plus une métadonnée. Si le capteur est agile, la position du capteur pourra être considérée comme une donnée ou encore comme une métadonnée dynamique selon le contexte d'exploitation et la fréquence de ses changements.

c) Axe 3 : Processus

Les flux de données spatio-temporelles issues de capteurs subissent généralement différents processus dans un système d'information temps réel. L'impact de ces processus sur les données et notamment sur la qualité des données n'est pas anodin et peut donc fournir des informations utiles à stocker dans des métadonnées. Nous distinguons trois processus de traitement temps réel (Section 4.1) :

- l'acquisition en temps réel,
- la transformation ou le traitement des données en temps réel,
- l'utilisation ou exploitation des données en temps réel.

En effet, lors de chaque processus, l'exécution en temps réel modifie les propriétés des données et donc la qualité de ces données. La proposition de représenter l'information de chacun de ces processus dans des métadonnées permettra à l'utilisateur

de mieux gérer les ressources du système, les méthodes de traitement et de stockage, les moyens d'utilisation de l'information, tout au long de l'application.

d) Axe 4 : Granularité

Dans le contexte temps réel, la question de la définition des métadonnées en fonction de la granularité des données est un aspect important. La granularité des données versus la granularité des métadonnées est d'ailleurs une problématique étudiée pour des données géographiques statiques [Deville's'06].

Dans des applications spatiales exploitant des bases de données géographiques et cartographiques, par exemple, les métadonnées peuvent porter sur la collection complète des données, sur chaque thème (routes, fleuves, altitude, etc.) ou encore sur chaque couche géométrique ou topologique voire sur chaque primitive géométrique.

Au niveau temps réel, il est également possible d'envisager des métadonnées portant sur une donnée, sur un ensemble de données ou sur un flux de données. De plus, ces métadonnées peuvent porter des informations plus complexes comme les agrégations (i.e. moyenne). L'ensemble de données peut être défini selon des critères spatiaux, temporels ou spatio-temporels. Par exemple, il peut être nécessaire d'évaluer la qualité d'informations sur une zone ou/et pendant une période de temps, ou la qualité sur les températures dépassent un seuil donné.

Nous pouvons définir que pour un capteur agile par exemple, à une position donnée et à une mesure, peut être attachée une métadonnée si l'utilisateur le demande. De la même façon, une métadonnée peut porter sur un ensemble de mesures faites par un ou des capteurs, lors d'une campagne de mesures voire lors d'une période de temps au sein d'une campagne de mesures (utile notamment lors du suivi de situations de crise). Également, nous pouvons avoir des métadonnées qui portent des informations complémentaires sur le flux des données d'un capteur.

Actuellement, nous pouvons souligner le manque des données associées aux données (métadonnées) en temps réel, les limitations dans le processus d'acquisition, et la qualité de données elle-même rend la gestion et la communication de la qualité beaucoup plus restreinte.

En conséquence, certains défis dans la structuration et la gestion des métadonnées en temps réel surgissent. Par exemple, dans des applications exploitant des mesures issues de capteurs, certaines métadonnées sont distribuées dans les capteurs, donc des techniques adéquates pour l'extraction et le stockage de ces métadonnées sont nécessaires. Certaines des informations contenues dans les métadonnées issues de capteurs sont dynamiques (i.e. l'état du capteur, leur énergie, la capacité de stockage) et d'autres restent statiques (i.e. identification, description...). Par ailleurs, certaines métadonnées sont historiques et d'autres générées en temps réel, donc la gestion des métadonnées est rendue plus difficile. Conjointement, la gestion de métadonnées en temps réel consomme une quantité importante de ressources, d'où le besoin de proposer des bonnes stratégies de gestion.

En résumé, la difficulté globale dans la structuration et la gestion des métadonnées en temps réel est alors, d'établir la granularité des données voire la granularité des métadonnées (portée des métadonnées) et de définir la faisabilité d'extraction ou d'annotation automatique de ces informations en temps réel.

En effet, le problème de la masse d'information à gérer se pose. Il est alors évident que, dans ce contexte, la quantité des métadonnées à gérer peut égaler la quantité des données. Techniquement et actuellement, cette solution reste à éviter.

Dans le paragraphe suivant, nous abordons ces défis. Nous proposons une structuration des métadonnées selon les spécificités des données et du contexte d'application, ainsi que des principes sur leur gestion et mise à jour.

5.2.1.2 Structuration des métadonnées spatio-temporelles issues de capteurs

Les applications dans le domaine de la surveillance exploitant des réseaux capteurs, notamment dans le domaine environnemental, requièrent des métadonnées à jour, car l'accès à l'information doit être adaptatif à l'évolution des caractéristiques du phénomène au cours du temps. Ces métadonnées présentent une fréquence de mise à jour en continu, c'est-à-dire, que la production de métadonnées est intégrée au processus de mise à jour du système. Également, l'information qualité contenue dans les métadonnées peut changer dans des délais de temps assez courts, donc, les valeurs représentatives de la qualité ne sont pas validées pour une longue durée. Les métadonnées doivent donc respecter les contraintes des niveaux de conformité de la qualité avec les valeurs qualité mesurées, c'est-à-dire, la comparaison entre les valeurs de référence et les valeurs réelles mesurées à l'instant t .

Nous définissons dans notre approche, les métadonnées spatio-temporelles en temps réel comme étant : « *des données sur les données qui répondent aux besoins de l'utilisateur pour une localisation et un temps (date) précis* » [Gutiérrez'07].

Afin de fournir à l'utilisateur une description complète sur les données exploitées, nous proposons une structuration conceptuelle pour définir les types de métadonnées à exploiter. Dans notre contexte, cette structure doit fournir le cadre nécessaire pour décrire les données spatio-temporelles issues de capteurs, en considérant les paramètres d'une observation, les éléments qui la produisent et la qualité qui les caractérise. La définition d'une spécification des métadonnées dans un contexte temps réel est une tâche difficile, car elle implique la catégorisation des éléments à décrire, ainsi que la description de chacun des éléments. La catégorisation des éléments impliqués dans les métadonnées dépendent fortement de leurs objectifs en termes d'utilisation et des objectifs prévus.

En effet, en déterminant une classification des métadonnées, le problème de la masse d'information à gérer se pose. Également, il est important de considérer la façon dont le système et l'information devront être décrits par les métadonnées.

Dans ce cadre, pour la structuration des métadonnées spatio-temporelles issues de capteurs, nous considérons deux aspects principaux : *la dynamique*

(temps et espace) et la *granularité* (niveau d'abstraction). Traditionnellement, l'utilisateur a besoin de l'information sur un phénomène qui évolue soit en temps soit à des niveaux d'abstraction spatiaux (coordonnées, zone...). Des métadonnées au niveau global sont aussi importantes que les métadonnées plus détaillées. A cette fin, les métadonnées dans notre contexte d'application doivent prendre en considération l'information suivante :

- L'information concernant l'observation, avec une description générale de leurs spécificités.
- L'information concernant l'identification des capteurs et leurs capacités (l'opération, l'état, spécificités, position, etc.).
- L'information concernant la qualité des données, permettant à l'utilisateur d'exploiter les critères internes et externes pour évaluer leur qualité.

Selon cette spécification, afin de structurer ces métadonnées, nous avons pris en considération les attributs spatiaux (un objet dans l'espace), temporels (un objet dans le temps), et les attributs orientés qualité (attributs du capteur, du processus de traitement, etc.). Nous proposons donc quatre classes de métadonnées pour la découverte d'informations dans un système de surveillance en temps réel : *Métadonnées Observation*, *Métadonnées Capteur (statique et dynamique)* et *Métadonnées qualité dynamique* (Figure 39).

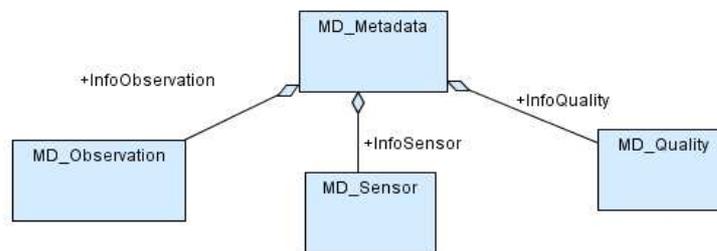


Figure 39 – Modèle global des métadonnées spatio-temporelles proposées

Dans la suite, nous décrivons notre proposition de catégorisation et les éléments pris en considération dans chacune des classes (plus en détail sur les annexes).

a) Métadonnées observation

Dans le même esprit que l'OGC et le SWE [SWE'09], nous proposons dans cette classe de fournir suffisamment d'informations sur la généalogie d'une observation, afin de déterminer la fiabilité d'exploitation pour l'aide à la décision (Figure 40).

Ces métadonnées concernent l'information décrivant l'observation d'un phénomène. Le contenu de ces métadonnées est un ensemble d'éléments catégorisés en *GeneralInfo* et *Contact*.

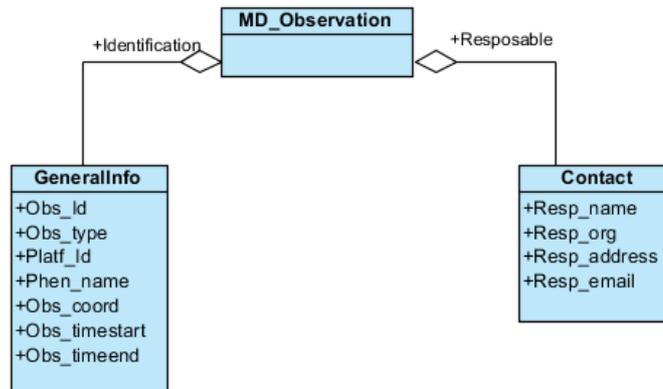


Figure 40 – Composition des Métadonnées Observation

- **GeneralInfo** : contient l'information nécessaire pour l'identification d'une observation, décrit comme suit.

[Obs_Id, Obs_Type, Platf_Id, Phen_name, Obs_coord, Obs_timestart, Obs_timeend], dont chaque ensemble porte l'information nécessaire pour l'identification d'une observation (détails de ces éléments en la section Annexe).

Cette information est obligatoire pour toute observation, stockée une fois, et valable pour toute l'observation. Des mises à jour sur cette information sont effectuées à chaque déclenchement d'une observation.

- **Contact** : contient l'information du responsable de l'observation. Cette métadonnée fournit l'information suivante :

[Resp_name, Resp_Org, Resp_address, Resp_email]

Cette information est optionnelle et est incluse dans l'application lorsque le personnel scientifique en a besoin. Des mises à jour sur cette information sont effectuées selon les besoins.

b) Métadonnées capteur

Chaque type de capteur a tendance à être accompagné par ses propres métadonnées, ses propres formats de données, et son propre système d'exploitation.

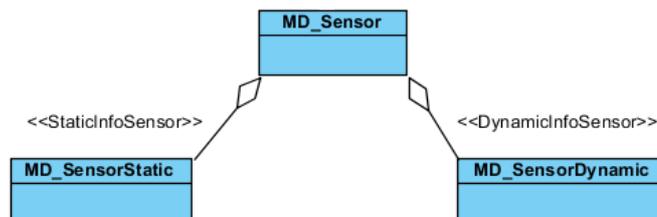


Figure 41 – Composition des Métadonnées capteur

L'objectif des métadonnées capteur est de fournir la possibilité d'identifier et d'évaluer les capacités d'un capteur. Nous proposons deux types de métadonnées liées au capteur : *Métadonnées Capteur Statiques* et *Métadonnées Capteur Dynamiques* (Figure 41).

▪ **Métadonnées capteur statiques**

Ces métadonnées représentent l'information concernant les propriétés et les caractéristiques des capteurs déployées pour effectuer une observation (Figure 42).

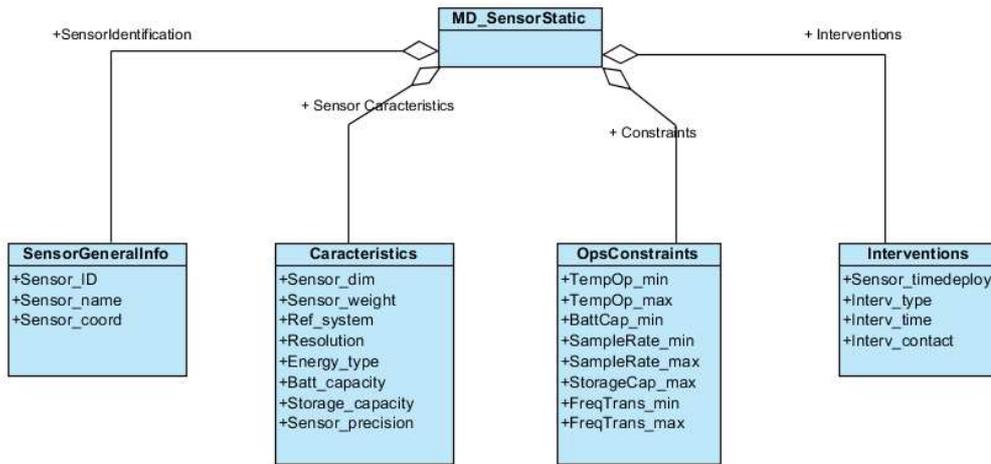


Figure 42 - Composition des Métadonnées Capteur Statique

Les informations concernant la description physique et d'opération du capteur sont données par le fournisseur lors de la fabrication. Ces informations sont stockées une fois, au début du déploiement du capteur. Toute mise à jour est effectuée lorsque le capteur change son état inactif en actif.

➤ **MD_SensorStatic** : contient l'information nécessaire pour l'identification du capteur déployé, en prenant en compte trois composantes :

SensorGeneralInfo, Characteristics, OpConstraints, Interventions (Figure 42).

○ **SensorGeneralInfo** – Information générale sur le capteur (cf. Annexe).

[Sensor_Id, Sensor_type, Sensor_coord]

○ **Caractéristiques** – Information concernant les caractéristiques physiques du capteur (cf. Annexe).

[Sensor_dim, Sensor_weight, Resolution, Energy_type, Batt_capacity, Storage_capacity]

○ **OpsConstraints** – Information concernant les contraintes opérationnelles du capteur.

[Temp_OpMin,Temp_OpMax, Batt_CapMin, Sample_RateMax, Sample_RateMin,
Storage_CapMin, Tx_FreqMax, Tx_FreqMin]

- **Interventions** – Information concernant les interventions réalisées sur les capteurs, voire les réparations, les déplacements, etc., décrite par (cf. Annexe) :

[Sensor_timedeployment, Interv_date, Interv_type, Interv_contact]

▪ Métadonnées capteur dynamiques

Ces métadonnées représentent l'information concernant les caractéristiques des capteurs qui varient au cours du temps, pendant une observation. Ces informations sont importantes pour comprendre l'état actuel du capteur (Figure 43).

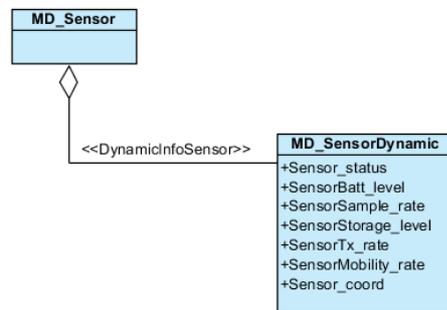


Figure 43 - Composition des Métadonnées Capteur Dynamique

- **MD_SensorDynamic** : contient l'information nécessaire pour vérifier l'état du capteur tout au long d'une observation, en prenant en compte les informations suivantes (Figure 43):

[Sensor_status, SensorBatt_level, SensorSample_rate, SensorStorage_level,
SensorTx_rate, SensorMobility_rate, Sensor_coord]

Les détails de chacun des attributs de cette classe sont donnés dans l'annexe.

c) Métadonnées qualité

Cette classe de métadonnées représente la source d'information pour caractériser la qualité de données et contient la description des éléments utilisés pour la mesurer. Nous considérons que la qualité des données peut changer quand une mesure est collectée ou quand la position de prise varie ou les deux.

Dans les sections précédentes, nous sommes partie de l'hypothèse que chaque donnée est associée à la valeur d'un critère pour une valeur d'attribut spécifique donnée (Section 5.1.3). Dans la suite de notre approche, nous introduisons un ensemble de métadonnées qualité qui caractérisent chaque valeur qualité de ces données. La qualité des données est associée à chaque donnée exploitée et est stockée avec les données pour faciliter leur découverte.

Plus spécifiquement, nous proposons pour décrire les valeurs qualité, un format générique des métadonnées qualité qui utilise un « indicateur » qui représente une ou plusieurs mesures qualité effectuées sur un critère spécifique. Ces métadonnées incluent la description des principes utilisés pour l'évaluation de la qualité.

Les métadonnées qualité sont décrites par (Figure 44) :

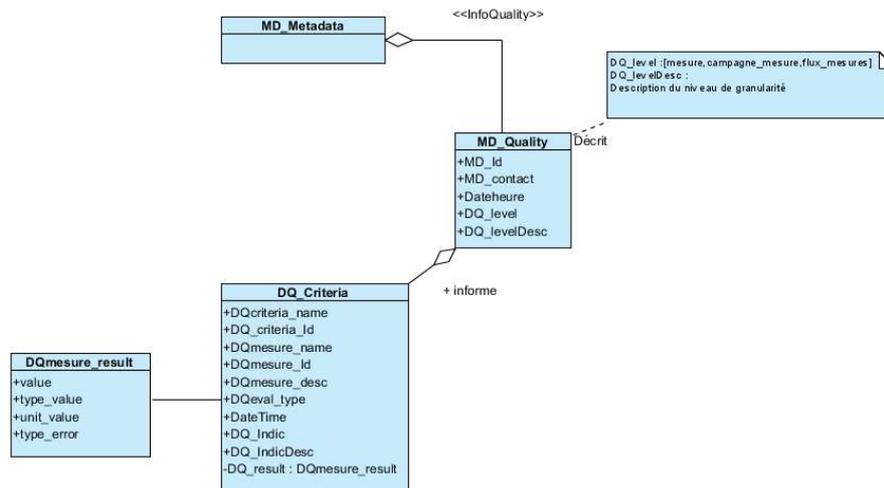


Figure 44 - Composition des Métadonnées Qualité

[DQcriteria_name, Qcriteria_Id, DQmeasure_name, DQmeasure_Id, DQmeasure_result, DQmeasure_desc, DQeval_type, DQeval_pro, DQDateTime, DQ_Indic, DQ_IndicDesc]

- DQcriteria_name, indique le nom du critère à évaluer.
- DQcriteria_Id, indique l'indicateur du critère à évaluer (optionnel).
- DQmeasure_name, indique le nom de la mesure utilisé pour évaluer le critère.
- DQmeasure_Id, indique l'identificateur de la mesure effectué (optionnel).
- DQmeasure_result indique la valeur représentative du critère évalué.
- DQmeasure_desc, décrit la mesure qualité utilisée.
- DQeval_type, décrit le type d'évaluation utilisé pour mesurer.
- DateTime est la date et heure de la mesure.
- DQ_Indic est le résultat d'une ou plusieurs mesures pour représenter un critère.
- DQ_IndicDesc décrit la signification du résultat de la mesure qualité appliquée.

A moins que chaque critère soit standardisé, différentes méthodes de mesure sont utilisées dépendant du fournisseur de données. DQmeasure_desc fournit une description de chacune des méthodes utilisées. En même temps, selon la mesure utilisée, chaque fournisseur de données peut donner différentes interprétations à ces mesures grâce à un indicateur DQ_Indic. Le DQ_IndicDesc doit communiquer aux

utilisateurs l'interprétation d'une valeur ou un résultat représenté par chaque indicateur. La combinaison de ces deux éléments fournit l'information nécessaire pour que l'utilisateur comprenne les caractéristiques de la qualité de la donnée.

Pour chaque critère, diverses métadonnées peuvent être nécessaires afin d'établir la spécification des niveaux de qualité. Comme nous l'avons spécifié dans la notre modèle de qualité (Section 5.1.3), les critères sont associés à des mesures ou flux de mesures issues de capteurs et décrits par des informations contenues dans des métadonnées. Pour ceci, nous décrivons dans la suite, l'ensemble des métadonnées nécessaires pour les catégories de la qualité proposées et leurs critères associées (Section 5.1).

Les informations comprises dans ces métadonnées, sont adaptées à l'ensemble de travaux proposés au sein de la gestion de la qualité de service et de la qualité dans les réseaux de capteurs [ISO'03a] [ISO'98]

▪ Métadonnées critères qualité contexte

- L'*exactitude* peut être considérée dans notre contexte comme la mesure selon laquelle les données sont « correctes » ou « précises » en fonction de la valeur de référence. Dans ce cadre, le `DQ_Indic` spécifie la valeur d'exactitude ou précision d'une donnée, qui peut être dans une échelle représentative, comme par exemple [0...1]. Le `DQ_IndDesc` décrit ce que cette échelle représente, par exemple 0 = erronée, 1 = exacte. Le `DQmeasure_desc`, décrit les méthodes utilisées pour leur mesure. Typiquement cette mesure est déduite de la comparaison de la valeur réelle avec une valeur de référence, ceci permettant d'économiser des ressources et de déterminer plus rapidement les problèmes en termes d'exactitude. Dans ce cas, les données qui ne sont ni précises ni exactes sont étiquetées pour que l'utilisateur puisse les reconnaître, les vérifier ou recalculer la mesure.
- Concernant la *complétude*, cette mesure adresse le problème de manque de valeurs causé par les défaillances ou le mauvais fonctionnement du capteur et estime le nombre des données arrivant à un point de consommation par rapport à la quantité maximale des données disponibles à ce point. Le `DQ_Indic` spécifié donc, le niveau de complétude que les données expérimentent à un instant donné. Le `DQ_IndDesc` décrit ce que l'échelle représentative de cet indicateur représente vis-à-vis la complétude analysée, par exemple complète ou pas. Le `DQmeasure_desc` décrit les méthodes ou tests utilisés pour évaluer telle critère.
- La *fiabilité* de la source représente la mesure selon laquelle la source de données (capteur ou réseaux) est fiable pour l'acquisition et le transfert des données. Un capteur est fiable en étant dans un environnement physique et d'opération convenables. Dans ce critère le `DQ_Indic` indique donc, le niveau de fiabilité de la source et `DQ_IndDesc` décrit de façon textuelle ce que la valeur de l'indicateur exprime. Le nom et le type de mesures utilisées pour évaluer ce critère peuvent être décrits grâce au `DQmeasure_name` et `DQmeasure_desc`.
- La *précision spatiale* qui représente la précision entre la position réelle du capteur et la position estimée peut être informée grâce à un indicateur de précision `DQ_Indic`

qui sera le résultat de la comparaison entre ces deux éléments. Leur niveau de précision, ainsi que la mesure ou mesures utilisées pour leur estimation pourront être également informées.

- La *fiabilité de la communication* sera établie comme la relation entre la quantité des données transmises et l'énergie dépensée lors de la transmission. Dans ce cadre, `DQ_Indic` spécifie la valeur de la fiabilité de la source pour une donnée, campagne ou flux collecté par un même capteur. Cette indicateur peut être représenté dans une échelle [0..10] dont les intervalles peuvent représenter « non-fiable, peu fiable, fiable, très fiable », par exemple. L'élément `DQ_IndicDesc` spécifie ce que les étiquettes du type « fiable » signifient. Le niveau de fiabilité est toujours relatif à l'application et à l'équipement. Également, dans ce type de contexte, il faut faire la relation entre cet indicateur et la source de données évaluée, donc l'identificateur de la source de données doit être inclus dans l'information (i.e. `Sensor_Id`).
- Nous considérons également les *conditions environnementales* comme une importante source d'informations dans ce contexte, notamment pour comprendre les comportements du capteur. Pourtant, nous exploitons ces informations issues des stations météorologiques situées à proximité du capteur ou issues des capteurs météorologiques situées dans les plateformes d'observation, ou encore exploiter les services météorologiques sur le web, pour extraire les conditions actuelles à un point géolocalisé bien précis. Ces informations sont à titre informatif. Dans ce cadre, le `DQ_Indic` est lié à un critère qualitatif, qui représente les bonnes ou mauvaises conditions environnementales près du capteur, également dans `DQmeasurement_name`, nous indiquons les divers éléments utilisés pour mesurer ce critère, comme la température extérieure, la pluie, la vitesse du vent, etc.

▪ **Métadonnées critères qualité interne**

Les métadonnées nécessaires pour les critères de la qualité interne :

- Le critère *actualité*, représente la mesure selon laquelle les données sont à jour. Dans ce cadre, le `DQ_Indic` spécifie la valeur d'actualité d'une donnée, qui peut être dans une échelle représentative, comme par exemple [0,1]. Le `DQ_measurementDesc`, décrit les méthodes utilisées pour leur mesure. Dans ce cas, normalement les données qui ne sont pas actuelles ou à jour sont étiquetées pour que l'utilisateur puisse les reconnaître (actuelle = 1, pas actuelle = 0).
- *Volatilité*. Selon la dynamique de la temporalité au sein des systèmes de surveillance, nous considérons une fonction basée sur [Scannapieco'04] qui représente la probabilité qu'une donnée peut varier selon le temps dans un intervalle du temps qui débute par l'instant d'entrée de la donnée et se termine par le temps d'expiration. Nous pouvons utiliser un `DQ_Indic` qui peut indiquer, par exemple : une volatilité faible (une valeur change très rarement au cours du temps), une volatilité moyenne (la valeur peut changer à une fréquence gérable), et une volatilité haute (une valeur varie de façon importante au cours du temps).

▪ **Métadonnées critères qualité d'utilisation**

Dans cette catégorie de métadonnées, nous remarquons que les critères qualité sont plus subjectifs, et leur composition entraîne une sélection beaucoup plus difficile des indicateurs, car ces critères dépendent énormément du type d'application. Cependant nous considérons ces critères fortement liés à la prise de décision de l'utilisateur en temps réel.

- *Actualité.* Mesure qui représente le niveau auquel les données sont à jour pour une tâche quelconque. Dans ce critère, l'information utilisée comme source pour l'évaluation de ce critère sont la date de la dernière mise à jour et la période du temps considéré comme valable pour une nouvelle mise à jour. Alors, le DQ_Indic informera est à jour ou pas, ainsi que la description des éléments utilisés pour l'établir.
- *Ponctualité.* Mesure qui représente le niveau auquel les données sont adéquates et disponibles pour une tâche. Dans ce cadre, le DQ_Indic spécifie la valeur de ponctualité, qui peut être représentée comme haute ou faible. Nous pouvons considérer dans ce cas, une ponctualité faible comme une prise de temps plus large pour mettre à disposition les données. Une ponctualité haute sera dans ce cas, un temps beaucoup plus rapide pour rendre disponible un ensemble de données.
- *Interprétabilité.* Mesure à laquelle les données sont dans la langue appropriée et les unités et définitions de données sont claires. Concernant ce critère, le DQ_Indic décrit la valeur qui permet à l'utilisateur de comprendre si les représentations graphiques ou numériques dans le système sont adéquates pour leur exploitation. Cet indicateur est qualitatif et complexe car il ne dépend pas d'une valeur numérique, sinon de l'évaluation de plusieurs éléments représentatifs pour l'utilisateur (i.e. langue adéquate, résolution de l'image, annotations...). Evidemment, ces éléments sont très divers et l'intervention de l'utilisateur est indispensable pour leur évaluation.

5.2.1.3 Gestion des métadonnées spatio-temporelles en temps réel

Les métadonnées sont actuellement utilisées afin de réaliser une meilleure gestion des données concernées et également pour archiver une information estimée pertinente. Comme nous l'avons cité dans la Section 5.2.1.1, dans un contexte temps réel, la pertinence de l'information est très relative en raison notamment de l'évolution en continu des valeurs et également en raison de la diversité des applications et des besoins.

Ces métadonnées peuvent être classées en différents types, mais des solutions pour leur gestion doivent être adoptées. A la différence du traitement des métadonnées traditionnelles, la gestion des métadonnées en temps réel, est caractérisée par la dynamique des métadonnées.

Afin de permettre une meilleure découverte de l'information et la représentation du monde réel, notre approche se concentre sur l'exploitation des métadonnées

conjointement aux informations dynamiques dans un système de surveillance [Gutiérrez'08]. Nous proposons donc d'intégrer les informations complémentaires, notamment sur la qualité des données au modèle des données capteurs (cf. Figure 15 – Section 4.2.1.1). Cette intégration est présentée dans la Figure 45.

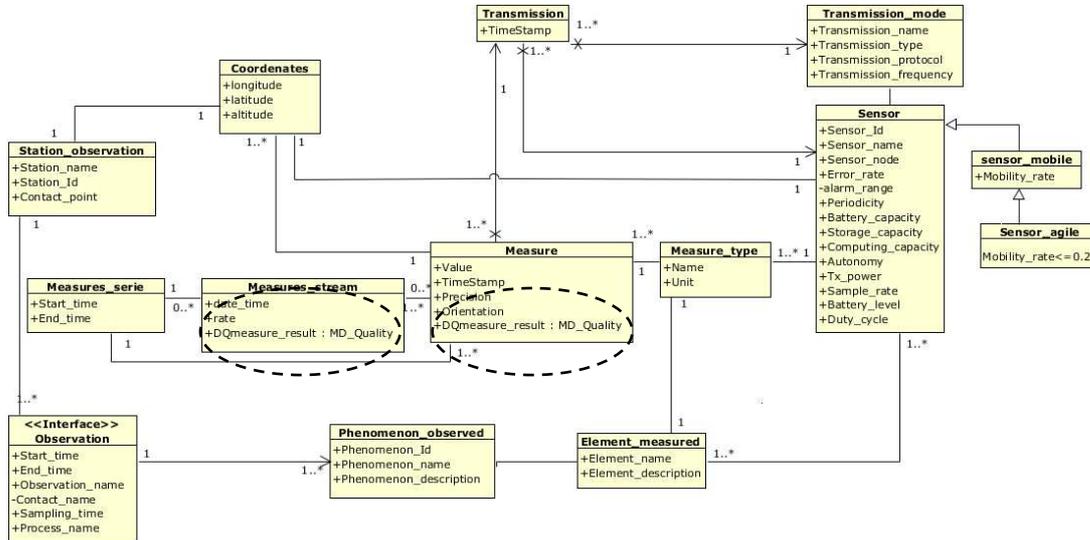


Figure 45 – Modèle intégral de données capteurs et la qualité de données

Nous décrivons plus en détail dans la suite, les spécificités prises en compte concernant l'extraction de ces métadonnées vis-à-vis de la dynamique de l'information, ainsi que la procédure de mise à jour adaptée à chaque catégorie de métadonnées.

5.2.1.3.1 Extraction des métadonnées spatio-temporelles en temps réel

La problématique est d'extraire de l'information en temps réel à partir de multiples sources disponibles et de l'organiser afin de permettre une exploitation optimale des métadonnées spatio-temporelles en temps réel par l'utilisateur. Comme nous l'avons décrit auparavant, les métadonnées dans ce contexte doivent tenir compte de la dynamique de l'information, des particularités des cas d'application ainsi que des besoins des utilisateurs. Nous identifions ainsi deux modes d'extraction des métadonnées : *l'extraction automatique* ou *l'extraction à la demande (utilisateur)*. En effet, les métadonnées sont généralement utilisées dans deux cas principaux : pour accéder à l'information ciblée en réponse à des requêtes ponctuelles des utilisateurs (*à la demande*) ou pour fournir des informations spécifiées au préalable (*automatique*).

a) L'extraction automatique.

L'extraction de façon automatique et en temps réel des métadonnées spatio-temporelles nécessite l'évaluation des spécificités. Nous avons précédemment présenté deux classifications des métadonnées suivant deux dimensions : *métadonnées génériques vs applicatives* et *métadonnées statiques vs dynamiques*. L'extraction automatique de

métadonnées statiques semble pertinente et réalisable ; cette extraction étant réalisée une seule fois à chaque campagne de mesures.

L'extraction automatique des métadonnées dynamiques n'est pas elle-même automatique et devra être décidée en fonction du contexte de l'application. En effet, il est nécessaire de prendre en considération le nombre de métadonnées à renseigner, la fréquence d'extraction, et la capacité de traitement (ressources du système) des informations données et métadonnées. Au regard de la classification des métadonnées : *génériques ou applicatives*, il semble pertinent de considérer que les métadonnées génériques devraient être extraites automatiquement. Alors que les métadonnées applicatives seront extraites à la demande.

En résumé, nous considérons que l'information extraite de façon automatique dans un contexte temps réel, doit se limiter à une information indispensable au fonctionnement du système.

b) L'extraction à la demande de l'utilisateur.

L'extraction à la demande permet à l'utilisateur d'interagir avec les éléments disponibles et de déterminer l'information qu'il considère pertinente pour les métadonnées dans une situation donnée. L'extraction de métadonnées à la demande doit demeurer ponctuelle et contextuelle. Il s'agit alors de disposer d'informations complémentaires lors de phases critiques ou de conditions exceptionnelles par exemple.

5.2.1.3.2 Mise à jour des métadonnées spatio-temporelles en temps réel

La mise à jour des métadonnées implique différentes techniques qui permettront à l'utilisateur d'exploiter les données les plus actuelles (i.e. trigger, manuelle, automatique). Contrairement aux métadonnées statiques, les métadonnées dynamiques doivent se maintenir à jour afin de capturer les changements temporels dans les métadonnées. On s'appuie sur les méthodes classiques de mise à jour des données pour proposer des mécanismes adéquats à la mise à jour des métadonnées dynamiques. Avec les mécanismes que nous proposons dans notre approche concernant la gestion des métadonnées, nous essayons principalement d'éviter la surcharge d'informations.

Nous pouvons observer que si une information est acquise à haute fréquence, la quantité des données par unité du temps augmente de façon significative. Alors, en comparant avec la génération des métadonnées, nous pouvons déduire que le traitement des métadonnées devient plus difficile dans ces circonstances. Il faut impérativement donner des priorités aux données ou proposer des méthodes de mises à jour adéquates.

Dans ce cadre, en relation avec les métadonnées statiques, quelques éléments des métadonnées peuvent être maintenus pour l'évaluation *à la demande*, comme c'est le cas des métadonnées génériques. La valeur est aussi mise à jour seulement quand l'élément est accédé. Cependant, ce type de mise à jour n'est pas toujours le plus indiqué, notamment sur l'information dynamique, car les valeurs de ces éléments peuvent varier plus rapidement que l'utilisateur ne le demande.

Il est donc possible de réaliser une mise à jour plutôt *périodique*, qui calcule les valeurs de métadonnées dans une fenêtre de temps spécifique. Une telle valeur est considérée valide pendant la période de temps définie par la fenêtre. Cette fenêtre représente un paramétrage lors de la mise en marche du système. Ce type de mise à jour est idéal pour les flux des données à haute vitesse, dont des agrégations représentatives des données par fenêtres sont exploitées.

Une autre technique possible dans notre cas d'application est la mise à jour des métadonnées par déclenchement (*trigger*). Certaines valeurs ne changent pas si fréquemment donc, une mise à jour périodique occasionne une dépense de ressources inutiles. Ce mécanisme de mise à jour permet l'actualisation des valeurs de métadonnées quand un élément dépasse une valeur de référence par exemple. Un autre exemple peut être le déclenchement de ce mécanisme quand un changement dans la valeur est détecté.

Quand une observation est définie et déclenchée, l'identification de cette observation est gérée par le système et les métadonnées sont stockées une fois au niveau du serveur (Figure 46).

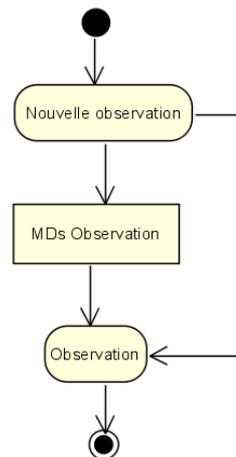


Figure 46 – Diagramme de séquence pour l'extraction et mise à jour des MD_Observation

Ces informations resteront valables tout au long de l'observation et seront stockées définitivement dans la base de données ou sur disque une fois que l'observation est finalisée ou son temps de validité terminé. La mise à jour de ces métadonnées est donc soit à la demande (à chaque nouvelle observation) ou soit périodique (si l'observation est pour un temps spécifique).

Ensuite, pour chaque capteur ayant été déployé, chaque nœud capteur initialise les opérations et toute l'information gardée en mémoire concernant l'information d'identification du capteur est communiquée à la station centrale (Figure 47).

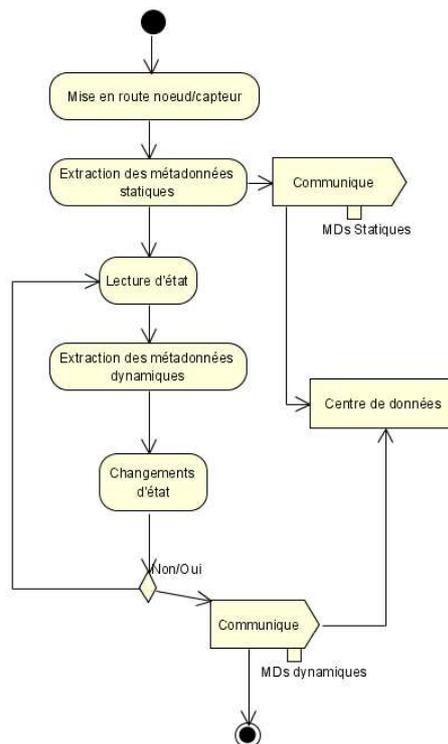


Figure 47 - Diagramme de séquences pour les métadonnées capteurs

Le système peut faire des requêtes de contrôle pour vérifier l'état du capteur, si non les informations concernant le capteur seront mises à jour, chaque fois qu'un changement est détecté.

Typiquement, ces informations sont réinitialisées quand le capteur passe du mode « inactif » au mode « actif ». Cela-dit, les mises à jour pour les métadonnées statiques du capteur sont plutôt faites par le déclenchement (de façon automatique), ou à la demande.

Les informations concernant la partie dynamique du capteur sont collectées principalement au sein des nœuds et ensuite communiquées au centre de données quand une mise à jour est effectuée automatiquement ou quand une nouvelle valeur est détectée (Figure 48).

Finalement, concernant l'information qualité, conjointement aux informations dynamiques du capteur et à l'arrivée d'une nouvelle mesure, le calcul des critères qualité correspondant à ces valeurs est effectué. Ainsi, chaque fois que le système détecte un changement dans l'ancienne valeur qualité associée à une donnée, à un ensemble ou à un flux de mesures, cette information est stockée en parallèle avec les données. La nouvelle valeur qualité est déterminée comme l'actuelle (Figure 48).

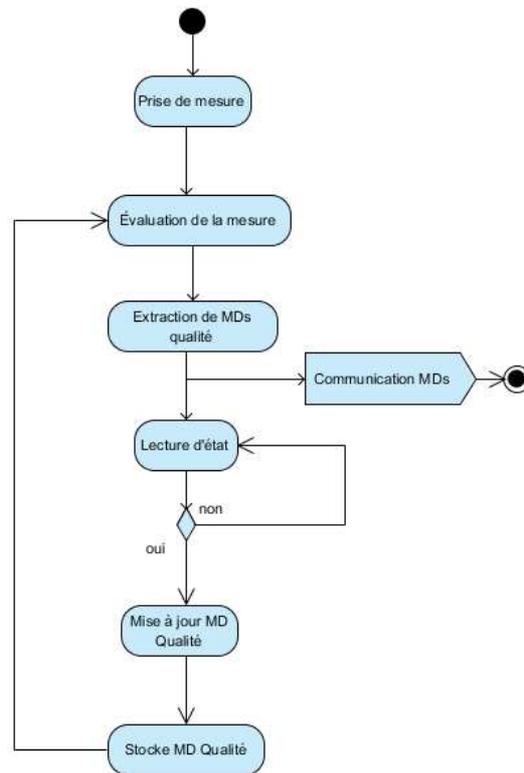


Figure 48 - Diagramme de séquences pour les métadonnées qualité

Après avoir sélectionné les divers critères qualité et les sources d'information utiles pour la caractérisation de la qualité des données capteurs, nous spécifions dans la section suivante, les éléments nécessaires pour effectuer l'évaluation de cette qualité en temps réel selon les spécificités et critères décrits précédemment.

5.2.2 Évaluation de la qualité des données issues de capteurs

Le but de cette étape est de qualifier et gérer l'information qualité issues des critères qualité (Figure 49).

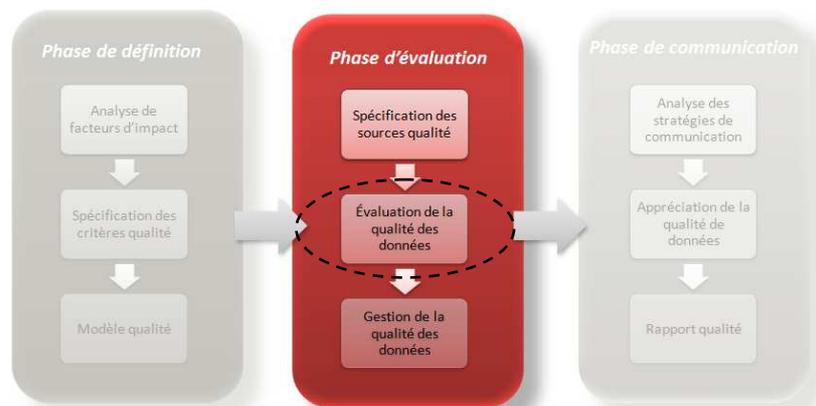


Figure 49 – Phase 2 – Evaluation de la qualité des données capteurs

Avec les différents concepts définis dans les sections précédentes, notamment le modèle de qualité (cf. Section 5.1.3) et les métadonnées (Section 5.2.1), l'utilisateur ou l'expert est capable d'évaluer la qualité des données de différentes façons. La première manière consiste à évaluer les critères associés à la source de données (l'exactitude, la fiabilité et la fiabilité de la communication) pour calculer l'erreur moyenne dans l'acquisition des données. La deuxième alternative consiste à utiliser la complétude pour connaître la quantité des données livrées. Une troisième façon serait que l'expert identifie le contexte temporel de la qualité de données, avec les critères liés à la temporalité, en considérant l'actualité et la volatilité des données. Finalement, une quatrième dont l'utilisateur peut choisir dans l'ensemble de critères disponibles ceux qu'il considère pertinents et les utiliser pour connaître la qualité des données exploitées. Nous partons de cette quatrième possibilité pour donner la liberté à l'utilisateur d'évaluer, selon leurs besoins, la qualité des données issues de capteurs exploitées.

Pour réaliser l'évaluation de la qualité de données, nous avons besoin d'un ou de plusieurs algorithmes d'évaluation qui prennent comme entrée les données capteurs (mesures ou flux de mesures) et calculent les valeurs qualité selon un critère ou plusieurs critères qualité sélectionnés par l'utilisateur.

Dans notre approche, nous proposons l'adaptation d'un algorithme utilisé pour l'évaluation de la qualité multicritères [Berti'99]. Cette algorithme permet l'évaluation d'un ou plusieurs critères qualité et d'assigner une valeur représentative à la qualité. Cette valeur représentative est estimée selon les poids que l'expert a signalés à chaque critère et les fonctions respectives de calcul à utiliser.

Cet algorithme appliqué dans la phase d'évaluation peut être implémenté au niveau de l'acquisition des données capteurs, avant ou après que les données soient prétraitées. Les résultats d'évaluation de tels critères qualité, accompagneront les données capteurs d'entrée à la phase de traitement ou de transformation. Ensuite, cet algorithme peut être réutilisé pour recalculer les critères au sein du serveur en prenant en considération les critères pertinents, ou l'appliquer de façon globale avant que les données soient exploitées par l'utilisateur. Nous estimons qu'il serait intéressant d'évaluer la qualité avant que les données soient exploitées par l'utilisateur au sein du serveur. De cette façon nous pourrions garantir à l'utilisateur des données analysées à différents niveau (acquisition et transformation), ainsi que l'utilisation de l'information qualité au même temps que les données capteur sont exploitées et

Pour calculer les poids de chaque critère nous nous basons sur le modèle LAW (*Linear Assignment of Weight*) qui permet de définir un coefficient pour annoter la qualité [Berti'99].

$$Q_{source}(M_i) = \sum_{k=1}^l W_k E_{ik} \quad (17)$$

Avec W_k le poids du critère k et E_{ik} l'évaluation du critère k pour une donnée m_i ou un ensemble de données $M\{m_1 \dots m_n\}$. Avec cette technique, nous pourrions fournir une valeur plus représentative de la qualité selon les critères sélectionnés pour leur évaluation.

Comme nous l'avons détaillé auparavant, chaque critère est mesuré selon des mesures directes ou indirectes avec des indicateurs objectifs ou subjectifs, notamment basés sur la normalisation des résultats. Ici, nous décrivons plus explicitement l'évaluation des données capteurs en s'appuyant sur le modèle d'assignation de poids (LAW) et selon les critères prédéfinis. Nous détaillons la procédure dans un l'algorithme décrit ensuite.

```

Algorithm_dataqualityeval
Let Qval[Qv] be the entry for the value v in the quality score storage
Init all the storage entries to 0
For each object o
    Find the set of attributes  $A\{a_1, a_2, \dots, a_n\}$  of o
/*n: number of attributes describing the object o */
    For each attribute  $a_i$  (with  $i=1, \dots, n$ )
        Find the set of values  $\{Qv_{ai,1}, Qv_{ai,2}, \dots, Qv_{ai,m_i}\}$  of  $a_i$ 
        For each value  $Qv_{ai,j}$  (with  $j = 1, \dots, m_i$ )
            /* $m_i$ : number of values for the attribute  $a_i$  proposed
            by the sources  $S_i$ */
                For each quality criterion  $QC_k$  in the metadata
                    associated with the value  $Qv_{ai,j}$  (with  $k=1, \dots, n_j$ )
                    /* $n_j$ : number of quality criteria for  $Qv_{ai,j}$ */
                        Find the set of values  $\{Qv_{ck,1}, Qv_{ck,2}, \dots, Qv_{ck,m_n}\}$  for which
                            the criterion  $c_k$  has been evaluated
                        /* $m_n$ : number of values for the quality criterion has
                        been evaluated ( $m_i \leq m_n$ )*/
                            For each value  $Qv_{ck,p}$  (with  $p = 1, \dots, m_n$ )
                                Find the source  $s$  of the value  $Qv_{ck,p}$ 
                                 $QVal[Qv_{ck,p}] = QVal[v_{ck,p}] + QVscore(Qv_{ck,p}, QC_k, s)$ 

```

Algorithme 1 – Algorithme pour l'évaluation de critères multiples de qualité de données inspiré de [Berti'99]

Le $QVscore(Qv_{ck,p}, QC_k, m)$ représente la fonction qui donne l'annotation de la valeur $Qv_{ck,p}$ pour le critère QC_k en prenant en compte l'annotation de la qualité de sa source. Cette fonction est réalisée avec la méthode LAW.

En s'appuyant sur ces éléments (fonctions et algorithme), nous pouvons procéder à l'évaluation de la qualité de données à exploiter. De façon générique, l'objectif de cette évaluation est l'identification des problèmes concernant la qualité à travers tout le traitement de l'information (erreurs ou faible qualité dans les données) et analyser les possibles causes.

En effet, nous considérons la possibilité d'effectuer une évaluation de façon différée ou *off-line*, ainsi qu'une évaluation en temps réel ou *on-line*. Donc, l'évaluation en différé, les fonctions d'analyse, typiquement déclarées en types de contrats, calculent les mesures issues de critères qualité, ces mesures sont par la suite stockées dans les métadonnées au sein de la base de données. Et, d'un autre côté, l'évaluation en ligne,

implique des contraintes sur la qualité de données qui sont déclarées comme des instances ou sous forme de requêtes (i.e. QWITH). Ces contraintes comparent les mesures pré-calculées du dépôt de métadonnées aux seuils attendus afin de construire un résultat de la requête de façon implicite. L'adaptation d'un langage pour la déclaration de contraintes impliquant la temporalité des données capteurs et leur qualité, représente une des perspectives importantes dans notre travail.

Une fois les contraintes et les instances déclarées, les mesures de la qualité de données sont vérifiées selon le niveau de granularité spécifié ou pour une source particulière de données à un instant t . Nous considérons que les contraintes sont basées sur la comparaison entre les valeurs mesurées et les valeurs attendues pour chaque critère de qualité. Au résultat obtenu de cette comparaison, un indicateur peut être associé pour représenter le niveau de conformité de la valeur mesurée.

Suite à la spécification d'une méthode d'évaluation de la qualité de données capteurs selon plusieurs critères, nous décrivons dans la section suivante comment cette information qualité peut être gérée en temps réel.

5.2.3 Gestion de la qualité de données issues de capteurs

Conjointement à la phase d'évaluation de la qualité des données issues de capteurs, d'autres aspects doivent être considérés, concernant principalement la gestion de la qualité des données (Figure 50) ; par exemple, les aspects concernant la mise à jour avec des changements dynamiques des valeurs qualité ou encore les méthodes d'accès à cette information.

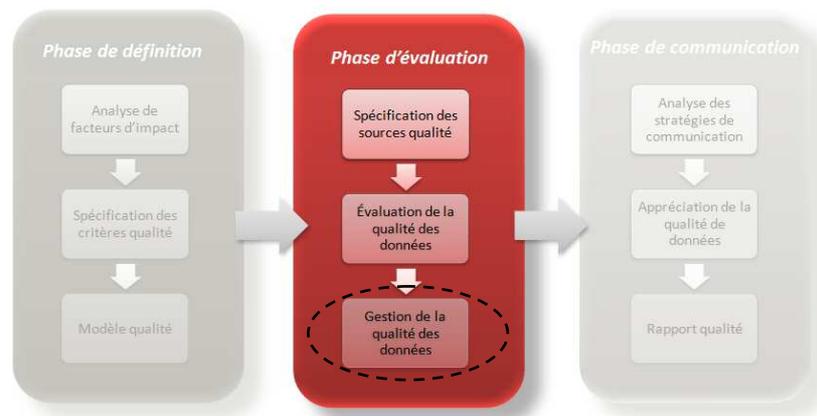


Figure 50 – Phase 2 – Gestion de la qualité des données capteurs

De façon générale, nous considérons que la gestion de la qualité des données est faite à partir de deux phases principales : l'action et la vérification. Ces phases sont typiquement composées de différentes tâches de gestion (i.e. définition des mesures, évaluation.) de qualité des données. Pour chacune de ces tâches des méthodes adéquates sont développées et/ou adaptés (i.e. évaluation interne et externe). De façon générale, on peut représenter la structure de cette gestion dans la Figure suivante (Figure 51).

Dans ce paragraphe, nous décrivons une structure pour la gestion de la qualité des données dans les systèmes de surveillance, en distinguant les principales activités qui doivent être prises en compte pour cette gestion.

La demande d'exploitation de plus en plus croissante dans le domaine de la surveillance, requiert une gestion et une exploitation plus flexible de cette information. Comme nous l'avons détaillé dans le début de cet chapitre, notre proposition de modèle qualité, consiste à exploiter l'information contenue dans les métadonnées décrivant des données capteurs et leur qualité. Ceci est dans le but de rendre l'information qualité facilement découvrable, avec un accès plus efficace et un stockage plus complet au sein des bases de données ou des entrepôts de données.

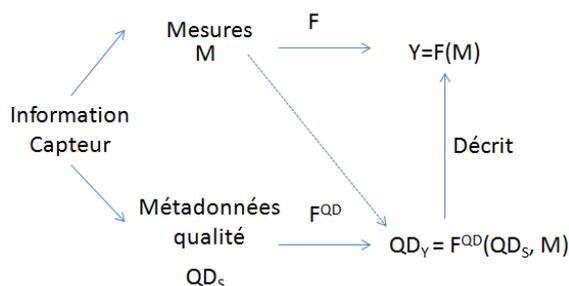


Figure 51 – Relations dans le traitement de la qualité de données capteurs (adapté [Klein'07])

Comme le décrit le travail de [Klein'07], la corrélation entre les données capteurs et le traitement de la qualité des données est évidente (Figure 51). Dans ce cadre, les capteurs S représentent la source de mesures M et des métadonnées relatives à la généalogie du capteur, mais aussi relatives à la qualité des données QD_s . Pendant le traitement des données capteurs, le résultat Y est dérivé des données brutes (mesures) M par l'application des fonctions complexes $F(M)$. La fonction de qualité des données F^{QD} est dédiée à calculer la qualité de données QD_Y , en décrivant les mesures dérivées $Y=F(M)$.

Dans notre approche nous considérons la qualité des données comme des métadonnées associées de manière relationnelle aux données. Chaque objet résultant d'une donnée, d'un ensemble de données ou d'un flux sera lié avec k critères de la qualité des données ($DQ_Criteria$). Dans ce cadre, il est important de maintenir les ressources au sein du serveur, donc l'information qualité n'est pas stockée à chaque valeur de mesure v_i . Les stratégies que nous supposons sont les suivantes :

- Les fonctions d'évaluation sont utilisées pour calculer les critères qualité et générer par la suite des métadonnées qualité associées à chaque valeur de mesure. Ces fonctions fournissent des indicateurs importants pour la caractérisation des problèmes qualité et la compréhension des anomalies dans les données.
- Chaque critère de qualité peut être évalué en continu ou selon la fréquence d'acquisition de la donnée, afin d'estimer sa pertinence à un instant t_i .

- L'information qualité est de préférence stockée chaque fois que sa valeur change au cours du temps (quand une nouvelle valeur de critère k pour un attribut a est détectée, une nouvelle valeur qualité est enregistrée), selon une période du temps spécifique (intervalle), ou chaque fois qu'un événement aléatoire est détecté.

Pour faciliter la gestion de la qualité de données variables dans le temps, les informations qualité accompagnent les données d'entrée et sont organisées dans des vues contenant l'information qualité conjointement aux données qui sont exploitées.

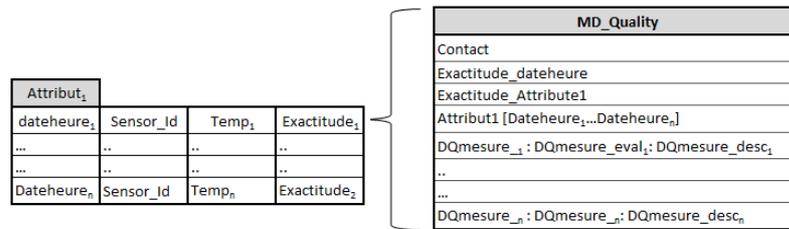


Figure 52 – Exemple d'interaction entre les données et les métadonnées en forme de table relationnelle.

De cette façon, la valeur du critère mesuré sera toujours associée à la mesure du capteur évaluée, permettant l'extension à tout critère qualité. Chaque nouvelle mesure qualité est agrégée et mise en correspondance avec des métadonnées qualité (MD_Quality).

La procédure de gestion consiste à séparer les informations qui viennent des données et celles qui viennent des métadonnées (Figure 52). Ensuite les données capteurs seront insérées dans leur schéma correspondant, et l'information qualité sera stockée si sa valeur est différente de la valeur précédente, à chaque instant de temps prédéfini ou pendant une période de temps également prédéterminée (début d'une observation, pendant un Δt , etc.).

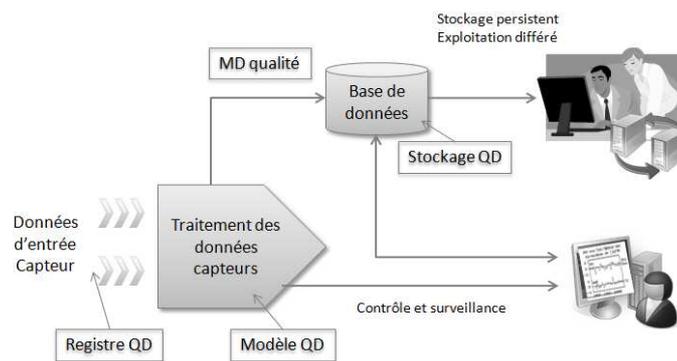


Figure 53 – Modèle de traitement de la qualité des données issues de capteurs

De façon générale, nous schématisons le déroulement de la gestion de la qualité des données jusqu'au moment du stockage dans la Figure 53.

Notre solution pour transférer et gérer la qualité des données consiste en l'enregistrement de la qualité de données à l'entrée des données capteurs. Pendant ce

processus, l'information qualité est capturée sur le capteur et l'environnement d'acquisition. Ensuite, le modèle de qualité doit être appliqué (Section 5.1.3, Figure 34), donc des critères qualité agissent sur des mesures et flux de mesures issus de campagnes. Egalement, à l'aide des métadonnées nous pourrions faire un stockage persistant de cette information au sein d'une base de données historique.

Plus spécifiquement, notre proposition de gestion de la qualité des données sur des données issues de capteurs se réalise dans des systèmes de surveillance intéressés par des questions du type : *Quelles sont les conditions du phénomène actuellement ? Qu'est-il arrivé il y a deux minutes ? Quelle est la différence entre le phénomène de maintenant celui de 3 jours auparavant ?...* D'une autre façon, nous devons avoir l'accès aux données historiques et aux données les plus récentes pour pouvoir répondre à ce types de questions, voire analyser et surveiller un phénomène quelconque.

Comme nous l'avons décrit au chapitre 3 de ce manuscrit des systèmes gestion de flux de données comme [Yao'02], [Arasu'03], [Abadi'03], proposent la gestion des requêtes en continu sur des flux des données. Cependant, ce type de systèmes ne prend pas en compte les caractéristiques spatiales des données capteurs. Dans cet esprit, une extension de ces travaux et qui permet la gestion de la dimension spatio-temporelle des données capteurs a été proposée [Koo'07]. Ceci est un système pour les réseaux de capteurs dans les systèmes de surveillance environnementaux.

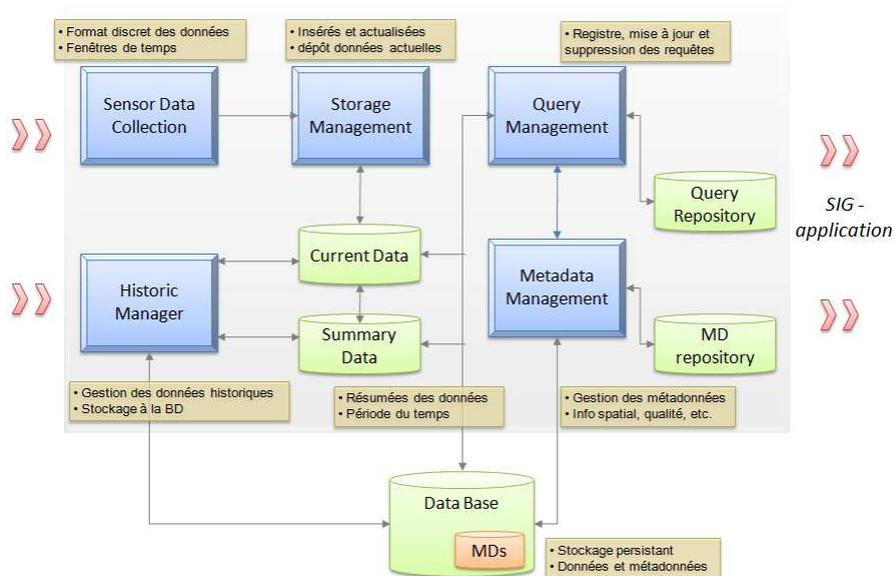


Figure 54 – Architecture du système de gestion des données issues de capteurs

Nous nous appuyons sur cette approche pour proposer une façon de gérer les données les plus récentes et les données historiques, ainsi que l'information contenue dans les métadonnées, concernant spécialement sur la qualité des données (Figure 54).

Comme nous le présentons dans cette Figure, dans ce système de gestion des données capteurs et des métadonnées, les données capteurs transmises par les capteurs sont collectées et rassemblées dans des paquets de données et transformées dans un format abstrait de données capteurs (i.e. *Sensor_Id*, *Measure_val*,

`date_time`, etc.). Le résultat est donc inséré dans des fenêtres de temps dans la section du *Storage Management*. Ici, les données issues de capteurs brutes sont insérées et stockées dans un dépôt de données actuelles. Les données expirées ou qui ont dépassé leur date de validité sont agrégées et stockées dans un dépôt pour des données agrégées. Ensuite le module *History Management* gère leur stockage dans une base de données, un entrepôt de données ou sur disque. D'autre part, la gestion de requêtes au sein du *Query Management* possède un dépôt de requêtes (*Query repository*) dont les requêtes sont enregistrées, mises à jour ou effacées. Le *Query processor* peut interroger les données actuelles et les métadonnées associées gardées en mémoire vive, ainsi que les informations gardées dans le «stockage persistant». De cette façon, l'accès aux informations historiques stockées est accessible. Finalement, une partie importante de ce système concerne la gestion des métadonnées (*Metadata Management*). Dans notre contribution, comme nous l'avons décrit précédemment, les métadonnées sont composées par des informations complémentaires, notamment sur l'information spatiale, l'origine des données, et la qualité des données. Cela dit, notre proposition de gestion des métadonnées ainsi que de la gestion de l'information qualité permettent à l'utilisateur d'accéder et exploiter les informations actuelles ainsi que historique sur les données issues de capteurs et leur qualité.

Dans la section suivante, nous allons détailler notre proposition pour communiquer l'information obtenue lors de l'évaluation et l'analyse de la qualité de données issues de capteurs.

5.3 Phase 3 : Communication de la qualité des données issues de capteurs

Au cours de ce chapitre, nous avons présenté les divers éléments à prendre en considération dans l'évaluation de la qualité des données issues de capteurs. Nous avons proposée un modèle de qualité composé par des catégories, critères, indicateurs et mesures qualité adéquates à notre contexte d'application. Egalement, nous avons proposé une méthode d'évaluation et de gestion de la qualité des données multicritères en temps réel. Ces contributions nous ont permis d'extraire l'information qualité sur des données capteurs.

Dans ce cadre, nous nous intéressons dans cette section à la façon de communiquer à l'utilisateur cette information qualité en temps réel. Actuellement, les systèmes de surveillance dans différents domaines exploitent les données issues de capteurs comme une source riche d'information pour comprendre les caractéristiques de l'environnement à observer. La visualisation d'une telle information représente un moyen d'interaction entre l'utilisateur et l'environnement. L'utilité de cette visualisation a été bien étudiée dans divers projets, par exemple le nowCOAST¹, ou le SwissExperiment [SwissExp'09], lesquels utilisent des interfaces web et services du mapping (WMS - Web Mapping Service ou le SensorMap du Microsoft Research) pour communiquer les données issues des stations d'observation dans un plan géospatial. En regardant l'utilité de ce type d'interfaces de communication, nous considérons important de réaliser une exploitation de l'information qualité selon ces paramètres. Plus précisément, la prise en compte de l'information qualité conjointement aux données issues d'observations, permettront d'aider à l'utilisateur lors de l'exploitation de l'information et la prise de décision.

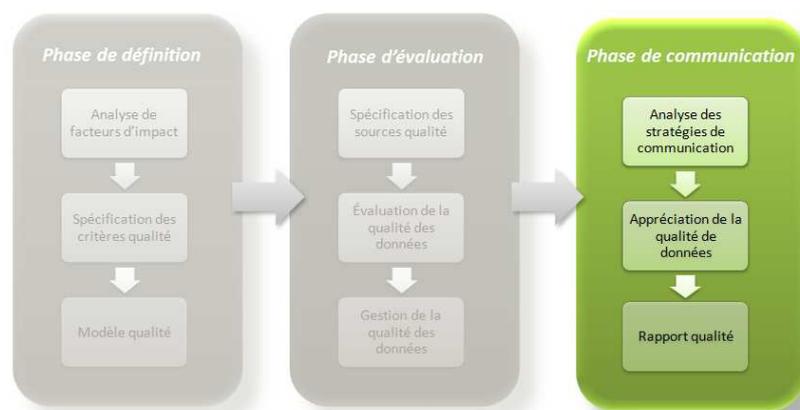


Figure 55 – Méthodologie SDaQ – Phase 3 : la communication

Dans ce contexte, afin de trouver la façon adéquate de communiquer l'information qualité en temps réel, nous présentons dans cette section la **phase 3 : la communication**. Cette phase est composée par trois tâches principales qui

¹ nowCOAST : GIS Mapping Portal to Real-Time Environmental Observations and NOAA forecasts

correspondent à l'analyse des stratégies de communication, la proposition d'une communication de la qualité des données issues de capteurs en temps réel basé sur l'exploitation des indicateurs qualité en forme graphique, ainsi que la proposition d'utilisation des rapports qualité sous forme de documents électroniques, dédiés à l'analyse et l'historisation de la qualité a posteriori (Figure 55).

Dans la suite, nous décrivons plus en détail les caractéristiques de chacune des tâches proposées dans cette phase. Tout d'abord nous mettons en évidence les stratégies possibles de communication pour la communication de la qualité des données issues de capteurs basés notamment sur l'exploitation des indicateurs qualité de façon visuelle. Ensuite, nous montrons l'appréciation de la qualité des données issues de capteurs selon l'utilisation de ces indicateurs. Nous finalisons cette section par la description de l'utilisation du rapport de qualité dans ce contexte.

5.3.1 Stratégies de communication de la qualité des données issues de capteurs : indicateurs qualité et rapport qualité

Nous allons décrire dans ce paragraphe comment nous percevons la communication de l'information qualité à l'utilisateur et sur quelles stratégies nous nous appuyons pour l'assister lors de la prise de décision (Figure 56).

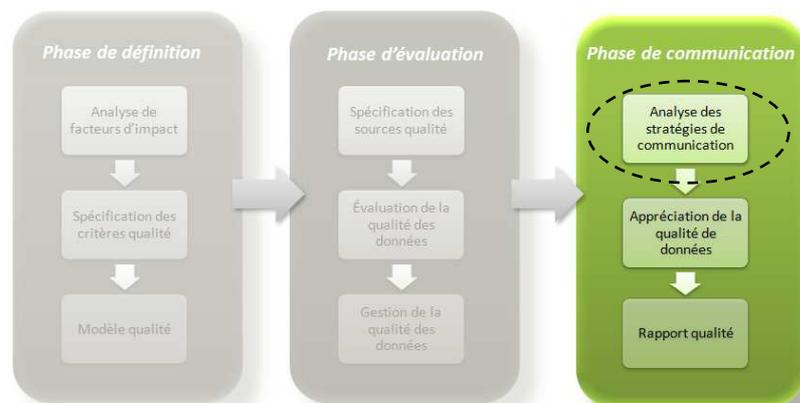


Figure 56 – Phase 3 – l'analyse des stratégies de communication

Du point de vue de l'utilisateur, la compréhension de l'information qualité vient typiquement des métadonnées transmises avec les données. Comme nous l'avons décrit auparavant (Section 2.2.2), l'objectif des métadonnées est de permettre aux utilisateurs d'évaluer l'adéquation à l'usage des données qu'il utilise. Cependant, les métadonnées ont été sujettes à diverses ambiguïtés, car souvent leur format et contenu sont inadéquats aux besoins de l'utilisateur final, et sont fournis séparément des données en limitant la possibilité d'exploiter facilement l'information qualité. Mais cela, n'indique pas que les métadonnées ne soient pas utiles, mais seulement qu'elles nécessitent des représentations plus adéquates aux objectifs de l'utilisateur et aux diverses applications. Alors, en fournissant des métadonnées appropriées, il serait possible d'évaluer plus

facilement le niveau de conformité des données. Nous pouvons donc communiquer et réaliser une évaluation plus efficace de l'information qualité, de façon que les utilisateurs puissent comprendre les défaillances dans leurs données exploitées.

En effet, dans le domaine géospatial, diverses approches se sont focalisées sur la recherche de meilleures façons de communiquer la qualité des données. Les diverses propositions reposent sur des méthodes procédurales [Hunter'99], sur des techniques de visualisation [Devilleurs'04], [Huth'07], ou encore sur la communication visuelle ou audio des signaux d'alertes aux utilisateurs [Gervais'06]. En général, le principe de ces travaux est de visualiser de façon graphique, les possibles erreurs en termes de qualité dans un espace géographique, ou exploitant les informations que certains critères qualité fournissent pour informer l'utilisateur de l'état d'une composante géométrique, d'une donnée, d'un jeu de données, etc. Cependant, nous remarquons que ces approches exploitent seulement des données statiques, et ne prennent pas en compte la dynamique des données capteurs en temps réel.

Dans notre contexte de recherche, nous sommes confrontés à des informations géolocalisées et variables dans le temps, avec des informations caractérisant l'état du capteur et la qualité des données issues de capteurs, entre autres. Lors de l'exploitation d'un système de surveillance, diverses informations changent au cours du temps, notamment le comportement du phénomène observé, les valeurs utilisées pour le déterminer, et par conséquent l'information relative à la qualité de données exploitées. Pour cela, nous avons besoin de choisir des moyens de représentation de la qualité plus adéquats qui s'adaptent à ces caractéristiques.

Traditionnellement, les systèmes exploitant des données dynamiques, utilisent des graphes pour montrer l'évolution des données au cours du temps. Par exemple, pour regarder l'évolution de la température dans une zone, ou encore la vitesse et la direction du vent (Figure 57).

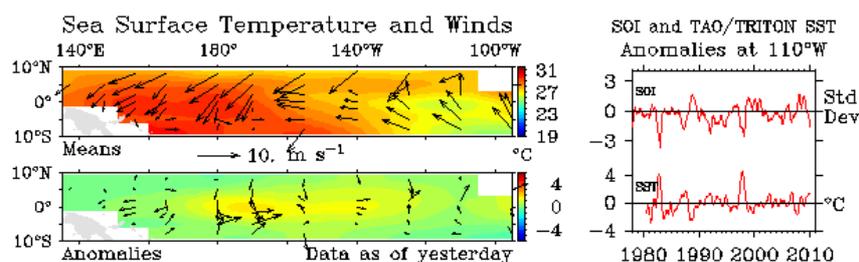


Figure 57 – Exemple des données en temps réel issues de bouées d'observation océaniques (TAO/TRITON)

Cependant, dans ce type de graphes il est très compliqué d'ajouter des informations complémentaires aux données, car l'espace représentatif est souvent limité et une surcharge d'informations peut facilement perturber l'utilisateur. Ainsi, dans des applications, dans lesquelles de grands volumes de données sont à traiter, il devient de plus en plus courant d'utiliser des indicateurs ou indices pour représenter de telles informations. Ceci permet de diminuer la quantité d'informations à communiquer.

Egalement, nous nous apercevons que dans le domaine géographique (i.e. ISO 19114), un rapport sur la qualité évaluée est indispensable. Ce rapport reflète les caractéristiques des données par la description des composantes de la qualité et des caractéristiques des données en termes des besoins.

Les rapports qualité sont très importants autant pour les utilisateurs de l'information que pour les producteurs d'information. Des rapports orientés selon les besoins des utilisateurs ainsi que vers les besoins des producteurs peuvent être générés. Les utilisateurs d'informations officielles (statistiques, agrégations, etc.) ont besoin d'avoir accès à un nombre important de mesures qualité et d'indicateurs pour mieux comprendre et mieux exploiter l'information. Egalement, ils ont besoin d'un aperçu sur la qualité afin d'observer et d'identifier les aspects à améliorer.

En conclusion, les rapports qualité et les indicateurs fournissent la documentation des caractéristiques sur la qualité de données nécessaires pour la prise de décision de l'utilisateur. Ces informations sont la référence pour l'évaluation de la qualité. Pour cette raison, ils représentent une forme importante d'aide à la décision.

Dans ce qui suit, nous présentons la description et l'utilisation des indicateurs visuels et des rapports qualité dans notre contexte d'application.

5.3.1.1 Indicateurs qualité

Afin d'éviter des mauvaises prises de décision, dans un contexte de surveillance, les utilisateurs ont besoin de comprendre d'une façon simple et rapide la qualité des données qu'ils exploitent. En effet, les utilisateurs sont exposés à des grandes quantités d'informations diverses et variables dans le temps, ce qui peut influencer la prise de décision de l'utilisateur dans un moment de crise. Dans ce cadre, notre approche propose l'adéquation et l'utilisation d'indicateurs qualité et leur représentation visuelle, pour aider les utilisateurs à mieux interpréter l'information concernant la qualité des données capteurs.

Comme nous l'avons décrit dans notre modèle de qualité (Section 5.1.3), la qualité des données capteurs est décrite par un ensemble de critères, lesquels sont représentés par des indicateurs. La valeur d'un indicateur peut être issue d'une mesure ou d'un ensemble de mesures sur des données (i.e. statistiques, algorithmes, etc.). Ces indicateurs peuvent caractériser des critères qualité qualitatifs ou quantitatifs par des valeurs dans des intervalles ou échelles représentatives (i.e. bon, mauvais). Dans le cadre d'un système de surveillance, l'importance de prendre une décision le plus rapidement possible est primordiale et dépend du niveau de criticité de l'application. En conséquence, nous considérons que l'utilisation des « *indicateurs* » (ou indices) pour communiquer la qualité sur des grands volumes de données dynamiques facilite le processus d'aide à la décision.

L'utilisation des indicateurs engendre principalement une méthode ou technique pour leur calcul ou estimation. Dans les travaux existants nous avons identifié deux façons de réaliser ce calcul [Aalders'98], [Pipino'02], [Devillers'04] : La première consiste à réaliser un calcul mathématique (i.e. statistique, une moyenne ou une

pondération) entre les résultats des mesures qualité réalisées afin de rendre une seule valeur à l'indicateur. Cependant, dans cette méthode, diverses hypothèses fortes sont souvent réalisées afin de rendre possible un tel calcul. Dans certains cas d'application comme la surveillance nucléaire, ce type d'hypothèses fortes est à éviter, et des algorithmes plus performants sont souvent souhaités. La deuxième façon est liée directement à l'utilisateur et ses besoins. En effet, il serait idéal que l'utilisateur soit capable de choisir (dans un rang limité) les éléments et paramètres à évaluer. Ceci lui permettra d'avoir une compréhension plus juste de la qualité des données exploitées, car chaque utilisateur possède une perception particulière sur les éléments qui peuvent la qualifier.

Ensuite, pour exploiter ces indicateurs, diverses représentations peuvent être utilisées (i.e. tables, images, symboles ou encore des alarmes visuelles ou audio). Les travaux proposés dans les divers domaines d'application (i.e. économique, médical, environnemental, géospatial...), montrent que ces représentations dépendent notamment du domaine d'application et peuvent être communiquées selon les besoins de l'utilisateur. Par exemple, dans le domaine géospatial, les indicateurs visuels sont exploités traditionnellement dans des tableaux de bord [Kaplan'92] [Devillers'02] [Huth'07] (Figures 58 et 59). Ce type de représentation de l'information est très utile, car il permet de concentrer toute l'information qualité sur un seul secteur, de façon à ce que l'utilisateur puisse se repérer facilement et évite la confusion avec le reste de l'information.



Figure 58 – Tableau de bord proposé par [Devillers'02]

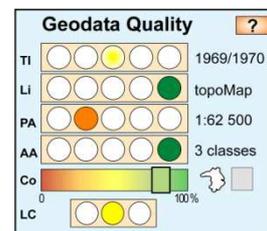


Figure 59 – Tableau de bord proposé par [Huth'07]

Avec ces pistes de communication de l'information qualité, nous pouvons compléter notre approche de définition de la qualité des données capteurs. Nous proposons donc, la mise à disposition des indicateurs paramétrables représentatifs des critères qualité et exploités de façon visuelle grâce à des symboles inspirés du domaine géospatial [Devillers'02], [Huth'07]. Cette approche permettra à l'utilisateur de choisir les aspects qui caractérisent la qualité de ses données exploitées selon ses besoins. Ainsi, ces indicateurs seront calculés selon les termes que l'utilisateur considère pertinents. Rappelons que chaque indicateur qualité pourra être lié à des mesures qualité relatives à une donnée particulière, un ensemble de données dans une période ou une région donnée, ainsi que sur un flux de données issues d'une source (capteur) quelconque, voire à des informations stockées dans les différentes classes de métadonnées.

Comme complément à ces indicateurs sur les critères qualité, nous nous sommes intéressés à la représentation d'un indicateur global représentatif de la qualité global des données exploitées. Cet indicateur global est basé sur un ensemble d'indicateurs sélectionnés, combinés et pondérés selon l'importance des critères qualité. La méthode d'évaluation de cet indicateur global a été définie en Section 5.2.3.

Finalement, en termes de visualisation, l'idée de notre approche est d'adapter les techniques proposées par [Devillers'04], [Huth'07] dans le domaine géospatial sur les données plus dynamiques comme les données issues de capteurs. Nous partons donc de leur étude et perception du concept des indicateurs représentatifs de la qualité des données, pour exploiter au mieux les techniques de communication de l'information qualité à l'utilisateur.

Comme nous l'avons décrit, afin d'évaluer la qualité des données, il est important dans un premier temps d'avoir une image claire et actualisée de la qualité des données via les indicateurs qualité. Cependant, pour mesurer de façon exacte la qualité des données, les indicateurs ne sont parfois pas suffisants. En conséquence des rapports qualité plus détaillés sont nécessaires. Nous décrivons dans la suite, les spécificités prises en compte concernant le rapport qualité dans notre contexte d'application.

5.3.1.2 Rapport qualité

Un rapport qualité fournit l'information sur les principales caractéristiques de la qualité des données qui permet à l'utilisateur d'être capable d'évaluer la qualité des données à exploiter. Dans un cas idéal, les rapports qualité sont basés sur les indicateurs qualité conjointement aux informations complémentaires ou contextuelles (i.e. type d'observation, état du capteur, les conditions environnementales).

Le principal avantage du rapport qualité est que les divers utilisateurs peuvent l'utiliser à des fins de surveillance ou d'analyse a posteriori. Cependant, dû à la diversité des besoins des utilisateurs, il est possible que certains utilisateurs ont besoin de plus d'information que d'autres. Dans ce cadre, il est possible d'envisager des rapports génériques ou standards avec l'information considérée comme indispensable dans un rapport qualité, et des rapports résumés qui contiennent l'information la plus pertinente et qui est plus simple à comprendre et disponible en temps. Un nombre important d'informations dans le rapport peut rapidement conduire à une surcharge et un accroissement du temps nécessaire à la génération du rapport. Certains travaux ont proposé l'utilisation de profils qualité [Batini'06] pour résoudre cette problématique. Au même niveau que pour les indicateurs, les profils qualité peuvent être utilisés pour les rapports qualité, permettant d'économiser du temps dans la configuration des besoins particuliers selon les types d'utilisateurs (i.e. expert ou non-expert). Dans notre étude, nous pouvons constater que dans un environnement dynamique, les préférences de l'utilisateur peuvent également changer. Pour cela, nous avons écarté la possibilité d'exploiter des profils, et nous préférons définir une structure standard assez générique dans son contenu qui rend la recherche de l'information pertinente pour la qualité pour tout type d'utilisateur.

Dans ce cadre, la norme ISO 19114, relative à l'évaluation des données géographiques, propose l'exploitation de métadonnées qualité pour documenter le rapport qualité. Comme source d'information qualité ils considèrent les éléments décrits dans le Tableau 18.

Information contenue dans le rapport qualité proposé par l'ISO 19114
<ul style="list-style-type: none">• L'identification du rapport (numéro/nom)• L'étendu observé (donné, ensemble...)• La mesure utilisée (formule, valeurs résultantes, unités, fiabilité et son unité)• Confiance au test de conformité (valeur, unité, documentation)• Type de méthode d'évaluation utilisé (direct, externe, échantillonnage...)• Description de la méthode utilisée (assumptions basiques, algorithmes de traitement et calcul, définition des paramètres, valeurs de paramètres, unités)• Possible agrégation des résultats (unité, valeurs résultantes, statistiques utilisés lors de l'agrégation, date de calcul...)• D'autres descriptions si besoin

Tableau 18 – Information contenue dans le rapport qualité proposé par l'ISO 19114

Après la sélection des éléments à inclure dans le rapport qualité, un aspect important est à traiter : leur génération et mise à jour. Typiquement, la génération et la mise à jour des rapports qualité dépendent fortement de la fréquence des observations et de la stabilité des caractéristiques sur la qualité des données. Ceci est effectué tout en gardant un équilibre entre les besoins d'information à jour et la génération excessive des rapports qualité.

Pour certaines applications dont la fréquence d'observation est plus élevée, les rapports qualité doivent être générés plus rapidement et mis à jour. Cependant, si les caractéristiques sur la qualité sont stables, l'utilisation des indicateurs qualité sera suffisante pour mettre à jour le rapport. Finalement, une autre possibilité consiste à fixer une fréquence moins élevée et qui concerne seulement les caractéristiques qui viennent d'être mises à jour. Généralement, ces rapports sont générés au format .html ou .pdf pour leur exploitation.

Dans les paragraphes suivants, nous présentons plus en détail l'adaptation de ces approches : les indicateurs et le rapport qualité, pour communiquer la qualité de données capteurs en temps réel à l'utilisateur et l'assister lors de la prise de décision.

5.3.2 Appréciation de la qualité des données capteurs

Généralement, les méthodes utilisées pour évaluer la qualité des données se basent sur les résultats des mesures de qualité et la documentation du processus d'évaluation. Ces méthodes fournissent l'information qui permet aux utilisateurs de contrôler systématiquement la qualité de données à n'importe quel niveau d'abstraction (selon le type d'application). Les résultats de l'évaluation de la qualité sont les sources principales pour n'importe quel type d'actions d'améliorations futures au sein du système.

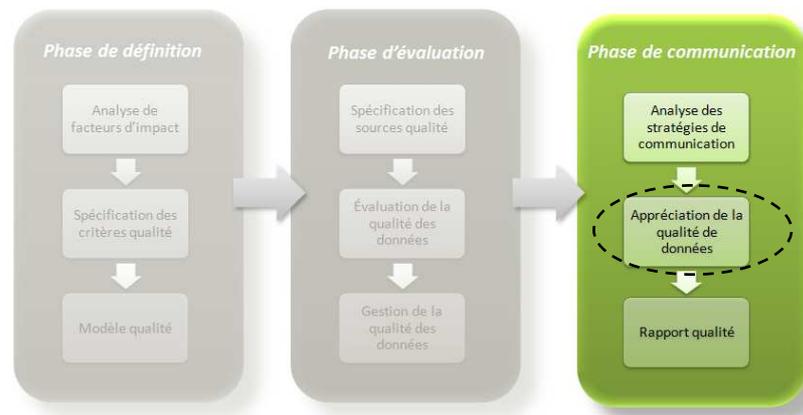


Figure 60 – Phase 3 – Appréciation de la qualité de données capteurs

Dans ce qui suit, nous décrirons plus en détail les aspects pris en considération pour le calcul et la représentation graphique des indicateurs qualité utiles pour la communication de la qualité des données capteurs (Figure 60).

5.3.2.1 Spécification des indicateurs qualité

Dans notre approche, les données capteurs sont évaluées selon huit critères fondamentaux (cf. Section 5.1.2). Ces critères sont représentés de façon générique par des indicateurs inclus aussi dans le rapport qualité.

Pour le calcul de ces indicateurs nous avons suivi les règles suivantes :

- **Règle 1.** Pour l'estimation des critères, nous avons sélectionné des échelles d'attribution de valeur qui vont nous aider à représenter de façon cohérente et uniforme la qualité de données. Cette échelle va permettre un classement plus générique du point de la comparaison pour le critère qualité.
- **Règle 2.** Les valeurs des indicateurs doivent être comprises dans l'intervalle $[0, 1]$, à l'exception de l'indicateur correspondant au critère complétude, qui est plus compréhensible en termes de pourcentage $[0, 100]\%$. Également, nous considérons dans tous les cas, deux valeurs supplémentaires $[-2, -1]$, telles que la valeur -2 représente que l'indicateur n'est pas applicable et -1 quand la valeur de l'indicateur n'est pas disponible. Ceci peut être présent dans le cas où l'utilisateur fait la sélection d'un critère qui ne correspond pas à l'étendue analysée.
- **Règle 3.** Afin de rendre les valeurs représentatives de la qualité plus claires à l'utilisateur, nous avons repris le principe des multiples travaux sur la qualité des données, dont l'indicateur prend la valeur plus proche de 0 comme une mauvaise qualité et près de 1 (ou 100%) comme l'idéal en termes de qualité.
- **Règle 4.** Évidemment, pour certains calculs des indicateurs, les valeurs résultantes peuvent être un peu ambiguës en utilisant ces échelles de représentation ; dans ces cas nous proposons l'ajout d'étiquettes textuelles auxiliaires, qui peuvent fournir aux utilisateurs les valeurs réelles des mesures ou le niveau de conformité des résultats.

Pour développer ces indicateurs il est donc nécessaire que les indicateurs soient représentatifs des critères qualité (choix de notre approche), que leur calcul soit bien établi, en concordance avec les critères qualité, et faciles à comprendre (Section 5.2.2). Egalement, il est important de souligner, que l'applicabilité et l'importance des divers indicateurs qualité et d'ailleurs les critères qualité, dépend directement des types de données et les besoins des utilisateurs.

Plus spécifiquement dans notre approche, l'exactitude peut être représentée par une échelle de $[0..1]$, telles que 0 = *inexacte* et 1 = *exacte*. Ce type de classement a été retenu en raison du processus de calcul, dont l'exactitude représente la valeur de probabilité d'erreur comprise normalement entre 0 et 1. Dans ce cas, l'évaluation de l'exactitude est typiquement assumée à être moins de 1 pour éviter une inexactitude considérable lors de la prise de décision.

La fiabilité, nous pouvons l'évaluer selon une échelle similaire à celle de l'exactitude, entre $[0..1]$ dont 0 = *non-fiabilité* et 1 = *fiabilité*. Cette spécification est indépendante des éléments pris en compte pour leur calcul. Typiquement ce résultat est associé à des indicateurs de la source de données (*MD_Capteur*).

La complétude, peut être évaluée comme le pourcentage dérivé entre les données envoyées et les données reçues au sein du système, ou dans un point d'accueil. Dans ce cas nous pouvons utiliser le pourcentage de complétude comme indicateur, voire entre $[0..100]\%$. Dont 0% = *incomplète* et 100% = *complète*.

Egalement, la fiabilité de la communication, peut être évaluée dans une échelle de $[0..1]$, avec 0 indiquant une communication « *non-fiabilité* » et 1 indiquant une communication *fiabilité*. Cette échelle nous permettra de créer des classes plus détaillées représentatives de la fiabilité de la communication. Nous proposons donc :

- $[0..0,2]$ *Non-fiabilité*
- $[0,3..0,5]$ *Peu fiabilité*
- $[0,6..0,8]$ *Fiabilité*
- $[0,9..1]$ *Très fiabilité*

L'actualité représente la mesure selon laquelle une donnée est à jour pour son exploitation. Dans ce cas, nous la normalisons avec 0 = *non-actuelle* et 1 = *actuelle*. Dans le même esprit, la volatilité qui mesure le changement de la valeur de la donnée, nous proposons l'échelle $[0..1]$, par exemple:

- 0 une volatilité *faible*,
- 0,5 une volatilité *moyenne*
- et 1 représentative d'une volatilité *haute*.

Finalement, l'interprétabilité est un critère plus particulier du domaine d'application, dans notre cas nous considérons intéressant de la mettre dans une échelle représentative de $[0..1]$, dont par exemple, 0 indique une interprétabilité *négative* et 1 une interprétabilité *positive*.

En définissant la façon avec laquelle ces indicateurs sont établis, nous pouvons définir la manière avec laquelle ces indicateurs seront générés. Certaines indicateurs qualité peuvent être produits à chaque nouvelle valeur exploitable, d'autres par

exemple, doivent être produits avec une fréquence prédéfinie, pendant de longues périodes de temps ou quand des variations importantes dans les valeurs sont perçues. En effet, la fréquence ou périodicité de la génération de ces indicateurs qualité dépend principalement de l'objectif des indicateurs (i.e. monitoring en continu) ou du type de surveillance à effectuer (i.e. surveillance environnementale, surveillance routière...). Concernant cet aspect, nous proposons un contexte qui met à disposition des procédures pour communiquer et évaluer la qualité des données, par exemple une liste exhaustive d'indicateurs à évaluer et visualiser en temps réel. De tels éléments sont souvent fournis dans des listes à cocher et des formulaires qui facilitent l'interaction avec l'utilisateur. Avec ces informations l'utilisateur sera capable de choisir les indicateurs qu'il considère pertinents pour caractériser la qualité de ses données.

Dans la section suivante, nous faisons une analyse entre ces spécifications pour la représentation des indicateurs optimales pour chaque critère qualité. Nous mettons en relation les propositions existantes pour la représentation des indicateurs qualité et les caractéristiques des critères qualité considérés dans notre approche.

5.3.2.2 Représentation visuelle des indicateurs qualité

Les applications liées à la surveillance de l'environnement nécessitent des outils visuels pour communiquer les informations variables dans le temps. Souvent, ces applications utilisent seulement des valeurs numériques ou des textes pour la décrire, notamment pour informer sur l'exactitude ou la précision d'une valeur. Cela dit, les systèmes de surveillance ont besoin de l'adaptation de techniques visuelles qui aident à l'utilisateur d'accéder plus facilement à l'information qualité grâce aux indicateurs. Dans ce cadre, l'idéal est de communiquer l'information en faisant en sorte que l'élément à surveiller soit visible et rappelable facilement. Pour ceci, nous considérons que les symboles ou images sont les plus utiles.

		Utile	Rappelable	Texte complémentaire	Compréhensible	Exacte
	Rectangle coloré	++	++	++	++	--
	Barre	++	++	++	++	++
	Slider	++	++	+-	++	++
	Segment d'un cercle	--	--	+-	+-	--
	Feu	++	++	++	++	--
	Graph variable	+	+-	++	+-	--
	Ligne	--	--	++	--	--
	Fleche	--	--	++	+-	--
	Symbole plus/moins	++	++	+-	++	--
	Cercle	--	+-	++	++	++
	Compteur	++	++	+-	++	++
	Tête	++	++	--	++	--
	Numérique	++	++	+-	+-	++

Tableau 19 – Tableau comparatif des symboles et l'utilité pour la description des critères qualité des données capteurs (inspiré de [Huth'07] et [Devillers'07])

Dans notre contexte, les huit critères qualité sélectionnés pour décrire la qualité des données capteurs doivent être associés à une visualisation pertinente pour l'utilisateur, voire un indicateur qualité sous une forme visuelle.

Dans le Tableau 19, nous essayons de faire un comparatif, inspiré par [Huth'07], afin de déterminer les figures ou symboles adéquats pour la représentation des indicateurs de chaque type de critère qualité sélectionné pour la caractérisation de la qualité de données capteurs dans un contexte de surveillance.

Dans la première colonne, nous indiquons si les symboles proposés dans le domaine géographique permettent la description de la qualité des données capteurs. Ensuite, nous analysons si ces symboles sont facilement rappelables par l'utilisateur. Dans la troisième colonne, nous évaluons si l'ajout de texte complémentaire aux symboles est possible. Egalement, nous analysons si ces symboles sont compréhensibles et exactes pour la représentation des critères qualité considérés.

De façon visuelle, l'évaluation d'un indicateur peut dépendre de la saturation d'une couleur ou d'une gamme de couleurs dépendant de l'échelle de description. Pour d'autres données, il peut s'agir des combinaisons de ces deux éléments pour montrer la variation dans les deux sens, comme c'est le cas de l'information avec une valeur variable dans le temps. Dans ce cas, nous pouvons faire une combinaison entre une saturation et une gamme de couleurs.

Comme nous le mettons en évidence dans ce tableau, les symboles qui ne possèdent pas une dégradation de la couleur, ne sont pas tout à fait exacts dans leur communication, c'est-à-dire, ils peuvent représenter une source d'incertitude à l'utilisateur : *Est-ce que le rouge signifie 0 ou 0.3 ?* . Dans ce type de symboles alors, il faut impérativement utiliser des informations textuelles qui décrivent la valeur qui est caractérisée par l'indicateur. Le texte complémentaire doit s'ajouter en cas de besoin, et si l'espace dédié à la communication le permet. En effet, trop d'information sur un espace réduit peut perturber également la perception de l'utilisateur. Le texte doit être donc, à côté de l'indicateur avec des mots clés qui transmettent à l'utilisateur l'état de l'indicateur. De longs textes ne sont pas tout à fait souhaitables.



Figure 61 – Symboles slider, plus/moins et compteur dans la complétude

Dans le cas du critère complétude par exemple, le symbole *slider, plus/moins ou compteur*, permettent de comparer un pourcentage dans une échelle colorée, ou en indiquant la relation positive de la valeur représentée (Figure 61).

Dans le cas de critères comme l'exactitude, la précision et la fiabilité, les symboles *feu, tête, plus/moins* permettent d'apercevoir de façon moins ambiguë l'exactitude ou la précision d'une valeur. Dans ce sens, pour utiliser ce type de symboles il faut impérativement définir des échelles de représentation adéquates à chaque

instance de couleur. Par exemple, la couleur verte ou le symbole *plus/moins* peuvent être utilisés pour représenter des valeurs exactes, précises ou fiables ; dans le cas contraire la couleur rouge ou le symbole *plus/moins* pourront représenter des valeurs inexactes, imprécises ou non-fiables (Figure 62).

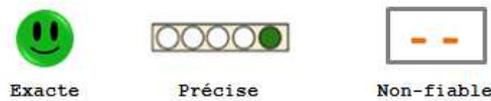


Figure 62 – Exemple de symboles pour l’exactitude, la précision et la fiabilité

Pour les critères liés au temps, notamment l’actualité, la volatilité et la ponctualité, la sélection de symboles est un peu différente car les indicateurs doivent montrer typiquement des valeurs numériques ajoutées, comme une date ou une fréquence. Dans ce cas, nous considérons que les symboles *slider*, *plus/moins* ou *numérique* peuvent communiquer de façon adéquate ce type de critères. Par exemple, dans le cas de l’actualité et la volatilité dont nous devons exprimer la dernière date de la mise à jour, nous pouvons utiliser un *slider* avec une fenêtre de temps fixe selon la mise à jour de l’application. Si ces critères sont mis à jour tous les 24 heures, chaque semaine ou mois, l’indicateur peut montrer l’échelle avec la valeur actuelle de la dernière mise à jour. On peut également représenter la date de la dernière mise à jour grâce au symbole numérique, en représentant la date à l’intérieur d’un rectangle coloré selon la perception de cet indicateur. Si la date est hors des limites d’actualité ou la ponctualité il sera visualisé, par exemple de plus en plus rouge (Figure 63).



Figure 63 – Exemple de symboles pour l’actualité et la ponctualité

Concernant les critères *volatilité et interprétabilité*, nous considérons qu’un symbole simple comme le *plus/moins* peut communiquer à l’utilisateur la variabilité dans les données exploitées et la facilité d’interprétation de cette information. Le fait d’utiliser une gamme de couleurs en parallèle pourrait gêner l’utilisateur. Dans ce cas, par exemple, si la variabilité est moyenne ou l’interprétabilité faible, la Figure 64 nous montre un exemple de sa représentation.



Figure 64 – Exemple du symbole indicateur pour le critère volatilité

En conclusion, nous pouvons considérer que l'ensemble d'indicateurs dédiés à communiquer la qualité des données dynamiques comme les données capteurs, est composé notamment par des gammes de couleurs et des textes brefs qui explicitent la qualité des données exploitées. L'aperçu que nous avons présenté dans cette section et la mise en relation des critères qualité des données capteurs vs les indicateurs visuels mettent en évidence, que des représentations simples via des symboles comme *le tête/feu*, *le plus/moins* sont adéquats pour ce type de critères. Cependant pour certains critères liés au temps, une meilleure perception sera possible si on considère une gamme variable de couleurs dans une fenêtre de valeurs qui permettent à l'utilisateur de se repérer plus facilement dans une échelle temporelle variable.

Cela-dit, il faut prendre en considération qu'une expression redondante de la qualité des données ou de l'information par couleur ou position (i.e. slider) peut engendrer des problèmes d'interprétation dus aux différences dans l'échelle de couleur. Dans ce cas, nous proposons d'utiliser une même échelle proportionnelle en couleur autant qu'en valeur.

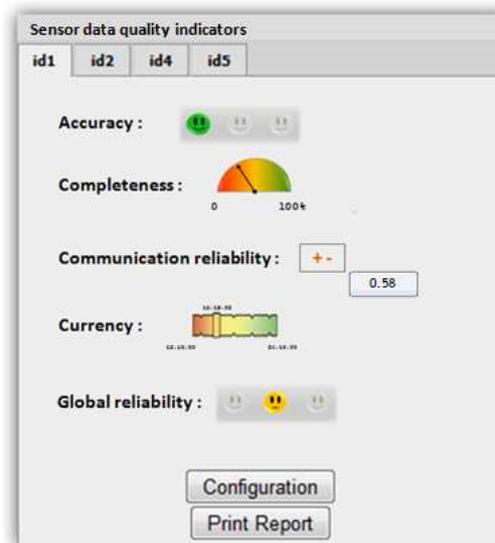


Figure 65 – Exemple de notre approche de représentation des indicateurs qualité.

Notre approche de visualisation de la qualité des données capteurs est donc basée sur la représentation des indicateurs correspondant aux huit critères qualité proposées. Ces indicateurs restent au choix de l'utilisateur dans sa visualisation via l'option « configuration ». Nous proposons la visualisation d'un indicateur pour chaque critère qualité de façon séparée comme le montre la Figure 65, ainsi d'un indicateur de fiabilité global qui permettra à l'utilisateur d'avoir un aperçu général sur la fiabilité des données exploitées lors d'une observation d'un ou plusieurs capteurs. L'information complémentaire dans les indicateurs est notamment affichée lors du passage de la souris sur les indicateurs et aussi au niveau des symboles, comme c'est le cas des légendes descriptives de symbole. Il est important de signaler que l'affichage des symboles peut être personnalisé selon les besoins de l'utilisateur. Par exemple, ajouter des légendes complémentaires, modifier les emplacements, échanger des symboles.

Dans ce paragraphe, nous avons introduit la visualisation de la qualité dans un système d'information exploitant des données capteurs. Cette proposition de visualisation est implémentée dans notre prototype de validation décrit dans le chapitre 6 de ce manuscrit.

Comme nous l'avons décrit au début de cette section, traditionnellement, associée à la communication de la qualité sous forme d'indicateurs (métadonnées), la communication de la qualité de données peut être réalisée grâce à un rapport de qualité généré selon les besoins qualité et exploité lors d'une observation quelconque. Dans le paragraphe suivant, nous détaillons les spécificités de ce rapport.

5.3.3 Spécification du rapport qualité

Dans l'évaluation de cette qualité, il existe un bon nombre de facteurs liés à la fiabilité des données capteurs ; il peut donc être intéressant par exemple, de communiquer le taux d'erreurs de chaque composante de la qualité quand la fiabilité de la communication est très faible. Dans cet esprit, des rapports qualité peuvent être générés de façon complémentaire à l'exploitation des indicateurs qualité en temps réel. Ensuite, nous décrivons les divers aspects pris en compte pour la définition d'un rapport qualité de données capteurs dans un contexte de surveillance (Figure 66).

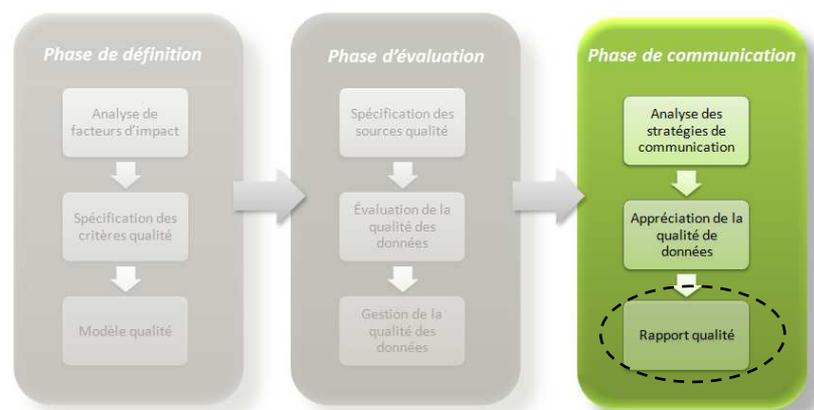


Figure 66 – Phase 3 – Spécification du rapport qualité

Les rapports qualité peuvent inclure les informations plus détaillées sur l'ensemble des critères mesurés, les valeurs de référence pour chaque mesure effectuée, les éléments observés, la temporalité des valeurs issues de ces mesures, ainsi qu'un résumé des réponses possibles aux requêtes prédéfinies qui aident l'utilisateur à mieux comprendre la qualité des données observées.

L'information contenue dans ces rapports doit être suffisante et doit être conservée de préférence de façon électronique pour des analyses futures. Nous considérons également, que les rapports doivent fournir l'information descriptive sur chaque source participant à l'évaluation (i.e. la généalogie de chaque capteur) afin de compléter le contexte de l'évaluation de la qualité des données.

Comme nous l'avons signalé dans la Section 5.3.1.2, le rapport qualité doit contenir l'information nécessaire pour décrire plus en détail le contexte de l'évaluation de la qualité des données issues de capteurs. A cause de l'hétérogénéité des besoins des utilisateurs en termes d'information descriptive de la qualité, nous considérons la spécification d'un rapport générique, qui englobe les informations de base pour la compréhension de la qualité des données, mais surtout qui sert comme point d'appui à l'utilisateur lors de l'évaluation de la qualité. Il est important de signaler que le contenu de ce rapport doit normalement être spécifié en collaboration avec les experts du domaine, qui peuvent définir avec une meilleure exactitude leurs besoins en termes d'information à exploiter. Dans notre approche, nous nous limitons à la proposition d'un rapport générique qui contienne essentiellement les informations relatives aux indicateurs qualité et l'information contextuelle de l'évaluation (Figure 67). Ceci permettra à l'utilisateur d'évaluer d'une façon plus claire la qualité des données exploitées et de garder une trace documentée sur l'évaluation des critères qualité et le contexte d'évaluation.

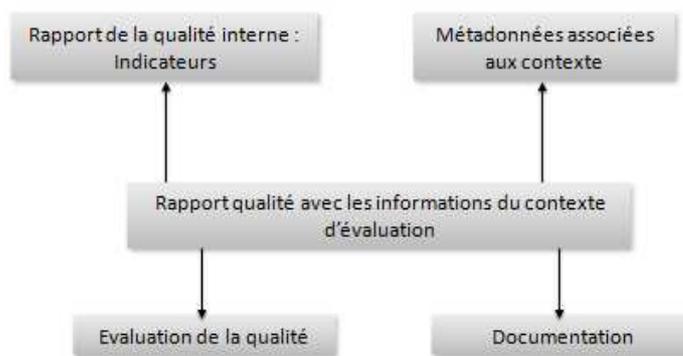


Figure 67 – Concept du rapport qualité au sein du traitement de données issues de capteurs

Nous considérons que le rapport qualité des données capteurs doit porter une description sur les mesures utilisées pour évaluer la qualité de ce type de données. Le contenu de ce rapport est en adéquation avec les informations proposées par l'ISO 19114, mais plus restrictif en termes de contenu.

Information contenue dans les rapport qualité de données issues de capteurs

- Identification du rapport (numéro, nom, date, périodicité du rapport)
- Information des responsables (nom, société, e-mail...)
- L'étendu observé (mesure, ensemble de mesures, flux de mesures d'un ou plusieurs capteurs, observation)
- La qualité globale aperçue et leur conformité selon les spécifications.
- Les erreurs de mesure et de traitement, concernant les critères qualité exploités dans les indicateurs et selon l'étendue observé.
- Détail de mesures utilisées (formule et valeurs résultantes)
- La généalogie de l'étendue observé (notamment les informations sur le contexte de l'évaluation)
- Possible agrégation des résultats (i.e. dans le cas que le rapport soit périodique), avec les unités, valeurs et statistiques utilisées.
- Dates importantes concernant la disponibilité, la ponctualité, l'actualité et la variabilité de l'information.

Tableau 20 – Proposition du contenu du rapport qualité orientée données issues de capteurs

Nous préférons communiquer l'information minimale pour comprendre l'application et ne pas surcharger l'utilisateur avec trop de détails sur l'information. Nous proposons donc une extension de ce rapport qui porte sur les éléments énoncées dans le Tableau 20.

Plus spécifiquement dans notre approche, due à la dynamique de l'environnement, ce rapport qualité est stocké indépendamment des données, car il peut être généré selon les préférences de l'utilisateur ou selon le comportement du système. Par exemple, un rapport peut être généré à chaque finalisation d'une observation où à la demande ponctuelle de l'utilisateur. Dans certains cas par exemple, quand le niveau de qualité est acceptable, il n'est pas forcément nécessaire de faire un rapport qualité de ces éléments. Dans d'autres cas, surtout lorsqu'il existe un nombre considérable de divergences entre les valeurs réelles et les valeurs de référence, il est important d'analyser et d'informer les éléments correspondant à l'utilisateur.

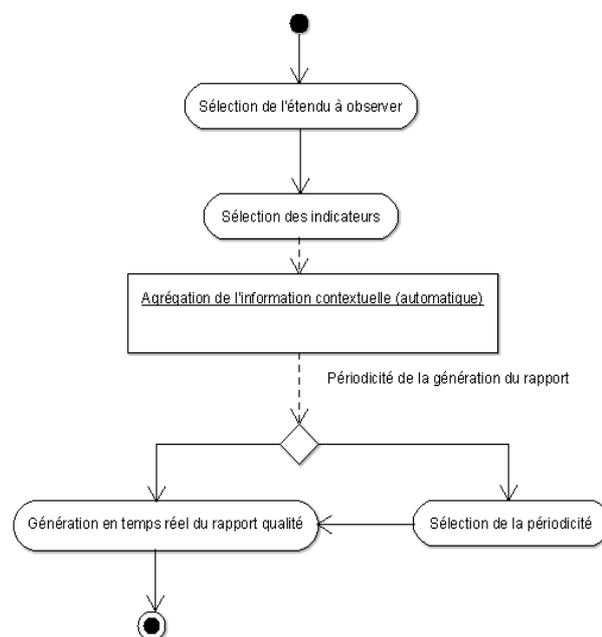


Figure 68 – Processus de génération du rapport qualité à la demande de l'utilisateur

Ainsi, pour ne pas entrer dans des problématiques de redondance d'information, dans notre approche, nous optons potentiellement pour la génération des rapports à la fin de chaque observation et à la demande ponctuelle ou périodique de l'utilisateur. Ce rapport se base donc, sur les indicateurs que l'utilisateur a jugé pertinents à l'instant t , en complément des informations contextuelles lors de l'évaluation.

Nous admettons qu'il existe un risque minimal en termes d'actualité de l'information qui peut surgir dans cette procédure, car la mise à jour des données se fait chaque fois que le rapport est généré et est basé sur les données actuelles. Ceci permet d'éviter la génération des rapports non-nécessaires. Sur ce même principe, des rapports périodiques ou historiques peuvent être générés selon les besoins de l'utilisateur (Figure 68).

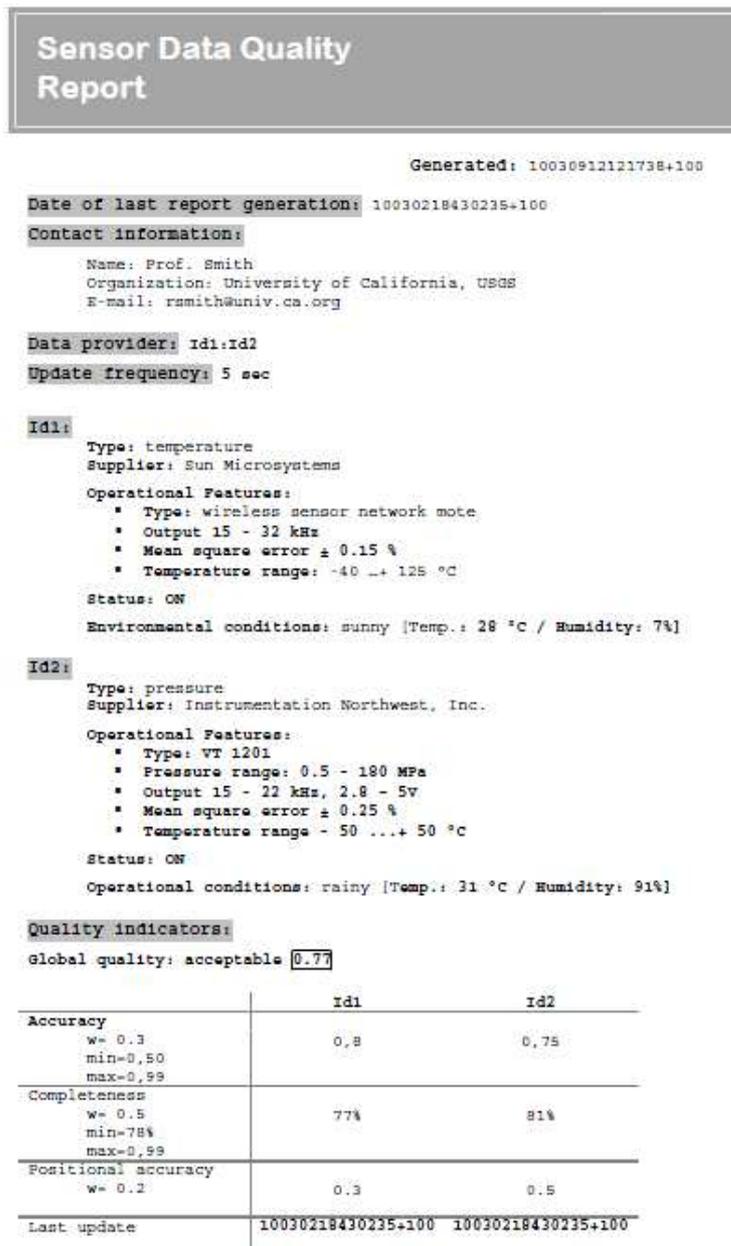


Figure 69 – Exemple du rapport qualité généré

Pour permettre à l'utilisateur une meilleure exploitation des informations qualité, nous proposons donc que ce rapport soit généré de façon numérique (document .pdf) à partir de l'interface de visualisation fournie à l'utilisateur, en parallèle de la visualisation des indicateurs en temps réel (Figure 69).

Pour cette génération nous avons utilisé dans notre exemple, l'outil libre Jasper Reports¹. Alors, les informations qualité et métadonnées associées sont extraites en temps réel en format .xml et couplées avec un template à l'aide de l'outil libre iReport².

¹ <http://jasperforge.org/projects/jasperreports>

² http://jasperforge.org/plugins/project/project_home.php?projectname=ireport

Les techniques de visualisation que nous utilisons dans notre approche, peuvent être utilisées dans des systèmes d'information ou interfaces web dans le domaine de la surveillance, notamment environnementale (i.e. plans, tables, graphes, etc.). Ces techniques permettent à l'utilisateur de naviguer dans la qualité des données capteurs à différents niveaux de détail, à travers des critères qualité représentés par des indicateurs et contenus dans des rapports qualités. Dans le chapitre suivant, nous décrivons plus en détail les aspects concernant l'implémentation de ces éléments au sein d'une interface utilisateur.

5.4 Conclusion

Dans ce chapitre nous avons présenté notre proposition de visualisation de l'information issue d'une observation et l'information qualité associée. Nous avons réalisé tout d'abord une étude sur les stratégies utilisées actuellement pour communiquer à l'utilisateur l'information qualité. Nous avons ensuite réalisé une comparaison sur leur possible application dans un domaine plus dynamique comme dans notre cas d'application. Ainsi, nous avons abouti à la proposition de l'adaptation des symboles graphiques dynamiques pour représenter de façon plus conviviale les indicateurs qualité, et l'utilisation d'un rapport de qualité pour résumer le contenu et résultats lorsqu'une évaluation est réalisée. Ce rapport est généré à la demande de l'utilisateur, ce qui permet de limiter la génération d'informations inutiles ou redondantes.

Grâce à notre approche, nous fournissons de façon graphique et textuelle l'information qualité obtenue de l'analyse des données issues de capteurs en temps réel. Utiliser ce type d'approche, cela permet de l'intégrer à n'importe quel système d'information géographique ou à des applications web traitant des données capteurs.

Dans le chapitre suivant, nous allons présenter l'implémentation de notre proposition de visualisation, ainsi que d'autres concepts fournis par notre travail de recherche.

Chapitre 6

MoSDaQ : Un prototype pour la visualisation de qualité des données dans les systèmes de surveillance environnementaux

Afin d'aider les utilisateurs dans l'évaluation de la qualité des données issues de capteurs, nous proposons un prototype d'interface web orienté utilisateur, qui permette la visualisation de données issues de capteurs et la communication des informations contextuelles à l'observation d'un phénomène quelconque, notamment sur la qualité des données exploitées.

Ce prototype permet de valider notre approche et implémenter les différents éléments qui le composent, notamment la sélection, le paramétrage, le calcul et la visualisation des indicateurs qualité, ainsi que l'exploitation des rapports qualité issus des évaluations effectuées. Ce prototype permet à l'utilisateur d'interagir avec l'information qualité à différents niveaux de détail selon le type d'observations qu'il effectue et les données qu'il veut analyser.

Ce chapitre est décrit de la façon suivante : dans la Section 6.1, nous présentons une description et architecture générale du prototype MoSDaQ (Monitoring Sensor Data Quality). La Section 6.2 introduit le cas d'application considéré pour le développement de ce prototype, la surveillance volcanique. Ensuite dans la Section 6.3, nous présentons les différents composants de cette interface et comment les utilisateurs peuvent l'exploiter. Finalement, nous concluons dans la Section 6.4.

6.1 Description du prototype

Notre prototype parte sur l'application d'un système de surveillance qui permet de caractériser l'environnement à l'aide de l'utilisation de capteurs, exécute le traitement d'information (l'acquisition, la surveillance, l'affichage) à travers l'utilisation d'interfaces, et la possibilité d'une prise de décision exécutée par les utilisateurs à distance (experts).

Plus spécifiquement, nous nous sommes focalisés sur les systèmes de surveillance environnementale, pour lesquels un ensemble de nœuds capteurs sont répartis dans l'environnement (géo-capteurs) et envoient leurs informations au centre de données (serveur central) et qui sont ensuite exploitées par l'utilisateur grâce à une interface web (Figure 70).

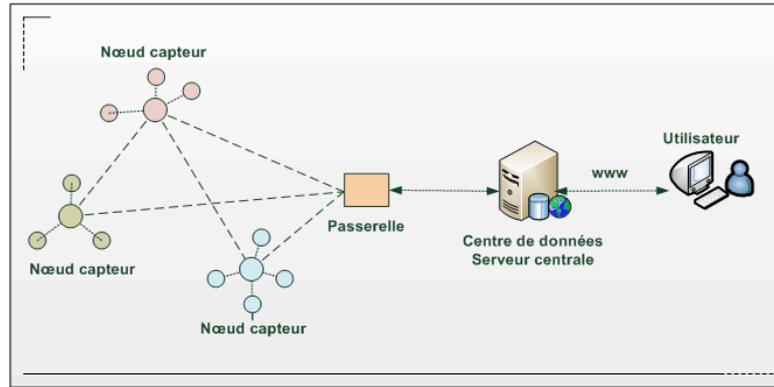


Figure 70 – Système de surveillance environnemental.

Dans un système de ce type, nous permettons l'accès et l'interrogation des données issues de capteurs en temps réel aussi qu'à leur qualité via le web. Dans notre proposition, nous fournissons des indicateurs qualité sous forme de symboles couplés avec une carte géographique, qui représente les divers capteurs éparpillés dans la zone d'observation. Ce système permet également à l'utilisateur de sélectionner les indicateurs les plus pertinents, les paramétrer et les appliquer sur les éléments de l'observation (mesures, flux, campagnes...). De cette façon nous contribuons à faciliter l'accès, la manipulation et la visualisation des informations qualité via les indicateurs et les rapports qualité.

Notre prototype (Figure 71) utilise la technologie Java pour le développement d'une application web dynamique. Plus spécifiquement nous avons utilisé AJAX pour le développement des pages dynamiques et l'application web du coté client (XHTML, CSS, JavaScript et XML) et la structure JSF (JavaServer Faces) pour le développement de l'application web coté serveur. Egalement, nous avons déployé ce prototype dans le serveur d'application GlassFish v. 3.0 et nous avons utilisé le système de gestion de base de données classique MySQL 5.4.1.

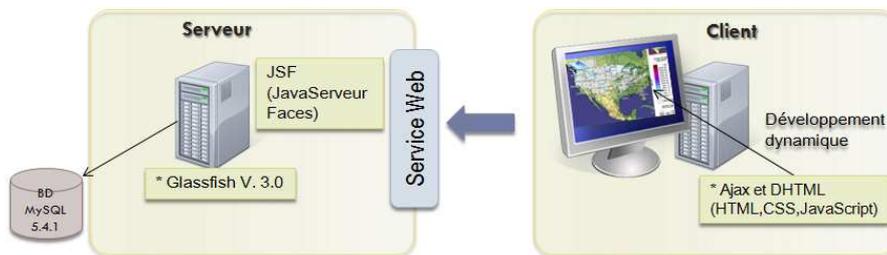


Figure 71 - Architecture générale du prototype MoSDaQ

Pour l'implémentation des indicateurs qualité nous restons dans le XHTML pour leur codification, SVG (Scalable Vector Graphics) comme format d'image et JavaScript pour assurer l'interaction entre les éléments. Pour la génération du rapport qualité nous avons utilisé la structure java JasperReports et iReport pour la partie conceptuelle du rapport.

L'interface web que nous proposons et qui peut être développée également pour être exploitée au sein des systèmes d'information géographiques, intègre le service Google Maps pour cartographier des objets géolocalisés dans un espace géographique et des technologies de bases de données pour la gestion d'informations proposés dans notre approche (Figure 71). Le résultat de cette implémentation est une application web avec des capacités de visualisation des informations en temps réel (données et métadonnées). Les principales technologies utilisées sont:

- **Glassfish V. 3.0** : est un serveur d'applications Open Source Java EE 5. GlassFish est une implémentation complète qui recouvre notamment: JPA (Java Persistence API), JSF (Java Server Faces), JSP et Servlet pour faire de l'injection de dépendance dans le serveur web. Nous l'avons utilisé pour déployer notre application.
- **AJAX** : est l'Asynchronous JavaScript and XML utilisé pour le développement de pages dynamiques et d'applications Web. Ajax évoque l'utilisation conjointe d'un ensemble de technologies libres couramment utilisées sur le Web : HTML (ou XHTML), CSS, JavaScript, XML.
- **MySQL 5.4.1** : système de gestion de bases de données que nous avons utilisé pour stocker les données historiques issues de capteurs, ainsi que les informations complémentaires (métadonnées), indicateurs qualité, etc.
- **JavaServer Faces (JSF)** : nous avons utilisé ce framework Java pour le développement de l'application web.
- **JasperReports** : cet outil d'exploitation libre se base sur des fichiers XML pour la présentation des états. Il peut être couplé à iReport (outil WYSIWYG - What You See Is What You Get) pour faciliter sa mise en œuvre dans une application Java, classique ou orientée web.
- **Google Maps API** : permet d'intégrer Google Maps dans un site web particulier avec nos propres données géolocalisées. C'est un service gratuit, mais les termes de services sont réservés par des droits. Pour utiliser ce service, une clé API est nécessaire, laquelle est liée au service web et à l'annuaire quand la clé est créée. Le fait de créer une interface d'un plan personnalisée, nécessite l'ajout du code JavaScript Google pour une page, et ensuite des fonctions Javascript pour ajouter des points sur la carte.

L'information utilisée pour l'expérimentation dans ce prototype est simulée et basée sur les différents standards ISO orientés information géographique, notamment l'ISO 19113, l'ISO 19114, l'ISO 19115 et l'ISO 19138, ainsi que ceux proposées par l'OGC orientés SensorWeb Enablement [SWE'09]. Nous avons décidé de coupler les informations représentées par ces deux importants standards afin de rendre le prototype beaucoup plus dynamique que dans les applications traditionnelles dans le domaine géographique. Nous avons utilisé comme cas d'application un système de surveillance volcanique.

Dans la suite, nous présentons plus en détails les spécificités de notre cas d'application pour ce prototype.

6.2 Cas d'utilisation

L'application à laquelle nous nous sommes intéressée est la surveillance et contrôle des phénomènes naturels, et plus particulièrement la surveillance volcanique. Ce type de systèmes de surveillance est de grande importance au niveau environnemental, urbain et social. Le principal objectif de ce type d'applications est d'effectuer une détection opportune des perturbations au sein de l'environnement pour éviter un désastre ou réduire ses effets. De même, ce type d'applications a aussi des objectifs secondaires comme la détection et le contrôle des changements autour du phénomène. Par exemple, observer le phénomène au cours du temps permet de déterminer et de construire des modèles de comportements selon certaines circonstances, et ainsi d'utiliser ces informations dans des cas préventifs pour prendre des décisions plus adéquates au à l'évolution et contexte du phénomène.

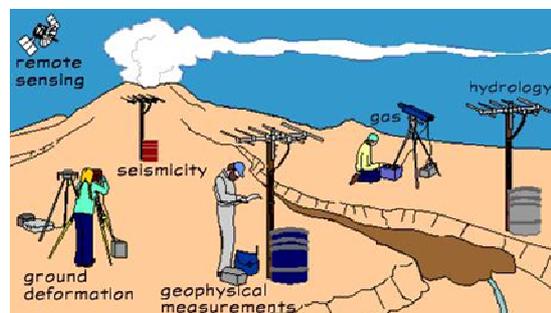


Figure 72 – Surveillance volcanique

Plus concrètement, les applications utilisées pour contrôler et surveiller un volcan (ou tout autre phénomène de ce type) doit inclure la détection des activités et conditions alarmantes qui peuvent accroître le risque d'une éruption par exemple ; la détection et la localisation d'une expulsion de gaz, ou le suivi de l'accroissement de la température dans une zone du volcan, etc. (Figure 72).

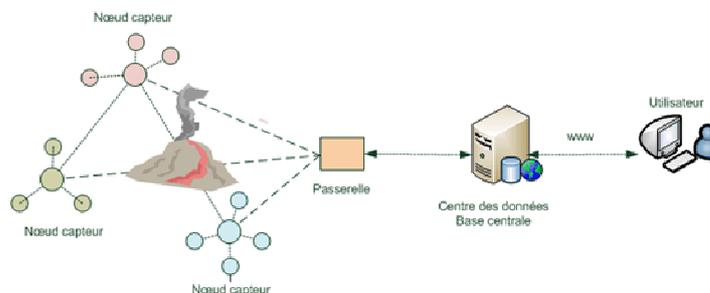


Figure 73 – Architecture physique d'un système de surveillance volcanique

Les données collectées par ces réseaux, sont appelées données observation. Ces données sont collectées par les nœuds capteurs et concernent des éléments comme l'humidité, la température, etc. Même si généralement un prétraitement est réalisé au sein des nœuds afin de filtrer et de réduire la quantité des données à gérer au sein de la mémoire principale, le traitement des données issues de capteurs et leur information complémentaire (métadonnées) est effectué dans un serveur et un système de gestion de base de données. Ces données collectées seront donc mises à disposition à des utilisateurs à distance via le web (Figure 73).

Nous prenons comme exemple dans notre cas d'application, un réseau de capteurs hétérogènes utilisé pour la surveillance du volcan Popocatépetl, au Mexique. Le volcan Popocatépetl est un volcan actif avec 13 éruptions enregistrées dans les 500 dernières années. Près de 20 millions de personnes habitent dans un rayon d'environ 80 kms au tour du volcan (Figure 74).

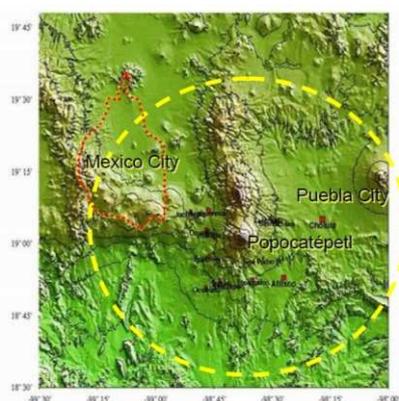


Figure 74 – Plan cartographique du volcan Popocatépetl et ses environs [Cenapred'06]

A 60 km du volcan, la station d'acquisition et de traitement des données CENAPRED (Centre National pour la Prévention des Désastres) [Cenapred'06], traite environ 64 signaux de télémétrie avec 16 ordinateurs dédiés à l'analyse et aux systèmes d'alertes et de communications surveillant de façon visuelle, sismique, géodésique et géochimique (Figure 75) le comportement du volcan.



Figure 75 – Surveillance du Cenapred au volcan Popocatepetl [Quaas'06]

Le CENAPRED travaille en collaboration avec l’Institut de Géophysique¹ et de Génie² de l’Université Nationale Autonome de Mexique (UNAM)³ et l’U.S. Geological Survey⁴.

Les données acquises par le réseau de surveillance du volcan (un système d’alertes) sont hétérogènes, résultant de divers capteurs : 9 sismomètres, 4 clinomètres, 3 détecteurs de flux, une caméra de haute résolution, une caméra infrarouge, un EDM (Electronic Distant Meter) contrôlés à distance avec quatre points de mesure, et des autres capteurs mesurant et analysant chimiquement des résidus de l’eau et des cendres, ainsi que des images satellites, etc., sont envoyées à la station centrale.



Figure 76 – Image du volcan Popocatepetl

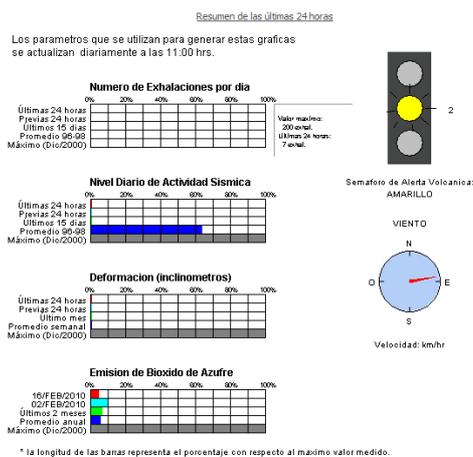


Figure 77 – Rapport de l’état de la surveillance

Comme résultat de la surveillance, ce projet communique via le web, dans un site internet dédié à la diffusion [Cenapred’06] deux aspects importants issus des observations : des images actuelles du volcan et un rapport historique (de deux mois à 24 heures) sur les différentes mesures utilisées pour déterminer l’état du volcan. Les Figures 76 et 77 nous montrent les exemples de ces aspects.

En prenant cet exemple, nous observons que l’intérêt de l’utilisateur dans ce type de systèmes considère les données d’observation obtenues grâce aux capteurs et aussi l’information qualité correspondant à ces données. Des requêtes dans le domaine spatial, temporel, spatio-temporel du type historique et en continu sont possibles.

Plus spécifiquement dans ce type de système, l’utilisateur est intéressé par la découverte des informations qui concernent principalement l’évolution du phénomène au cours du temps ainsi que l’état actuel du phénomène en question. Pour que l’utilisateur puisse analyser le phénomène, nous considérons que les requêtes possibles doivent prendre en considération l’échelle du temps. Ceci permettra de distinguer entre une analyse actuelle d’une analyse historique du phénomène.

¹ <http://www.geofisica.unam.mx/indexi.php>
² <http://www.ingen.unam.mx/default.aspx>
³ <http://www.unam.mx/index/fr>
⁴ <http://vulcan.wr.usgs.gov/home.html>

Pour évaluer l'état actuel du phénomène observé, nous considérons deux types d'interrogations : des requêtes ou interrogations en continu (à partir de l'instant t jusqu'à l'infini t_{∞}) et des requêtes « one-shot » [Gripay'09] que nous interprétons comme ponctuelles et qui nous permettent d'évaluer un paramètre ou état à l'instant actuel t_{now} . Pour analyser l'évolution du phénomène nous considérons un ensemble de requêtes traditionnelles liées au domaine du temps entre un instant du temps passé t_{-} jusqu'à l'instant du temps actuel t_{now} .

Ces deux types d'interrogations classifiées selon la temporalité des objets nous les combinons avec la dimension spatiale pour retrouver les objets présents dans notre plan d'observation spatiale à différents instants ou périodes du temps [Mokbel'05].

Dans ce cadre, nous considérons les requêtes suivantes comme possibles et auxquelles notre système essaye de répondre :

a) Requêtes spatio-temporelles ponctuelles et en continu

Dans le cadre d'une surveillance volcanique, l'utilisateur est intéressé par l'état actuel des éléments observés, l'état du phénomène et les informations contextuelles de telle observation. Pour ceci nous prenons quelques requêtes exemples :

Q1 : La sélection des capteurs localisés dans la zone d'observation à l'instant t_{now}

Q2 : La sélection de la température d'un capteur à une position donnée P à l'instant t_{now} .

Q3 : Informer en continu les mesures des capteurs dans la zone d'observation.

Q4 : Informer en continu les diverses mesures dans la zone d'observation.

A partir de ces requêtes exemples, nous pourrions les coupler au besoin en termes d'information complémentaire, notamment sur la qualité des données exploitées. Dans ce cadre, nous pouvons de plus considérer les requêtes suivantes :

Q1' : La sélection des capteurs localisés dans la zone d'observation à l'instant t_{now} avec une fiabilité de communication supérieure à 0.5

Q2' : La sélection de la température d'un capteur à une position donnée P à l'instant t_{now} , avec une précision spatiale entre 0.5 et 0.9.

Q3' : Informer en continu les mesures des capteurs actifs dans la zone d'observation.

Q4' : Informer en continu les diverses mesures dans la zone d'observation avec une complétude supérieur à 70%.

Pour cette analyse en continu et ponctuelle, nous considérons important de communiquer l'information qualité en temps réel de façon à ce que l'utilisateur puisse l'exploiter sans une demande spécifique et selon les données qu'il exploite. Cette information permettra à l'utilisateur de prendre une décision plus facilement. Nous expliquerons plus en détail cet aspect dans la description de notre interface (Section 6.4).

b) Requêtes spatio-temporelles historiques

Dans le cadre d'une surveillance volcanique ou d'un autre phénomène naturel l'utilisateur est également intéressé par la comparaison et l'analyse des données historiques obtenues dans une observation donnée. Ceci permet aux experts d'analyser le comportement du phénomène lors de périodes du temps et de détecter les possibles anomalies. Nous citons quelques exemples de requêtes historiques dans le domaine spatio-temporel :

Q5 : La sélection des capteurs localisés dans la zone d'observation $Zone_1$ les dernières 24 heures.

Q6: L'ensemble de mesures effectuées par le capteur $S1$ les trois derniers jours.

Q7 : Le mouvement de terrain plus élevé dans la $Zone_2$ enregistré entre t_1 et t_2 .

L'information contextuelle notamment sur la qualité des données peut être jointe dans les demandes de l'utilisateur sur les données historiques.

Q5' : La sélection des capteurs éparpillés dans la zone d'observation $Zone_1$ les dernières 24 heures avec une fiabilité de communication entre 0.7 et 0.9.

Q6' : L'ensemble de mesures effectuées par le capteur $S1$ les trois derniers jours avec une ponctualité de plus de 0.8.

Q7' : Le mouvement de terrain plus élevé dans la $Zone_2$ enregistré entre t_1 et t_2 obtenu par les capteurs avec une précision spatiale de plus de 0.8.

Pour réaliser cette étude plutôt historique, nous considérons que ce type d'information sera fourni à l'utilisateur à la demande et selon leurs besoins, afin de ne pas surcharger le système et produire des informations non exploitables. Nous considérons que certaines de ces requêtes peuvent être fixées et cataloguées comme périodiques, comme c'est le cas du système de surveillance volcanique du CENAPRED (Figure 77). Ces requêtes périodiques, notamment appuyées sur des données historiques permettent la constitution des rapports de phénomènes. Ceci peut être très bien adapté à notre système. Cependant, nous restons dans le cadre de requêtes ponctuelles à la demande de l'utilisateur sur des données historiques et sur des requêtes de base qui seront exécutés en continu pour surveiller l'état du phénomène et des éléments qui le mesurent.

Dans la section suivante, nous présentons l'ensemble des éléments que nous avons conçus pour achever notre approche orientée définition de la qualité de données capteurs.

6.3 Visualisation des données d'observation et de l'information qualité

En se reposant sur notre approche de traitement et de visualisation des données issues de capteurs et leur qualité, ce prototype permet aux utilisateurs d'enrichir leur connaissance à travers un outil d'exploitation visuel via le web. Comme nous l'avons décrit précédemment, le fait de communiquer l'information à différents niveaux de

détail, permet aux utilisateurs d'analyser et d'exploiter leurs données dans un environnement de surveillance. La Figure 78 présente l'aperçu global de cette interface appelé web MoSDaQ (Monitoring Sensor Data and Quality).

Cette interface est composée par cinq sections principales : section de cartographie d'objets géolocalisés, section d'information observation, section d'information capteurs, section d'interrogation, section indicateurs qualité (conditions environnementales et indicateurs qualité de données).



Figure 78 - Aperçu du prototype d'interface web MoSDaQ

Les caractéristiques et composition de ces sections seront décrites plus en détails dans les paragraphes suivants.

I. Section de cartographie des objets géolocalisés

La première section répond à la cartographie des objets géolocalisés dans l'espace, ce que nous appelons « zone d'observation » (Figure 79).

Dans cette section nous utilisons le service disponible de Google Maps API, qui nous permet d'intégrer des objets géolocalisés dans une zone géographique. Cette zone est délimitée par un point représenté en coordonnées qui sera pris comme le point central de la carte.



Figure 79 – Section de mapping des objets géolocalisés (capteurs) au sein du volcan

Nous avons décidée de fixer le périmètre afin de permettre un affichage plus clair et une bonne résolution. Ce service nous permet d'avoir différents types d'affichage : satellite (Figure 79), de type cartographique (Figure 80) et de type terrain (Figure 81).



Figure 80 – Affichage du type cartographique



Figure 81 – Affichage du type terrain

Comme le montre cet ensemble de Figures (72,73 et 74) nous avons placé les différents objets géolocalisés, des capteurs, indiqués par un symbole graphique. Nous avons utilisé un symbole particulier pour chaque type de capteur, et quelques autres symboles pour indiquer l'état du capteur. Voici quelques exemples :



Figure 82 – Symboles utilisés pour représenter les capteurs et signaux d’alerte au sein des capteurs

Pour chacun des capteurs représentés dans la zone d’observation nous ajoutons les informations qui les décrivent de façon à ce que l’utilisateur puisse cliquer sur le symbole et que les informations soient affichées dans une fenêtre informative (Figure 82).

Cette fenêtre informative contient les principales informations identifiant et décrivant le capteur. Nous utilisons dans l’entête l’identification du capteur qui peut être un nom ou numéro d’identificateur (ID). Cet identificateur est unique pour chaque capteur.

Ensuite, nous avons l’ensemble d’informations descriptives et informatives qui correspondent notamment aux caractéristiques particulières du capteur. Nous commençons par une courte description du fonctionnement du capteur (i.e. capteur de température), la position avec une latitude et longitude, le nom du fournisseur de l’appareil, et la date d’installation du capteur (Figure 83).



Figure 83 – Information sur le capteur sélectionné en temps réel

Ces informations restent fixes tout au long de la vie active du capteur et sont insérées au moment de l’installation du capteur dans la zone d’observation. Ensuite, nous avons un ensemble d’informations descriptives qui changent au cours du temps et qui décrivent l’état actuel du capteur, par exemple : le statut du capteur (actif : on / inactif : off), le niveau de batterie que le capteur informe (0-100%), ainsi que des informations relatives à la dernière prise de mesure effectuée : valeur prise, date et heure de la prise, et exactitude de la mesure prise.

Finalement, dans cette fenêtre informative nous proposons une fonction « Graph » qui permet de voir sous forme graphique l'évolution de la température en temps réel (Figure 84).

L'ensemble de ces valeurs est mis à jour selon son évolution au cours du temps mais actualisé au sein de l'interface toutes les 10 secondes. Ce paramètre est variable selon les besoins d'affichage et selon la dynamique des données.

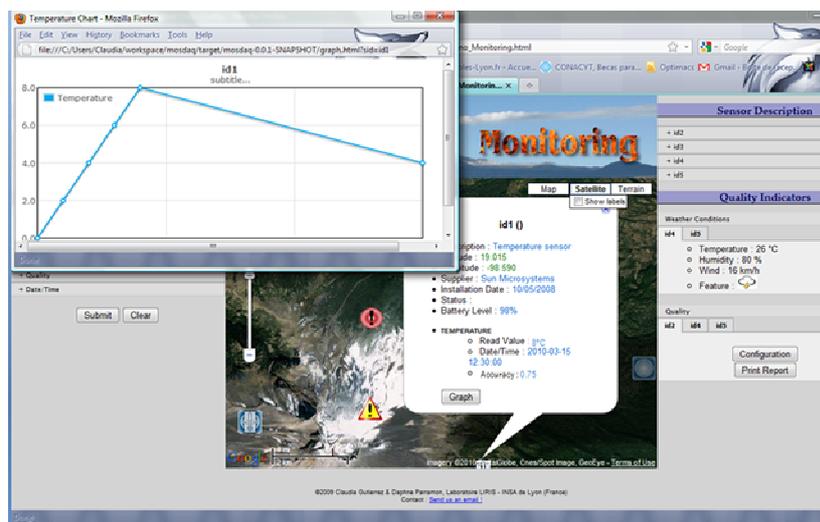


Figure 84 – Exemple de représentation en temps réel de la valeur mesurée du capteur de température

II. Section d'information des observations

L'objectif de cette section est de présenter les informations relatives aux phénomènes et éléments observés.



Figure 85 – Section d'information observation

Cette section (Figure 85) est composée de quatre aspects importants qui décrivent le type d'observation réalisé. Nous débutons par l'information correspondant au phénomène observé (i.e. volcan), ensuite nous additionnons le type de caractéristique évaluée sur le phénomène (i.e. activité volcanique).

Nous dédions également un espace pour citer les éléments mesurés pour évaluer le phénomène (i.e. température, séisme, ...) et finalement, nous fournissons les coordonnées et limites géographiques de la zone d'observation (latitude, longitude...).

Les informations dans cette section sont principalement extraites des caractéristiques de la zone d'observation. Les éléments mesurés et affichés dans cette section dépendent des capteurs qui sont présents dans la zone d'observation. La position de la zone d'observation peut être changée, zoomée en avant et en arrière, selon les besoins de l'utilisateur. Chaque fois que l'utilisateur fait une manipulation de ce type, les valeurs dans la position de la zone d'observation seront automatiquement mises à jour, ainsi que les éléments et caractéristiques qui la composent.

III. Section d'information capteurs

Cette section comprend les informations relatives aux capteurs (métadonnées capteurs statiques) qui permettent à l'utilisateur d'identifier leur équipement déployé sur la zone d'observation.

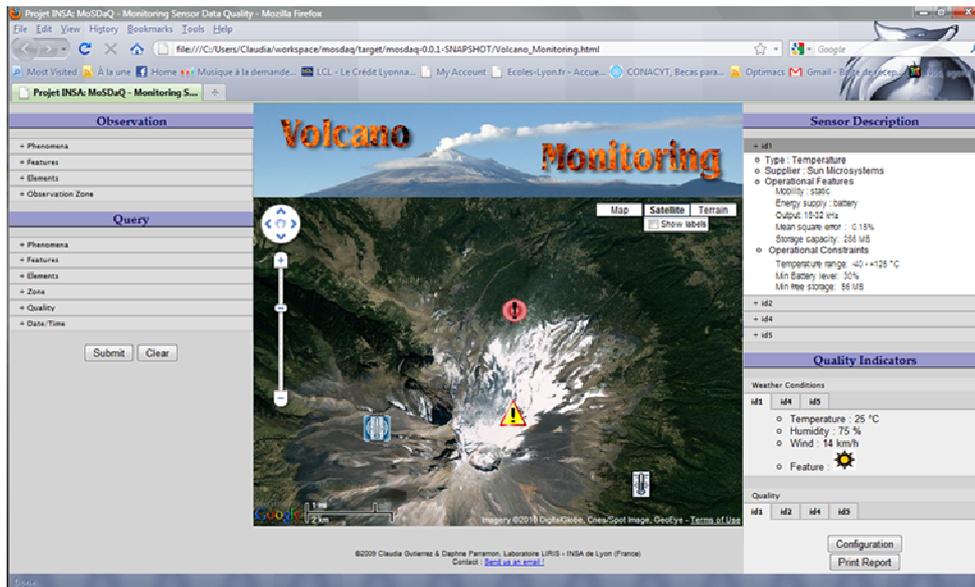


Figure 86 – Section Information des capteurs

Plus spécifiquement, ces informations incluent : le type de capteur, le fournisseur, les caractéristiques opérationnelles, les contraintes opérationnelles du capteur (Figure 86).

Ces informations sont données notamment par le fournisseur du capteur et l'expert qui réalise le déploiement. Ces informations sont mises à jour chaque fois que le capteur est installé ou chaque fois qu'il est mis en route après une calibration.

Ces informations sont également mises à disposition au sein de l'interface en fonction des capteurs qui sont affichés dans la zone d'observation. Nous présentons cet affichage de façon dynamique afin de présenter de façon plus claire et organisée les informations.

IV. Section des indicateurs qualité

Le contenu de cette section, nous l'avons divisé en deux sous-sections, une première indiquant les conditions environnementales présentes dans chaque localisation du capteur et deuxième qui correspond aux indicateurs de qualité de données exploitées.

Dans la première sous-section, nous fournissons des indicateurs sur les conditions environnementales grâce à des stations météorologiques déployées sur place et qui nous indiquent l'état de l'environnement du capteur. Pour cette section nous considérons trois éléments importants pour décrire l'environnement du capteur : la température, la pression atmosphérique et l'humidité. Ces éléments peuvent toujours être modifiables selon les besoins des experts.

Nous proposons pour cette sous-section, l'utilisation des indicateurs graphiques qui informent l'utilisateur de façon plus rapide sur l'état de l'environnement. Dans la Figure suivante nous montrons quelques exemples (Figure 87).

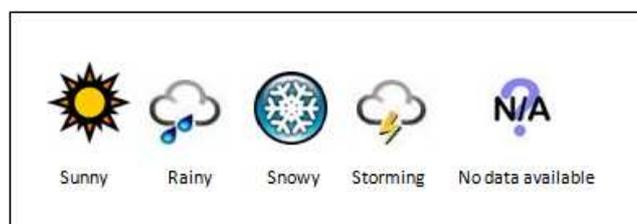


Figure 87 – Indicateurs de conditions environnementales

L'affichage de ces indicateurs d'environnement est réalisé également en considérant les capteurs qui sont déployés sur la zone d'observation. Cet affichage est donc dynamique et fonction des capteurs sélectionnés pour l'observation.

Dans la deuxième sous-section (Figure 88), nous communiquons les indicateurs qualité correspondant aux données issues de capteurs utilisés dans l'observation. Ici, grâce à la fonction « configuration », nous permettons à l'utilisateur de faire la sélection des critères qu'il veut évaluer et représenter de façon graphique pour leur exploitation. Comme dans les sections précédentes, les capteurs existant dans la zone d'observation sont automatiquement affichés dans la section d'indicateurs sur la qualité de données. Dans la suite, l'utilisateur devra procéder à la configuration des critères qu'il veut évaluer.



Figure 88 – Section d'indicateurs sur la qualité des données issues de capteurs

Au moment de la sélection, l'utilisateur peut aussi paramétrer ses choix en termes de valeurs maximales et minimales et le facteur de pondération de chacun des critères sur lesquels il s'appuie pour calculer la qualité globale (Figure 89). Selon ces valeurs, une qualité globale peut être ensuite calculée sur les données issues d'un seul ou de plusieurs capteurs, c'est-à-dire, la qualité des données issues d'une observation faite par un ou plusieurs capteurs.

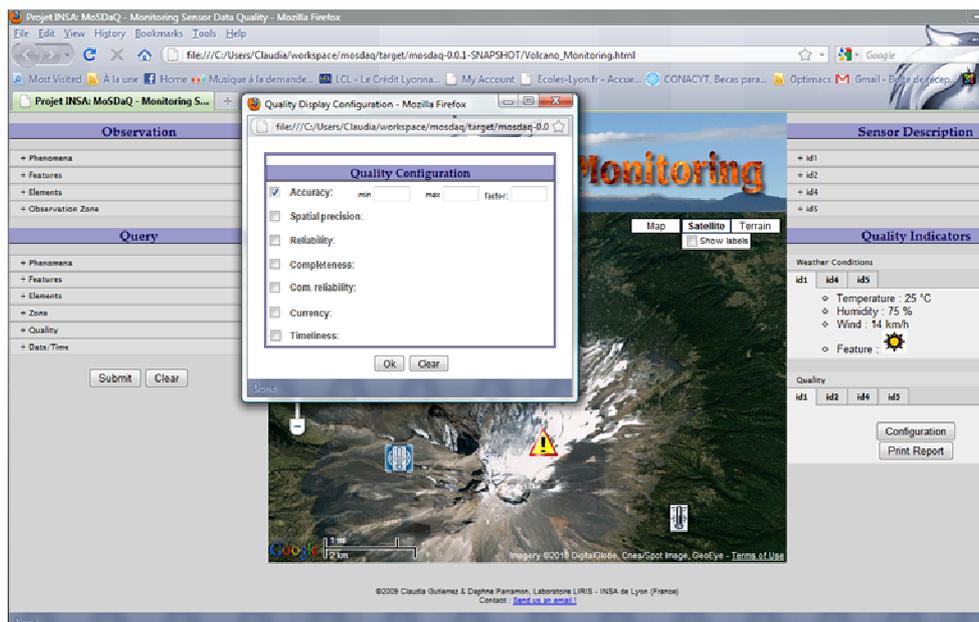


Figure 89 – Configuration des indicateurs qualité

L'utilisateur est toujours assisté lors de l'insertion des valeurs, notamment dans cette section. Des messages d'alerte sont générés chaque fois qu'il saisit une valeur hors des limites prédéfinies.

Dans notre prototype, nous restons sur le calcul en temps réel de la qualité des données dans les niveaux de granularité suivantes : une donnée ou flux de données issues d'un capteur ou de plusieurs capteurs. Nous considérons que les calculs plus ponctuelles ou sur un ensemble spécifique de données historiques seront faits a posteriori. Nous ne nous attachons pas à cet aspect lors du développement de ce prototype.

Suite à cette configuration, les indicateurs qualité avec leur valeur actuelle seront affichés dans la section dédiée à l'information qualité (Figure 90). L'utilisateur peut toujours reconfigurer ces indicateurs selon ses besoins.

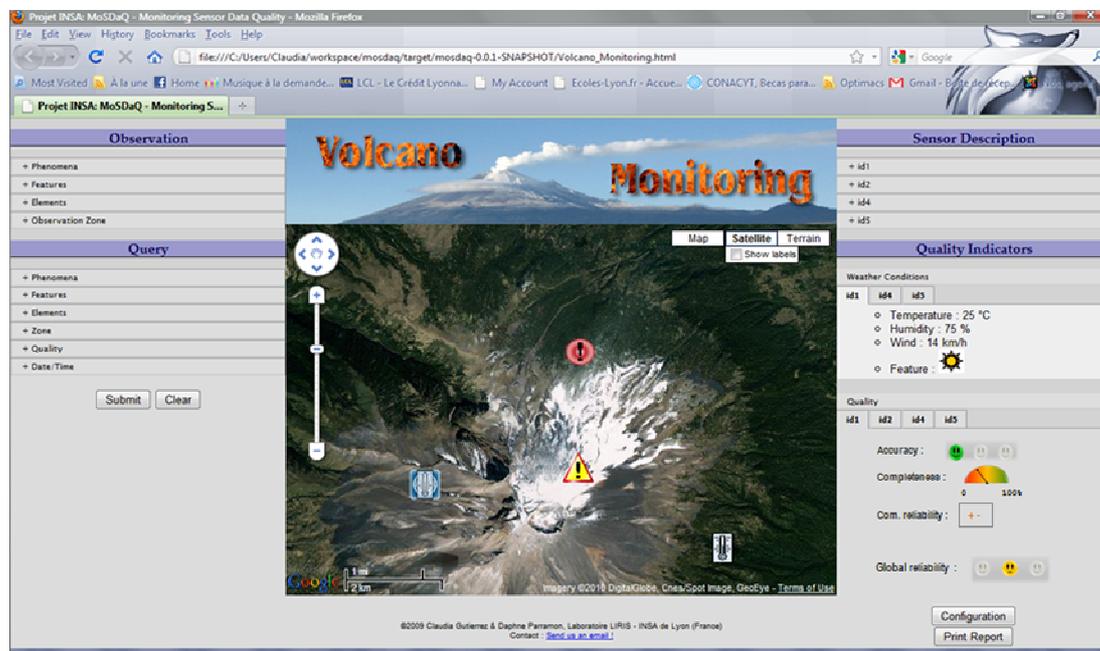


Figure 90 – Indicateurs sur la qualité de données issues de capteurs

Comme nous l'avons exprimé dans la Section 5.3.3, ces indicateurs sont calculés selon la variation de la valeur mesurée et affichés avec le même taux de rafraichissement prévu pour toute la page web ~10 seconds. Rappelons que ce paramètre est toujours adaptable aux besoins de l'utilisateur ou de l'application elle-même.

Finalement concernant cette section, nous fournissons à l'utilisateur le moyen de garder une trace ponctuelle de l'évaluation qu'il effectue sur l'ensemble des données exploitées et l'information contextuelle qui l'accompagne. Avec la fonction « imprimer rapport » (*print report*), l'utilisateur peut générer le rapport sur l'évaluation de la qualité des données. L'ensemble d'informations exploitées dans l'observation est extrait sous forme de graphe (xml) et utilisé pour générer un rapport qualité (Figure 91).

Ce rapport est généré avec un template prédéfini, lequel peut toujours être modifiable en contenu et structure. Nous proposons ce format pour sa simplicité et adaptabilité à l'agrégation de nouveau contenu. Suite à la génération de ce rapport, l'utilisateur aura la possibilité de l'enregistrer de façon numérique sur disque ou mémoire flash. Le rapport généré n'est pas gardé en mémoire avec le reste de l'information. Il représente la possibilité de garder des traces des évaluations ponctuelles à un moment ou événement donné.

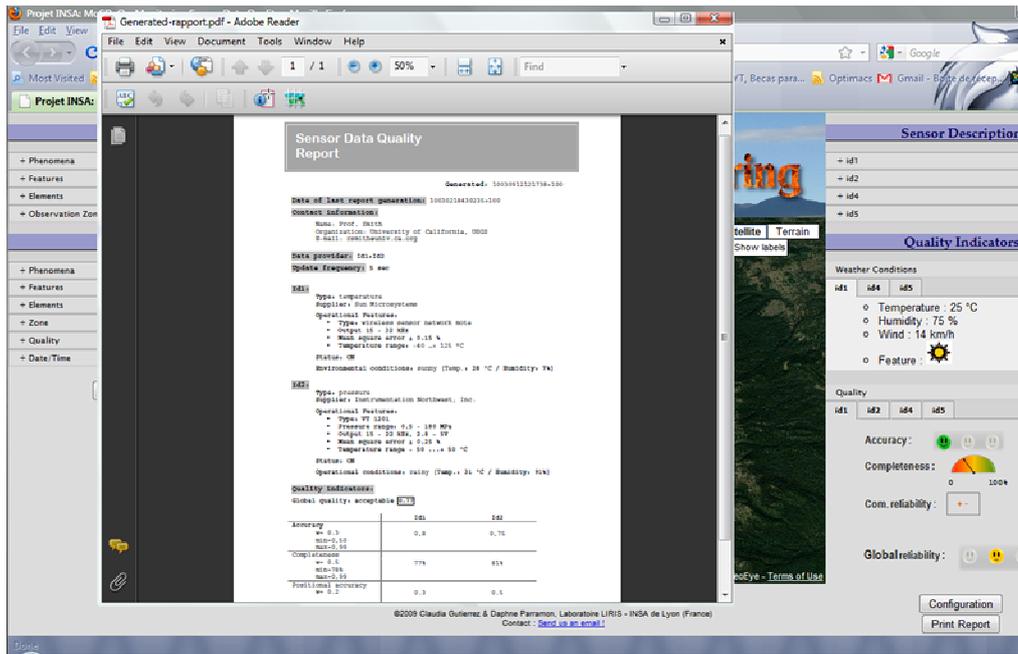


Figure 91 – Rapport qualité généré lors d'une évaluation

Comme nous l'avons cité auparavant, la possibilité de générer des rapports périodiques est une bonne stratégie, tout en le générant dans des périodes du temps gérables. Ce type de rapports est mieux exploité en format page web (html), gardant l'historique en ligne sous forme de pages (Figure 91).

V. Section d'interrogation

L'objectif principal de cette section est l'interrogation des différents éléments et caractéristiques de l'observation en cours ou encore de l'historique.

Notre proposition d'interrogation dans ce niveau est très basique car ce qui nous intéresse est que l'utilisateur ait accès aux informations qu'il considère les plus importantes. Dans cet esprit, nous proposons les éléments de base pour élaborer des requêtes spatio-temporelles en temps réel sur un phénomène observé. Premièrement, (1) nous proposons la sélection d'un phénomène à interroger ; ceci n'est pas obligatoire, juste informatif, mais de grand utilité quand diverses observations sont effectuées par le même expert. En deuxième lieu, (2) nous proposons à l'utilisateur l'ensemble des

caractéristiques qui sont liées au phénomène et parmi lesquelles doit sélectionner celles qui l'intéressent pour ses requêtes. Ensuite, (3) nous proposons à l'utilisateur l'ensemble des éléments qu'il peut interroger afin d'avoir plus d'informations sur le phénomène en question (Figure 92). Il peut alors délimiter une zone d'observation d'intérêt pour ses requêtes ou a la possibilité de cliquer sur un ou plusieurs capteurs qui sont déployés dans la zone d'observation ou sélectionner ceux qui sont actifs au moment de la requête (4).

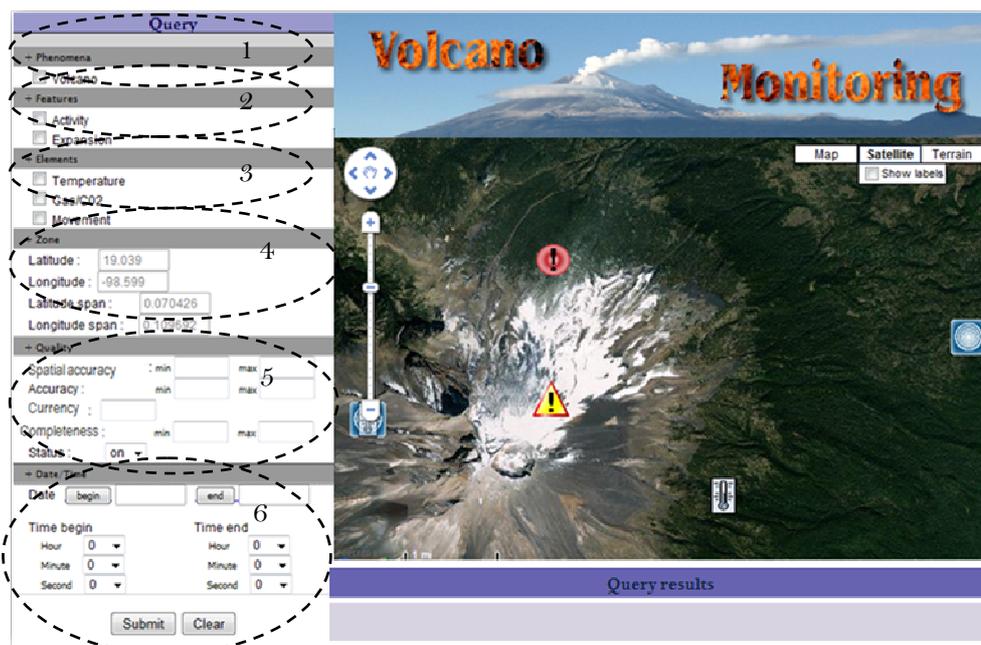


Figure 92 – Section d'interrogation

En complément à cet ensemble de paramètres utiles à l'élaboration de la requête de l'utilisateur, nous ajoutons les indicateurs qualité (5). Ces critères vont être pris en compte dans l'ensemble de données qui ont été déjà sélectionnées lors des choix précédents. Par exemple, les données qui sont issues d'une observation du type activité volcanique représentées par des valeurs de température d'un capteur spécifique et qui trouvent un indicateur d'exactitude entre 0.6 et 0.9.

Finalement, (6) nous ajoutons à cet ensemble d'aspects de l'information temporelle qui est de grande importance lors d'une surveillance. L'utilisateur peut déterminer la période ou instant sur lequel il veut obtenir de l'information. Nous considérons que ce cadre d'interrogation est préférable pour des requêtes hors d'un moment de crise, car le temps de sélectionner les paramètres à analyser peut demander un temps préjudiciable. Nous proposons à l'utilisateur dans ce type de cas, d'observer et exploiter l'ensemble des informations issues de requêtes en continu et prédéfinies que nous fournissons grâce aux métadonnées et données affichées en temps réel dans l'interface.

6.4 Conclusion

Nous avons présenté dans ce chapitre la description de notre proposition de visualisation MoSDaQ. Ce prototype aide les experts à améliorer leur connaissance sur les données afin d'évaluer la qualité des données exploitées via le web. Ce prototype est basé sur l'analyse des données issues de capteurs, ainsi que des informations contextuelles d'une observation. Ceci permet à l'utilisateur de mieux exploiter l'information qualité selon différentes perspectives (i.e. observation, capteurs déployés, conditions environnementales, indicateurs qualité, etc.).

Dans ce prototype, les données capteurs sont communiquées en même temps que les informations complémentaires (métadonnées), notamment concernant la qualité des données. L'information qualité est communiquée aux utilisateurs à travers des indicateurs qualitatifs et contextuels représentés avec des symboles graphiques dynamiques. Ces indicateurs sont affichés dans un tableau de bord intégré à l'interface web et au service de géolocalisation GoogleMaps.

Au cours de ce chapitre, l'architecture du prototype a été présentée ainsi que les différentes sections d'information qui la composent (section de cartographie des objets géolocalisés, section d'information observation, section d'information des capteurs, section d'indicateurs qualité, section d'interrogation). Ce prototype a permis la communication des informations qui aident l'utilisateur lors de la prise de décisions à des moments critiques. Ce prototype a été principalement conçu pour prouver les concepts proposés que nous avons considérés importants dans nos travaux de recherche. Nous fournissons ici, une première approche de ce qui peut être l'exploitation de la qualité des données dynamique via le web ou les systèmes d'information classiques.

Troisième partie : Conclusion et perspectives

7.1 Synthèse

La qualité des données dans les nouveaux systèmes d'information tels que les systèmes de surveillance reste un problème d'actualité. Au long de ce manuscrit nous avons identifié plusieurs problèmes liés à la qualité des données issues de capteurs dans un contexte de surveillance. Ces problèmes concernaient plus particulièrement la spécification des facteurs d'impact sur la qualité des données capteurs, ainsi que sur la gestion de cette qualité en temps réel. Nous avons identifié ces problèmes suite à une étude détaillée sur les approches, les standardisations et les méthodologies existantes concernant l'évaluation de la qualité des données dans les systèmes d'information et les spécificités de la gestion des données capteurs. Il ressort de l'état de l'art trois aspects importants à considérer : les modèles qualité (la relation entre les dimensions, critères et mesures qualité), les méthodologies et standardisations utilisées pour l'évaluation de la qualité dans les systèmes d'information, et les techniques de visualisation pour communiquer la qualité des données à l'utilisateur à base d'objets graphiques. Nous avons donc détaillé ces aspects en décrivant leur pertinence dans notre contexte de recherche. Nous avons remarqué que les modèles, méthodologies et standards existants orientés qualité, montrent une structuration solide et des applications diverses. Cependant, ces aspects sont utilisés pour l'évaluation des données statiques dans des applications traditionnelles et ne traitent pas la problématique de la qualité dans sa totalité, notamment dans des contextes dynamiques comme le traitement des données issues de capteurs. Quant aux techniques de visualisation existantes, notamment dans le domaine géographique, elles permettent la communication de l'information qualité, mais ne sont pas tout à fait adaptées à la dynamique de l'information. Dans ce cadre, nous avons défini la *dynamisme* comme étant *un contexte dans lequel les données ainsi que leur information complémentaire, spécialement sur la qualité, varient selon le temps*. Un bon nombre de travaux dans le domaine de la surveillance ont mis en évidence cette problématique d'évaluation de la qualité de données dynamiques, mais aucune approche selon notre perspective n'a été énoncée.

Nous avons donc tenté de résoudre cette problématique en abordant les spécificités des données issues de capteurs dans les systèmes de surveillance (cf. Chapitre 4). Dans cette étude, nous avons fourni les caractéristiques principales de ce type de systèmes et les données traitées. Nous avons abouti à la proposition d'un modèle de données capteurs, leur formalisation et la spécification des caractéristiques prises en compte dans leur traitement. Ceci a été effectué afin de structurer notre cadre d'évaluation de la qualité des données. Les caractéristiques et besoins ayant été ainsi spécifiés lors du chapitre 4, nous proposons une méthodologie pour la définition de la qualité des données issues de capteurs supportée par l'interface de visualisation MoSDaQ. La méthodologie que nous avons proposée lors de ce travail de recherche est orientée sur l'analyse de la qualité des données issues de capteurs dans un contexte de surveillance. Cette méthodologie aboutit à la définition d'un modèle qualité adapté à l'évaluation de la qualité selon les spécificités des données capteurs, à la gestion et visualisation en temps réel de cette information.

Cette méthodologie est spécifiée selon trois phases : *phase de définition*, *phase d'évaluation* et *phase de communication* de la qualité des données issues de capteurs. Dans ce manuscrit nous avons présenté en détail l'objectif de chacune des phases. Dans la première phase : *phase de définition*, nous avons étudié les facteurs qui peuvent impacter la qualité des données dans un contexte de surveillance. Selon ces facteurs, nous avons spécifié les critères qualité adéquats pour la caractérisation de la qualité dans ce contexte, ce qui nous a amenée à la proposition d'un modèle de qualité pertinent au regard des spécificités des données capteurs. La deuxième phase de cette méthodologie : *phase d'évaluation*, s'attache principalement à l'évaluation et la gestion de la qualité des données en temps réel. A ces fins, nous avons proposé l'utilisation des métadonnées qualité comme source d'information qualité. En accord avec les spécificités du domaine, cette proposition nous a conduit à l'analyse des métadonnées suivant quatre axes d'étude (la généralité, l'espace et le temps, les processus de traitement en temps réel et la granularité des données) et une proposition de structuration et gestion de ces métadonnées en temps réel. Grâce aux différents axes d'étude, nous avons pu constater que les métadonnées des données dynamiques possèdent de caractéristiques différentes aux métadonnées traditionnelles et ne peuvent pas être ni classifiées ni gérées de la même façon. Avec cet étude nous fournissons les différentes perspectives de classification des métadonnées dans un contexte dynamique, ce qui abouti à la proposition d'une structuration adéquate pour la gestion des métadonnées issues de données capteurs. Ces information seront donc établies comme sources primaires pour l'évaluation de la qualité de données capteurs. Lors de cette phase, nous avons également proposé l'évaluation de la qualité globale des données grâce à une approche multicritères et une gestion de l'information qualité obtenue en temps réel. Cette évaluation globale de la qualité en temps réel peut ainsi être évaluée à partir de plusieurs critères pondérés selon l'optique de l'utilisateur. Dans la troisième phase de cette méthodologie : *phase de communication*, nous essayons d'apporter des pistes pour la représentation visuelle de la qualité des données en nous basant sur les approches existantes dans le domaine géographique. Dans cette phase, nous présentons une analyse sur les stratégies de communication de la qualité existantes et les avantages et d'avantages de leur utilisation dans notre approche. Selon cette analyse, nous proposons la représentation des indicateurs qualité par des objets visuels dynamiques et la construction d'un rapport qualité sous la forme de document numérique. Suivant cette proposition, nous décrivons le type d'objet graphique qui correspond le mieux à chaque critère évalué et l'impact qu'il pourrait avoir lors de la prise de décision de l'utilisateur. Finalement, nous présentons lors de cette phase, la construction d'un rapport qualité et l'ensemble d'informations et éléments à prendre en considération.

Notre approche s'est appuyée sur l'interface de visualisation MoSDaQ (Monitoring Sensor Data and Quality). Cette interface offre la visualisation en temps réel et via le web l'ensemble des informations issues d'une observation en temps réel. Composée de cinq sections principales : *section de cartographie d'objets géolocalisés*, *section d'information observation*, *section d'information capteurs*, *section d'interrogation*, *section indicateurs qualité*. Cette interface permet à l'utilisateur d'exploiter les données issues de capteurs déployés dans une zone d'observation en temps réel et de prendre des décisions avec le support d'information complémentaires grâce à l'exploitation des métadonnées spatio-temporelles en temps réel. Ces métadonnées sont distribuées dans

toute l'interface graphique et concernent les caractéristiques de l'observation, les caractéristiques des capteurs déployés dans la zone d'observation, ainsi que l'information qualité. Dans cette approche, les requêtes de type continu concernent les données capteurs impliqués dans l'observation et leurs métadonnées associées. Grâce à notre proposition d'interface, nous fournissons une perspective actuelle de l'exploitation des données capteurs issues des systèmes de surveillance environnementaux et en adéquation avec les besoins de l'utilisateur.

7.2 Discussion et perspectives

Tout au long de ce travail de recherche, afin de résoudre notre problématique de la qualité des données capteurs, nous avons parcouru un grand nombre de travaux dans des domaines très vastes comme le traitement de flux de données, les bases de données, le domaine géographique et la qualité. Lors de notre étude, nous avons remarqué l'impact que la technologie a sur n'importe quel système d'information. Nombreuses techniques et approches proposées dans ces domaines, se concentrent sur l'optimisation : amélioration de la vitesse de calcul, économie de ressources, miniaturisation de l'équipement, etc. Mais les grandes avancées technologiques ont laissé à l'écart certaines de ces approches. Dans quelques années, la gestion de ressources de calcul et d'exécution ne se posera plus. Nos travaux s'inscrivent d'ailleurs dans cette perspective en cherchant à mieux informer l'utilisateur dans un contexte de prise de décision. Nous considérons que des approches plus innovatrices concernant l'accès, l'analyse et l'exploitation des informations seront toujours d'actualité.

Dans ce cadre, les perspectives suscitées par nos travaux sont multiples : elles concernent tout d'abord l'évaluation de notre approche auprès des experts dans le domaine géomatique et de la surveillance environnementale puis, les méthodes et les techniques d'analyse et d'évaluation de la qualité sur les données avec une forte dynamique et hétérogénéité, ou encore la découverte de connaissances et la prise de décision, ainsi que l'extension du prototype MoSDaQ.

7.2.1 Evaluation de la qualité de données capteurs et les experts

De façon générale, notre approche a été conçue pour permettre la communication aux utilisateurs, des différentes caractéristiques de la qualité des données capteurs qu'il exploite. Notre proposition reste à la fin de cette période de recherche, à un niveau empirique. De ce fait, nous sommes fortement intéressés et motivés à la mise à disposition de notre approche auprès des experts dans la géomatique et la surveillance environnementale qui exploitent des systèmes traitant des données capteurs. Dans ce cadre, des possibles collaborations avec l'Institut de Physique du Globe de Paris (IPGP), l'Institut de Recherche en Sciences et Technologies pour l'Environnement (Cemagref) et le Centre National de Prévention de Désastres (Cenapred) au Mexique seront concrétisées.

7.2.2 La qualité des données : dynamique et hétérogénéité

Comme nous l'avons exprimé dans ce manuscrit, la dynamique des données implique une dynamique de leur information qualité. Ainsi, quand la dynamique de ces données devient importante, comme c'est le cas des applications utilisant des flux des données à hautes fréquences, le défi devient encore plus évident. Dans ce cadre, les applications dans le domaine multimédia, les services web, le trafic routier et des navires, le contrôle et la surveillance, ou encore les systèmes pervasifs, montrent un fort besoin pour utiliser des méthodes et des techniques qui améliorent leur analyse et l'évaluation de leurs données en termes de qualité. Les données de ces domaines d'application sont issues de différentes sources d'information et stockées dans des entrepôts des données et bases de données distribuées. Toutefois, ces données sont extraites puis intégrées alors qu'elles ne possèdent pas le même niveau d'abstraction et la même description. Ceci nécessite donc une très grande attention lors de l'analyse de ces données.

Comme contribution à cette problématique, nous avons approfondi la prise en compte de la dynamique des informations et les possibles impacts que l'environnement et les processus de traitement ont sur la qualité des données issues de capteurs. Même si ces données sont dynamiques, elles restent à une fréquence gérable et une quantité raisonnable pour les systèmes de traitement actuels. Dans ce cadre, nous avons proposé une méthode de gestion des informations qualité au même niveau que le traitement des données en temps réel. Cependant, la problématique est actuellement plus large et ne se concentre pas seulement sur les données capteurs, mais concerne toute source d'information dynamique et hétérogène, comme les vidéos, l'audio, les transactions web, les données multidimensionnelles et distribuées, etc.

Concernant ces deux spécificités il est donc nécessaire donc d'approfondir et mettre en place des approches pour l'analyse et le nettoyage des données, ainsi que des méthodologies de prétraitement et métriques d'évaluation adéquates qui prennent en compte la dynamique des données et leur hétérogénéité.

7.2.3 L'extraction de connaissances et la prise de décision

La qualité des données joue un rôle très important dans l'extraction de connaissances et la prise de décisions. Quand l'utilisateur est confronté à l'extraction et la prise de décisions à partir des données avec une qualité médiocre, les conséquences sur le système et les activités propres de l'utilisateur quelque soit son domaine d'application peuvent être d'une gravité considérable. Par exemple, dans le domaine géospatial, il existe de nouvelles méthodes d'extraction des modèles de terrain ou d'imagerie satellitaire en deux ou trois dimensions, ainsi que l'analyse et recommandation des trajectoires routières ou de navires, entre autres. Dans ce cadre, un bon nombre d'approches se sont appuyées sur les techniques de fouille des données pour proposer à l'utilisateur, des modèles et informations extraites automatiquement. En effet, l'importance de la qualité des connaissances extraites dans ce domaine n'est pas anodine, surtout quand l'utilisateur base ses décisions sur ces aspects.

Dans notre approche présentée dans ce manuscrit, nous contribuons au problème de la prise de décision en fournissant de l'information contextuelle à l'utilisateur, notamment sur la qualité des données qu'il exploite. Nous évitons de réaliser des agrégations ou des hypothèses qui peuvent mettre en cause la représentation du monde réel. Lors de notre étude, nous avons détecté certaines propositions qui basent leur approche sur la création de « patterns » ou de règles d'extraction pour déterminer des situations à risque lors d'une observation ou diminuer au maximum la quantité des données à traiter. La plupart de ces travaux n'explorent pas totalement la qualité des techniques utilisées ni le traitement que ces données ont subi pour leur exploitation, rendant l'information douteuse et pouvant conduire à une prise de décision erronée. Nous considérons donc que la qualité des données et des connaissances reste d'actualité autant dans le domaine de la recherche que les domaines applicatifs.

7.2.4 L'extension du prototype MoSDaQ

Nous envisageons d'améliorer et d'étendre le prototype MoSDaQ dans des travaux futurs selon plusieurs aspects :

- Tout d'abord, en ce qui concerne le traitement des flux de données. Nous envisageons d'intégrer notre prototype MoSDaQ à des systèmes de gestion de flux existants (i.e. STREAM, PEMS...) afin d'étendre leur application et proposer notre approche à d'autres domaines d'application.
- Nous proposons également, de coupler notre prototype avec un système d'annotations qui permettra à l'utilisateur de faire des commentaires sur les informations visualisées lors d'une observation, et qui peuvent être prises en compte pour assister les utilisateurs lors de la prise de décision.
- Ensuite, pour compléter notre prototype, nous envisageons d'enrichir la partie interrogation intégrée dans notre interface. Pour cela, nous envisageons de le coupler avec un moteur de requêtes plus adapté aux spécificités des données spatio-temporelles dynamiques (i.e. PLACE, SoCQL), ainsi que de fournir des options graphiques plus ergonomiques et conviviales qui permettent à l'utilisateur de réaliser des requêtes plus complètes. De plus, nous envisageons de garder dans l'historique une trace des requêtes effectuées afin de les mettre à disposition de l'utilisateur grâce à un thesaurus.
- Finalement, nous aimerions adapter l'interface graphique à des autres cas d'applications et tester d'autres perspectives visuelles et d'interaction avec l'information qualité et géospatiale afin que notre interface soit suffisamment flexible pour s'adapter à n'importe quel type d'application dans le domaine de la surveillance.

Bibliographie

[Aalders'98] Aalders H.J.G.L., Morrison J. Spatial data quality for GIS. **In:** H. C. Onrud, Geographic Information Research : Trans-Atlantic Perspectives. Taylor and Francis, 1998, pp. 463–475.

[Abadi'03] Abadi A., Carney D., Cetintemel U., Cherniack M., Convey C., Erwin C., Galvez E., Hatoun M., Hwang J., Maskey A., Rasin A., Singer A., Stonebraker M., Tatbul N., Xing Y., Yan R., Zdnik S. Aurora: a data stream management system (demonstration). **In:** Proceedings of the ACM SIGMOD International Conference on Management of Data (SIGMOD'03).

[Aboelaze'05] Aboelaze M., Aloul F. Current and future trends in sensor networks: a survey. **In:** Sencond IFIP International Conference on Wireless and Optical Communications Networks, WOCN'05, 2005, pp. 551-555.

[Agumya'97] Agumya A., Hunter G.J. Determining fitness for use of geographic information. ITC Journal , 1997, pp. 109-113.

[Akoka'07] Akoka J., Berti-Équille L., Boucelma O. , Bouzeghoub M. , Comyn-Wattiau I. , Cosquer M.A. Framework for Quality Evaluation in Data Integration Systems. **In:** Proceedings of the 9th International Conference on Enterprise Information Systems - ICEIS, 2007. Madeira, Portugal, pp. 170-175.

[Akyildiz'02] Akyildiz I. F., Su W., Sankarasubramaniam Y., Cayirci E. A survey on sensor networks. IEEE Communications Magazine. 2002, pp. 102–114.

[Arasu'03] Arasu A., Babcock B., Babu S., Datar M., Ito K., Nishizawa I., Rosenstein J., Widom J. STREAM: The Stanford stream data manager demonstration description - short overview of system status and plans. *Proceedings of the ACM Intl Conf. on Management of Data (SIGMOD 2003)*.

[Babcock'02] Babcock B., Babu S., Datar, M., Motwani R., Widom J. Models and issues in data stream systems. **In:** Proceedings of Principles of Database Systems – PODS, 2002, pp. 1-16.

[Bastini'02] Visualizing uncertainty in multi-spectral remotely sensed imagery. Computers & Geosciences.

[Batini'09] Batini C., Cappiello C., Francalanci C., Maurino A. Methodologies for data quality assessment and improvement. ACM Comput. Surv. 41 (3), 2009, pp. 1-52.

[Batini'06] Batini C., Scannapieco M. Data Quality. Concepts, Methodologies and Techniques (Data-Centric Systems and Applications). 2006. Springer-Verlag New York, Inc. ISBN 3540331727.

[Bauman'05]. Bauman R. Radiation-Induced Soft Errors in Advances Semiconductor Technologies. IEEE Trans. Device and Materials Reliability , 2005, pp. 305-316.

[Beard'91] Beard M.K., Battenfield B. P., Clapham S.B. NCGIA Research Initiative 7: Visualization of Data Quality. Santa Barbara, Etats Unis, 1991.

- [**Bédard'95**]. Bédard Y., Valliere D. Qualité des données à référence spatiale dans un contexte gouvernemental. Rapport de recherche, Département des Sciences Géomatiques, Université Laval, Ste-Foy, 1995.
- [**Berti'99**] Berti L. Quality and Recommendation of Multi-source Data for Assisting Technological Intelligence Applications. *DEXA'99 LNCS 1677*, Springer-Verlag, 1999, pp. 282-291.
- [**Berti-Equille'02**] Berti-Equille L. Annotation et recommandation collaboratives de documents selon leur qualité. *Ingénierie des Systèmes d'Information 7(1-2)*, 2002, pp. 125-155.
- [**Berti-Equille'04**] Un état de l'art sur la qualité des données. *Ingénierie des Systèmes d'Information 9(5-6)*, 2004, pp.117-143.
- [**Biswas'06**] Biswas, J., Naumann F., Qiu Q. Assessing the Completeness of Sensor Data . *Database Systems for Advanced Applications* , 2006, pp.717-732.
- [**Bonnet'00**] Bonnet P., Gehrke J., Seshadri P. Querying the Physical World. *IEEE personal Communication. Special Issue - Networking the Physical World (7)*, 2000, pp. 10-15.
- [**Bonnet'01**] Bonnet, P., Gehrke J. Seshadri P. Towards Sensor Database Systems. **In:** MDM'01: Proceedings of the Second International Conference on Mobile Data Management. 2001, London, UK: Springer-Verlag. pp. 3-14.
- [**Bouzeghoub'04**] Bouzeghoub M. A Framework for Analysis of Data Freshness. **In:** Proceedings of Proceedings of the 2004 international workshop on Information quality in information systems – IQIS, 2004, Paris, France, pp. 59-67.
- [**Bovee'01**] Bovee M., Mak, B., Srivastava R.P. A conceptual framework and belief-function approach to assessing overall information quality. **In:** Proceedings of the 6th International Conference on Information Quality. 2001, pp. 311-328.
- [**Brodie'80**] Brodie M.L. Data quality in informations systems, *Information and Management. Communications of the ACM* , 1980, pp. 245-258.
- [**Bulusu'02**] Bulusu N., Heidemann J., Bychkovskiy V., Estrin D. Density adaptive beacon placement algorithms for localization in ad hoc wireless networks. *IEEE Infocom 2002*. New York, NY.
- [**Bychkovskiy'03**] Bychkovskiy V., Megerian S., Estrin D., Potokonjak M.A. Collaborative Approach to In-Place Sensor Calibration. **In:** Proc. Int'l Workshop Information Processing in Sensor Networks (IPSN '03), 2003, pp.301-316.
- [**Carter'94**] Carter B., Fox C. The notion of data and its quality dimensions. *Information Processing and Management vol 30 (1)*, 1994, pp. 9-19.
- [**Cenapred'06**] CENAPRED. Cenapred, Secretaría de Gobernación de México. Monitoreo Volcánico. Consulté avril 2010 sur <http://www.cenapred.unam.mx/es/Instrumentacion/InstVolcanica/MVolcan/>
- [**Challen'09**] Challen, G. W., Waterman J., Welsh M. Integrated distributed energy awareness for wireless sensor networks. **In:** SenSys '09: Proceedings of the 7th ACM Conference on Embedded Networked Sensor Systems. Berkeley, California. ACM, 2009, pp. 381-382.
- [**Chen'04**] Chen D., Varshney P.K. QoS Support in Wireless Sensor Networks: A Survey. **In:** Proceedings of the International Conference on Wireless Networks, 2004 , pp. 227-233.

- [**Choi'03**] Choi Y., Gouda M.G., Kim M.C., Anish A. The Mote Connectivity Protocol. Department of Computer Sciences, the University of Texas at Austin. 2003.
- [**Chrisman'83**] Chrisman N. The role of quality information in the long-term functioning of a Geographic Information System. **In:** Proceedings of AUTOCARTO 6, 1983, pp. 303-321.
- [**CIHI'09**] The Canadian Institut for Health Information. The CIHI Data Quality Framework. 2009. Ottawa, Ont. ISBN:978-1-55465-656-7 (PDF).
- [**CISCO'01**] *Quality of Service (QoS)*. Internetworking Technology Handbook. 1-58705-001-3. Consulté avril 2010.
<http://www.cisco.com/en/US/docs/internetworking/technology/handbook/QoS.pdf>
- [**Culler'04**] Culler D.E., Mulder H. Smart Sensors to Network the World. Scientific American, Inc, New York, 2004.
- [**Dasu'03**] Dasu T., Johnson T. Exploratory Data Mining and Data Cleaning. J. Wiley Series in Probability and Statistics. 2003.
- [**David'06**] David B., Fasquel P. Qualité d'une base de données géographique : concepts et terminologie. (IGN, Éd.) Bulletin d'information de l'IGN, 2006.
- [**Davis'93**] Davis R., Shrobe H., Szolovits P. What is a Knowledge Representation? AI Magazine, 14(1), 1993, pp. 17-33.
- [**Delin'00**] Delin K.A., Jackson S.P. Sensor web for in situ exploration of gaseous biosignatures. **In:** Proceedings of the 2000 IEEE Aerospace Conference, 2000, p. 16.
- [**Delin'05**] Delin, K.A. Sensor Webs in the Wild. **In:** Wireless Sensor Networks: A Systems Perspective, N. Bulusu and S. Jha, (eds.), Artech House, 2005.
- [**Devillers'02**] Devillers R., Gervais M., Bédard Y., Jeansoulin R., Spatial Data Quality : From Metadata to Quality Indicators and Contextual End-User Manual. **In:** Proceedings of OEEPE-ISPRS Joint Workshop on Spatial Data Quality, 2002.
- [**Devillers'04**] Devillers R., Bédard Y., Gervais, M. Indicateurs de qualité pour réduire les risques de mauvaise utilisation des données géospatiales. Revue Internationale de Géomatique, vol.14, no.1, 2004, pp. 35-57.
- [**Devillers'06**] Devillers R., Jeansoulin R. (eds.). Fundamentals of Spatial Data Quality. ISTE Publishing Company. April 2006. ISBN: 1905209568, pp.312.
- [**Devillers'07**] Devillers R., Bédard, Y., Jeansoulin R., Moulin B. Towards spatial data quality information analysis tools for experts assessing the fitness for use of spatial data. International Journal of Geographical Information Science, 2007, pp. 261-282.
- [**Dey'99**] Dey A.K., Abowd G.D. Toward a better understanding of context and context-awareness. College of Computing, Georgia Institute of Technology, 1999.
- [**Elnahrawy'03a**] Elnahrawy E., Nath, B. Cleaning and Querying Noisy Sensors. **In:** Proc. Second ACM Int'l Workshop Wireless Sensor Networks and Applications (WSNA '03), 2003.
- [**Elnahrawy'03b**] Elnahrawy E. Research direction in sensor data streams: solutions and challenges. Rutgers University, 2003.
- [**English'99**] English, L. P. Improving Data Warehouse and Business Information Quality. John Wiley & Sons, Inc. 1999.
- [**Fisher'94**] Fisher P.F. Visualization on the reliability in classified remotely sensed images. Photogrammetric Engineering & Remote Sensing, 1994, pp. 905-910.

- [**Fisher'99**] Fisher P.F. Models of Uncertainty in Spatial Data. Dans M. M. Goodchild, *Geographic Information Systems* (pp. 191-205). New York: John Wiley and Sons.
- [**Faïz'99**] Faïz, S.O. Modélisation, exploitation et visualisation de l'information qualité dans les bases de données géographiques . Thèse. France: Université Paris-Sud.
- [**Frank'00**] Frank R. Understanding Smart Sensors. Artech House, Inc. 2000.
- [**Fries'07**] Fries E.F., Karra V.K., Sammarco J.J., Paddock R. A technology review of smart sensors with wireless networks for applications in hazardous work environments. Department of Health and Human Services. 2007.
- [**Gaber'05**] Gaber M.M., Zaslavsky S., Krishnaswamy S. Mining data streams: a review. *ACM SIGMOD Record* , 2005, pp. 18-26.
- [**Gama'07**] Gama J. and Gaber M. Learning from Data Streams - Processing techniques in Sensor Networks. Springer. 2007.
- [**Ganz'04**] Ganz A., Ganz Z., Wongthavarawat K. Multimedia Wireless Networks: Technologies, Standards, and QoS. Prentice Hall. 2004.
- [**Gervais'06**] Auditing spatial data suitability for specific applications: professional and technological issues. **In:** Workshop on quality assurance in geographic data production. Marne-la-Vallée, France, 2006.
- [**Goodchild'98**] Goodchild, M. J. Data quality in geographic information : from error to uncertainty. Goodchild M., Jeansoulin R., 1998.
- [**Govindan'02**] Govindan R., Ratnasamy S., Estrin D. Data-centric storage in sensornets. **In:** Proceedings of ACM Int. Workshop on Wireless Sensor Networks. 2002.
- [**Gripay'09**] Gripay Y., Laforest F., Petit J-M. SoCQ: A Framework for Pervasive Environments. **In:** ISPAN 2009, 10th International Symposium on Pervasive Systems, Algorithms and Networks, 2009, pp. 1-6. Kaohsiung, Taiwan.
- [**Günther'98**] Günther, O. Environmental Information Systems. Springer-Verlag, 1998.
- [**Guitton'07**] Guitton A., Skordylis A., Trigoni N. Utilizing correlations to compress time-series in traffic monitoring sensor networks. **In:** IEEE Wireless Communications and Networking Conference (WCNC), 2007.
- [**Guptill'95**] Guptill S.C. Metadata and data catalogues. In: Longley, P., Goodchild, M.F., Maguire, D.J., Rhind, D.W. (Eds.), *Geographical Information Systems*, Wiley, Chichester. 1995, pp. 677-692.
- [**Gutiérrez'07**] Gutiérrez C., Servigne S., Laurini R. Towards Real Time Metadata for Network-Based Geographic Databases. **In:** ISSDQ2007, 5th International Symposium of Spatial Data Quality, 2007, p. 8. Enschede, The Netherlands: ISPRS.
- [**Gutiérrez'07a**] Gutiérrez C., Servigne S. Métadonnées Spatiotemporelles Temps-Réel. *Revue des sciences et technologies de l'information : Ingénierie des Systèmes d'Information* vol. 12(2), 2007, pp. 97-119, Lavoisier, ISBN 978-2-7462-1913-.
- [**Gutiérrez'08**] Gutiérrez C., Servigne S. Bases de données capteurs et Métadonnées. **In :** Conférence Québeco-Française pour le développement de la Géomatique, SAGEO 2008, Montpellier, France.
- [**Gutiérrez'09**] Gutiérrez C., Servigne S. Métadonnées et Qualité pour les Systèmes de Surveillance en Temps-Réel. *Revue Internationale de Géomatique*, vol.19(2), 2009, pp. 1-168, Lavoisier. ISBN 978-2-7462-2479-7.

- [**Haas'96**] Haas, P.J. Large-Sample and Deterministic Confidence Intervals for Online Aggregation. *SSDBM* (pp. 51–63). IEEE Computer Society Press.
- [**Hart'06**] Hart J. K., Martinez K. Environmental Sensor Networks: A revolution in the earth system science? *Earth-Science Reviews* . 2006.
- [**Hellerstein'02**] Hellerstein J., Hong W., Madden S., Franklin M. Tag : Tiny aggregate queries in adhoc. **In:** Proceedings of the USENIX Symposium on Operating Systems, 2002.
- [**Henson'07**] Henson C., Sheth A., Jain P., Pschorr J., Rapoch T. Video on the Semantic Sensor Web. **In:** W3C Video on the Web Workshop, 2007.
- [**Hoes'09**] Hoes R., Basten T., Tham C., Geilen M., and Corporaal H. Quality-of-service trade-off analysis for wireless sensor networks. *Performance Evaluation*, 2009, pp. 191-208.
- [**Holdridge'04**] Holdridge G.M., Gould S. Sensors for Environmental Observatories. *Report of the NSF-Sponsored Workshop*. WTEC, Inc., 2004.
- [**Huijsing'94**] Huijsing J.H., Riedijk F. R., Horn G. v. d. Developments in integrated smart sensors," *Sensors and Actuators A* 43, 1994, Special issue on Transducers'93.
- [**Hunter'99**] Hunter G.J. Managing uncertainty in GIS. *Geographical Information Systems*, 1999, pp. 633-641.
- [**Hunter'01**] Hunter G. J. Spatial Data Quality Revisited. **In:** *Proceedings of GeoInfo*. 2001.
- [**Huth'07**] Huth K., Mitchel N., Schaab, G. Judging and Visualising the Quality of Spatio-temporal Data on the Kakamega - Nandi Forest Area in West Kenya. **In:** 5th International Symposium on Spatial Data Quality (ISSDQ2007). ITC, Enschede, The Netherlands.
- [**ISO'94**] ISO. Quality management and quality assurance - Vocabulary. ISO 8402, 1994.
- [**ISO'99**] ISO. Information technology -- Quality of service -- Guide to methods and mechanisms. JTC 1/SC 6. ISO/IEC TR 13243:1999, 1999.
- [**ISO'00**] ISO. Quality Management Systems - Requirements . ISO 9000, 2000.
- [**ISO'02**] ISO. Geographic Information - Quality Principles. ISO 19113. 2002.
- [**ISO'03**] ISO. Geographic information - Quality evaluation procedures. Technical Committee. ISO 19114, 2003.
- [**ISO'03a**] ISO. Geographic information - Metadata. Technical Committee. ISO19115, 2003.
- [**ISO'04**] ISO. Geographic information – Geography Markup Language. Technical Committee. ISO 19136, 2004.
- [**ISO'06**] ISO. Geographic information – Data quality measures. Technical Committee. ISO 19138, 2006.
- [**Johnson'03**] Johnson T., Cranor C., Spatscheck O. Gigascope: A stream database for network application. **In:** Proceedings of the 2003 ACM SIGMOD International Conference on Management of Data, 2003, pp. 647-651.
- [**Kahn'99**] Kahn J. M., Katz R. H., Pister K. S. J. Next century challenges: Mobile networking for “smart dust”. **In:** Proceedings of 5th Annual ACM/IEEE International Conference on Mobile Computing and Networking (MOBICOM), 1999, pp. 271-278.

- [**Kahn'02**] Kahn B.K., Wang R.Y., Lee Y.W., Strong D.M. AIMQ : A methodology for information. *Information Management*, 2002, pp. 133–460.
- [**Kaplan'92**] Kaplan R., Norton D. The balanced scorecard: Measures that Drive Performance. *Harvard Business Review* , 1992, pp. 71-79.
- [**Karl'03**] Karl, H., Willig, A. A short survey of wireless sensor networks. Berlin: Technical University Berlin - Telecommunication Networks Group, 2003.
- [**Khemapech'05**] Khemapech I., Duncan I., Miller A. A survey of wireless sensor networks technology. **In:** PGNET - Proceedings of the 6th Annual PostGraduate Symposium on the Convergence of Telecommunications, 2005.
- [**Klein'07**] Klein A., Do H.H., Lehner W. Representing data quality for streaming and static data. . *In Proceedings of the International Workshop on Ambient Intelligence, Media, and Sensing (AIMS). AIMS Workshop*, (pp. 3-10).
- [**Konstantinidis'07**] Konstantinidis A., Kun Y., Chen H., Qingfu Z. Energy-aware topology control for wireless sensor networks using memetic algorithms. *Computer communications*. 2007, pp. 2753-2764.
- [**Koo'07**] Koo L.Y., Jin J. Y., Ho R. K. Design and Implementation of a System for Environmental Monitoring Sensor Network. **In:** APWeb/WAIM Workshops, 2007, pp. 223-228.
- [**Lam'01**] Lam Y., Kuo T.W (Eds). *Real-Time Databases systems: Architecture and Techniques*. Kluwer Academic publishers, 2001, ISBN 0-7923-7219-2.
- [**Laurini'94**] Laurini, R. Towards Visual Interfaces for Multi-Source Updating of Geographic Information. **In:** NATO Advanced Research Workshop, Cognitive Aspects of Human Computer Interfaces for Geographic Information Systems, 1994, pp. 147-162.
- [**Lenzerini'02**] Lenzerini M. Data Integration: A Theoretical Perspective. **In:** Proceedings of PODS, 2002.
- [**Lian'07**] Lian J., Naik K., Chen L., Liu Y., Agnew G. Gradient boundary detection for time series snapshot construction in sensor networks. *IEEE Transactions on Parallel and Distributed Systems (TPDS)*, 2007.
- [**Lohningen'99**] Lohningen H. *Teach me Data Analysis*. Springer-Verlag. 1999.
- [**Lorincz'04**] Lorincz K., Malan D., Fulford-Jones T.R.F., Nawoj A., Clavel A., Shnayder V., Mainland G., Moulton S., Welsh M. Sensor networks for emergency response : challenges and opportunities. *IEEE Pervasive Computing, Special Issue on Pervasive Computing for First Response*, 2004, pp.16-23.
- [**Ma'05**] Ma L., Viglas S.D, Li M., Li Q. Stream Operators for Querying Data Streams. *Advances in Web-Age Information Management*, 2005.
- [**MacEachren'92**] Maceachren A.M. *Visualizing uncertain information*. Cartographic Perspectives, 1992.
- [**Mainwaring'02**] Mainwaring A., Polastre J., Szewczyk R., Culler D., Anderson J. Wireless sensor networks for habitat monitoring. **In:** WSNA'02, 2002.
- [**Meguerdichian'01**] Meguerdichian S., Koushanfar F., Potkonjak M., Srivastava M.B. Coverage Problems in Wireless Ad-hoc Sensor Networks. **In:** Proceedings of IEEE Infocom, 2001.
- [**Milenković'06**] Milenković A., Otto C., Jovanov E. Wireless sensor networks for personal health monitoring: Issues and an implementation. *Wireless Sensor Networks: Performance, Reliability, Security, and Beyond*, 2006.

- [**Münzenmaier'02**] Münzenmaier P., Eppler M. Measuring information quality in the web context: A survey of state-of-art instruments and an application methodology. **In:** In Proc. of 7th International conference on Information Systems (ICIQ), 2002.
- [**MMI'07**] Marine Metadata Interoperability. Marine Metadata Interoperability Project. Consulté avril 2010 sur <http://marinemetadata.org/>.
- [**Mokbel'05**] Mokbel M.F., Xiong X., Hammad M. A., Aref W. G. Continuous Query Processing of Spatio-Temporal Data Streams in PLACE. *Geoinformatica*, 2005, pp. 343-365.
- [**Mukhopadhyay'04**] Mukhopadhyay S., Panigrahi D., Dey S. Data Aware, Low Cost Error Correction for Wireless Sensor Networks. **In:** Proceedings of IEEE Wireless Communication and Networking Conference (WCNC '04), 2004.
- [**Naumann'99**] Naumann F., Leser U., Freytag J.G. Quality-driven Integration of Heterogenous Information Systems. **In:** Proceedings of VLDB'999. Edinburgh, UK. 1999.
- [**Naumann'02**] Naumann F. Quality-driven query answering for integrated information systems. Lecture Notes in Computer Science . Springer-Verlag New York, Inc. 2002. ISBN 3-540-43349-X.
- [**Naumann'04**] Naumann F., Linden U.D., Rolker C. Assessment Methods for Information Quality Criteria. **In:** Proceedings of ICWE 2004. Springer-Verlag Heidelberg. 2004.
- [**Nittel'09**] Nittel, S. A Survey of Geosensor Networks: Advances in Dynamic Environmental Monitoring. *Sensors*, 2009.
- [**NOAA'09**] National Oceanic and Atmospheric Administration, NOAA. Consulté avril 2010 sur <http://www.noaa.gov/>
- [**Noël'06**] Noël, G. Indexation dans les bases de données capteurs temps réel. Application à la surveillance de phénomènes environnementaux et de risques naturels [En ligne]. Thèse : Institut National des Sciences Appliquées de Lyon, 2006, 213 p. Disponible sur <http://docinsa.insa-lyon.fr/these/pont.php?&id=noel>
- [**OGC'09**] OGC. The Open Geospatial Consortium, Inc. Consulté avril 2010, sur <http://www.opengeospatial.org/>
- [**Onchaga'04**] Onchaga R. Quality of service provisioning in geographic information service. *International Journal of Applied Earth Observation and Geo-Information*, 2004.
- [**Palpanas'03**] Palpanas T., Papadopoulos D., Kalogeraki V., Gunopulos D. Distributed devitaion detection in sensor networks. **In:** ACM SIGMOD 2003, pp. 77-82.
- [**Peralta'04**] Peralta V., Bouzeghoub M. On the evaluation of data freshness in data integration systems. **In:** BDA 2004, pp. 287-307.
- [**Perillo'03**] Perillo M., Heinzelman W. Providing Application QoS Through Intelligent Sensor Management. **In:** Proceedings of the 1st IEEE International Workshop on Sensor Network Protocols and Applications (SNPA'03), 2003.
- [**Piattini'00**] Piattini M., Genero M., Calero C., Polo C., Ruiz, F. Advanced Database Technology and Design. Dans *Database Quality*. Artech House, 2000.
- [**Pipino'02**] Pipino L., Lee Y.W., Wang R.Y. Data quality assessment. *Communications ACM* 45(4), 2002, pp. 211-218.

- [**Pister'02**] Pister K.S.J., Warneke B. Mems for distributed wireless sensor networks. **In:** 9th International Conference on Electronics, Circuits and Systems . Dubrovnik, Croatia, 2002.
- [**Quaas'06**] Quaas R., Guevara E. Monitoring and Warning System: The Mexican Experience. **In:** Third International Conference on Early Warning from Concept to Action (EWCIII). Bonn, Germany, 2006.
- [**Ratnasamy'03**] Ratnasamy S., Karp B., Shenker S. Estrin D. Govindan R., Yin L., Yu F. Data-centric storage in sensornets with GHT, a geographic hash table. *Mobile Networks and Applications*, vol. 8(4), 2003, pp. 427-442.
- [**Redman'97**] Redman T. C. *Data Quality of the Information Age*. Artech House, Inc., 1997, ISBN 0890068836.
- [**Redman'01**] Redman T. C. *Data Quality: The field Guide*. Digital Press, Elsevier, 2001. ISBN 1-55558-251-6.
- [**Rocker'04**] Assessment methods for information quality criteria. **In:** Proceedings ICWE 2004, 2004, pp. 147-160. Springer-Verlag Heidelberg.
- [**Savvides'01**] Savvides A. Han C.C., Srivastava M. Dynamic fine-grained localization in ad-hoc networks of sensors. **In:** 7th ACM Int. Conf. on Mobile Computing and Networking (Mobicom), 2001, pp. 166 -179. Rome, Italy.
- [**Scannapieco'04**] Scannapieco M., Virgillito A., Marchetti C., Mecella M., Baldoni R. The DaQuinCIS Architecture: A Platform for Exchanging and Improving Data Quality in Cooperative Information Systems. *Information Systems*, 2004, pp. 551-582.
- [**Sergey'03**] Sergey Y. Smart Sensor for Electrical, and Non-electrical, physical and chemical variables: State-of-Art. *Smart sensors and MEMS*, 2003.
- [**Servigne'06**] Servigne S., Lesage N., Libourel T. Spatial data quality components, standards and metadata. *Fundamentals of Spatial Data Quality*. International scientific and technical, 2006, pp. 179-208.
- [**Servigne'09**] Servigne S., Devogèle T., Bouju A., Bertrand F., Gutierrez C., S. Laucius S., Noël G., Ray C. Gestion de masses de données temps réel au sein de bases de données capteurs. *Revue international de Géomatique*. 2009.
- [**Seshadri'95**] Seshadri P., Livny M., Ramakrishnan R. SEQ: A Model for Sequence Databases. **In:** Proceedings of ICDE'95, 1995, pp. 232-239.
- [**Shankaranarayan'00**] Shankaranarayan G., Wang R., Ziad M. Modeling the Manufacture of an Information Product with IP-MAP. **In:** Proceedings of 5th International Conference on Information Quality (IQ 2000), 2000.
- [**Shenker'97**] Shenker S., Partridge C., Guerin R. *Specification of Guaranteed Quality of Service*, 1997.
- [**Sherwood'04**] Sherwood R., Chien S. Sensor Web Technologies: A New Paradigm for Operations. **In:** International Symposium on Reducing the Cost of Spacecraft Ground Systems and Operations (RCSGSO 2007). Moscow, Russia. 2004.
- [**Simonis'05**] Simonis I., Sliwinski A. *Quality of Service in a Global SDI*, 2005. Kairo, Egypt.
- [**Smith'98**] Smith, I., Pipino, L. Proceedings of the Third Information Quality Conference MIT, October 1998.
- [**Stavrou'05**] Stavrou, E. *Wireless Sensor Networks - Part 2 and 3 - Limitations*. Consulté avril 2010, sur Web Hosting Articles:

<http://webhosting.devshed.com/c/a/Web-Hosting-Articles/Wireless-Sensor-Networks-part-2-Limitations>

[**Storey'98**] Storey V. C., Wang R. Y. An Analysis of Quality Requirements in Database Design. **In:** Proc. 4th International Conference on Information Quality, 1998.

[**Storey'01**] Storey V. C., Wang R. Extending the ER Model to Represent Data Quality Requirements. Dans R. Z. Wang, Data Quality. Kluwer Academic Publishers.

[**Strong'97**] Strong D., Lee Y., Wang R. Data Quality in Context. **In:** Proceedings of the Second Information Quality Conference MIT. 1997, pp. 103-110.

[**SWE'09**] OGC. Sensor Web Enablement. Consulté avril 2010 sur Sensor Web Enablement : <http://www.opengeospatial.org/ogc/markets-technologies/swe>

[**SwissExp'09**] Swiss Experiment. Interdisciplinary Environmental Research. Consulté avril 2010 sur <http://www.swiss-experiment.ch/index.php/Main:Home>

[**Szewczyk'04**] Szewczyk R., Osterweil E., Polastre J., Hamilton M., Mainwaring A. Habitat monitoring with sensor networks. Communications of ACM , 2004, pp. 34-40.

[**Taylor'99**] Taylor, J.R. An Introduction to Error Analysis: The Study of Uncertainties in Physical Measurements,1999. University Science Books.

[**Thomas'08**] What's in a message? interpreting geo-referenced data for the visually-impaired. **In:** Proceedings of INLG08. 2008.

[**TAO'06**] Pacific Marine Environmental Laboratory. Tropical Atmosphere Ocean project (TAO). Consulté avril 2010, sur <http://www.pmel.noaa.gov/tao/>

[**Taoussi'03**] Taouss A. Les serveurs de métadonnées géographiques : Moyen organisationnel et outil de diffusion de l'information spatiale. **In:** 2nd FIG Regional Conference Marrakech, 2003. Morocco.

[**Thirunarayan'09**] Semantic information and sensor networks. **In:** SAC '09: Proceedings of the 2009 ACM symposium on Applied Computing, 2009, pp. 1273-1274. Honolulu, Hawaii: ACM.

[**Timpf'96**] Timpf S., Raubal M., Khun W. Experiences with Metadata. **In:** Proceedings of Spatial Data Handling 96, Advances in GIS Research II. 1996.

[**Trevisan'08**] Trevisan J. Un SIG, c'est quoi ? Consulté avril 2010, sur Géoazur: <http://geoazur.oca.eu/spip.php?article208>

[**Tubaishat'03**] Tubaishat M., Madria S. Sensor networks: an overview. IEEE Potentials , 2003, pp. 20-23.

[**USGS'00**] U.S. Geological Survey. Content Standard for Digital Geospatial Metadata Workbook, version 2. FGDC. 2000.

[**Vasilescu'05**] Vasilescu I., Kotay K., Rus D., Dunbabin M., Corke P. Data collection, storage and retrieval with an underwater sensor network. **In:** Proceedings of IEEE SensSys, 2005.

[**Veregin'99**] Veregin H. Data quality parameters. Dans M. G. Longley, Geographical Information Systems John Wiley and Sons, Inc, 1999, pp. 177-189.

[**Wac'07**] Wac K., Van Halteren A., Bults R., Broens T. Context-aware QoS provisioning in an m-health service platform. International journal of Internet protocol technology, 2007, pp. 102-108.

[**Wand'96**] Wand Y., Wang R. Anchoring data quality dimensions in ontological foundations. Communications ACM, 1996.

- [Wang'90] Wang R., Madnick S.E. A polygen Model for Heterogeneous Database Systems: The Source Tagging Perspective. **In:** Proceedings of VLDB'90, 1990. Brisbane, Queensland, Australia, pp. 519-538.
- [Wang'93] Wang, R., Kon, H.B., Madnick, S.E. Data quality requirements analysis and modeling. **In:** Proceedings of the 9th International Conference on Data Engineering, 14, 1993, pp. 670-677.
- [Wang'95] Wang R.Y., Storey V.C., Firth C.P. A framework for analysis of data quality research. *IEEE Transactions on Knowledge and Data Engineering* , 7 (4), 1995, pp. 623-638.
- [Wang'96] Wang R.Y., Strong D.M. Beyond Accuracy: What Data Quality Means to Data Consumers. *Journal of Management Information systems*, 1996.
- [Wang'98] Wang R. Y. A product perspective on total data quality management. *Communications of the ACM*, vol 4(2), 1998, pp. 58-65.
- [Wang'01] Wang R. Y., Ziad M., Lee Y.W. *Data quality*. Kluwer Academic Publisher, 2001.
- [Wilson'05] Wilson J.S. *Sensor Technology Handbook*. Elsevier, 2005.
- [Worboys'06] Worboys M., Duckham M. Monitoring qualitative spatio-temporal change for geosensor networks. *International Journal of Geographical Information Science*, 2006.
- [Xiang'04] Xiang J., Hongyuan Z. Sensor Positioning in Wireless Ad-hoc Sensor Networks Using Multidimensional Scaling. **In:** Proceedings of INFOCOM'04, 2004.
- [Xu'04] Xu N., Rangwala S., Chintalapudi K.K., Ganesan D., Broad A., Govindan R., Estrin D. A Wireless Sensor Network for Structural Monitoring. **In:** Proceedings of the 2nd international conference on Embedded networked sensor systems.
- [Xu'09] Xu N. A Survey of Sensor Network Applications. Survey Paper for CS694a, Computer Science Department, 2009.
- [Yao'02] Yao Y., Gehrke J. The cougar approach to in-network query processing in sensor networks. **In:** PACM SIGMOD, 2002, pp. 9-18.
- [Yong'08] Yong G., Bai H., Sanping L. Geo-spatial Data Analysis, Quality Assessment and Visualization. **In:** Computational Science and Its Applications – ICCSA 2008 , pp. 258-267.
- [Zhou'05] Zhou Z., Bai S., Hu B., MC Guire N., Li L. A novel portable multimedia QoS monitor: independent and high efficiency. *Wireless Communications and Mobile Computing*, 2005.

Annexe

Structuration des métadonnées spatio-temporelles issues de capteurs

a) Métadonnées observation

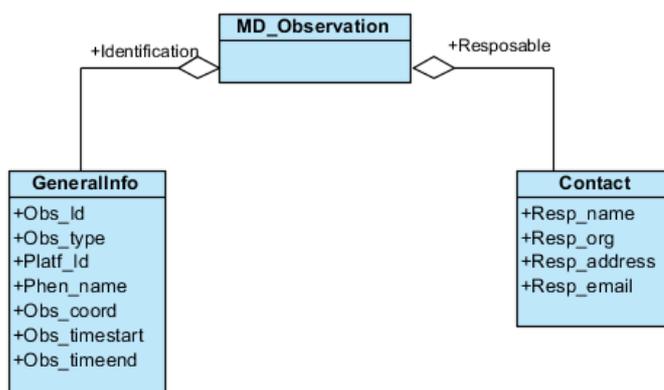


Figure 93 – Composition des Métadonnées Observation

```

...
<xs:group name="MD_Observation">
<xs:annotation>
<xs:documentation>Métadonnées observation</xs:documentation>
</xs:annotation>
<xs:sequence>
<xs:group ref="GeneralInfo" minOccurs="0"/>
<xs:group ref="Contact" minOccurs="0"/>
</xs:sequence>
</xs:group>
...
  
```

Tableau 21 – Extrait du schéma de métadonnées observation

GeneralInfo :

[Obs_Id,Obs_Type,Platf_Id,Phen_name,Obs_coord,Obs_timestart,Obs_time end]

- Obs_Id est l'identification de l'observation. Chaque fois qu'une observation est déclenchée un numéro d'identification est affecté.
- Obs_Type est la description du type d'observation réalisée. Les observations peuvent être du type : reconnaissance, contrôle, vérification, continues, programmés.
- Platf_Id contient l'identificateur de la plateforme de surveillance. Typiquement, les capteurs ou réseaux de capteurs utilisés pour surveiller un phénomène sont attachés à une plateforme d'observation. Il est important de garder cette information afin de lier les résultats obtenus à chaque plateforme.

- Phen_name correspond à l'identification du phénomène à surveiller. Dans un environnement d'observation, plusieurs phénomènes peuvent être surveillés en même temps (i.e. éruption, séisme...). Chaque phénomène observé possède des spécifications particulières, lequel est important à identifier.
 - Obs_coord, chaque observation est liée à une position géographique, zone, région ou point dans l'espace.
 - Obs_time_start, est la date/heure du début de l'observation.
 - Obs_time_end, représente la date/heure de la fin de l'observation.
- **Contact** : [Resp_name, Resp_Org, Resp_address, Resp_email]
- Resp_name correspond au nom du responsable de l'observation. Le personnel scientifique et technique doit signaler leur participation au sein d'une observation aux fins administratives.
 - Resp_Org est le nom de l'organisation qui est en charge de l'observation.
 - Resp_address correspond à l'adresse de l'organisation ou du responsable de l'observation.
 - Resp_email se réfère à l'adresse électronique du contact.

b) Métadonnées capteur

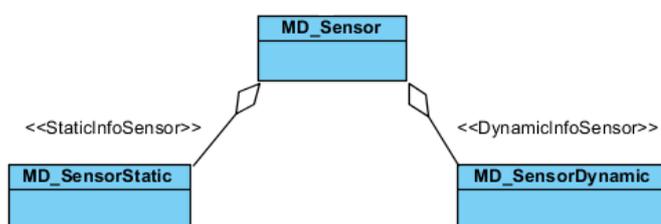


Figure 94 – Composition des Métadonnées capteur

```

...
<xs:group name="MD_Sensor">
<xs:annotation>
<xs:documentation>Group de métadonnées capteurs</xml:documentation>
</xs:annotation>
<xs:sequence>
<xs:group ref="MD_SensorStatic" minOccurs="0"/>
<xs:group ref="MD_SensorDynamic" minOccurs="0"/>
</xs:sequence>
</xs:group>
...

```

Tableau 22 - Extrait du schéma des métadonnées capteur

▪ Métadonnées capteur statiques

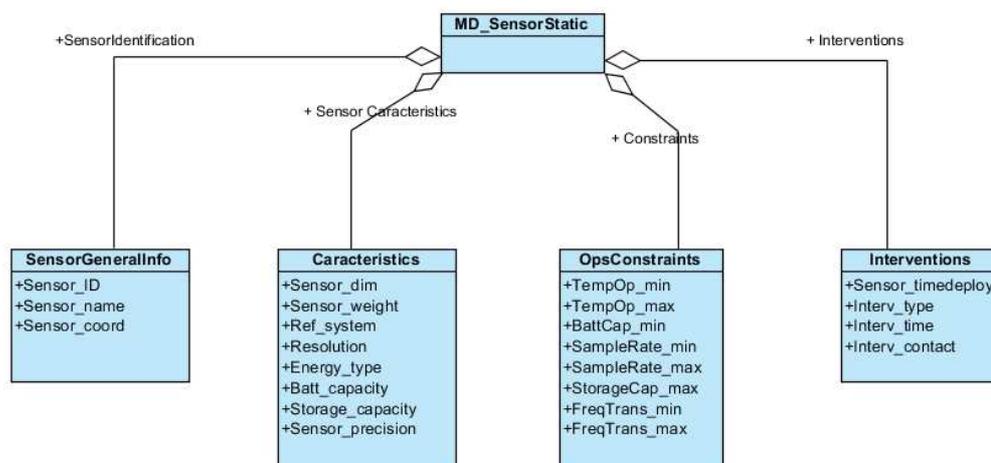


Figure 95 - Composition des Métadonnées Capteur Statique

```

...
<xs:group name="MD_SensorStatic">
<xs:annotation>
<xs:documentation>Group de métadonnées capteur statique </xs:documentation>
</xs:annotation>
<xs:sequence>
<xs:group ref="SensorGeneralInfo" minOccurs="0"/>
<xs:group ref="Characteristics" minOccurs="0"/>
<xs:group ref="OpsConstraints" minOccurs="0"/>
<xs:group ref="Interventions" minOccurs="0"/>
</xs:sequence>
</xs:group>
...
  
```

Tableau 23 – Extrait de la composition des métadonnées capteur statique

- MD_SensorStatic :

SensorGeneralInfo, Characteristics, OpConstraints, Interventions (Figure 95).

- **SensorGeneralInfo** – Information générale sur le capteur.

[Sensor_Id, Sensor_type, Sensor_coord]

- Sensor_Id est l'identificateur du capteur.
- Sensor_type correspond à la description du type de capteur (température, pression...).
- Sensor_coord, indique la position du capteur.

- **Characteristics** – Information concernant les caractéristiques physiques du capteur.

[Sensor_dim, Sensor_weight, Resolution, Energy_type, Batt_capacity, Storage_capacity]

- `Sensor_dim` correspond aux dimensions physiques du capteur.
 - `Sensor_weight` est le poids du capteur.
 - `Ref_system` est le système de référence utilisé pour la localisation du capteur.
 - `Resolution` est le minimum de changement que le capteur peut détecter dans la quantité qu'il mesure.
 - `Energy_type` correspond à la description du type d'énergie utilisé pour le capteur.
 - `Batt_capacity` indique la capacité des batteries (heures).
 - `Storage_capacity` indique la capacité de stockage de mémoire du capteur (octets).
 - `Precision`, indique la précision du capteur fournit par le fabricant (unités d'approximation).
- **OpsConstraints** – Information concernant les contraintes opérationnelles du capteur.

[Temp_OpMin,Temp_OpMax, Batt_CapMin, Sample_RateMax, Sample_RateMin, Storage_CapMin, Tx_FreqMax, Tx_FreqMin]

- `Temp_OpMin`, indique la température minimale d'opération du capteur
 - `Temp_OpMax`, indique la température maximale d'opération du capteur
 - `Batt_CapMin`, indique le niveau de batterie nécessaire pour le fonctionnement du capteur.
 - `Sample_RateMax` indique la fréquence maximale de mesure supportée par le capteur.
 - `Sense_RateMin`, indique la fréquence minimale de mesure supportée par le capteur.
 - `Storage_CapMin`, indique la capacité de stockage minimum avec laquelle le capteur peut être disponible.
 - `Tx_FreqMax`, représente la fréquence de transmission maximale que le capteur peut effectuer.
 - `Tx_FreqMin`, représente la fréquence minimale de transmission du capteur.
- **Interventions** – Information concernant les interventions réalisées sur les capteurs, voire des réparations, déplacements, etc. Décrit par :

[Sensor_timedeployment, Interv_date, Interv_type, Interv_contact]

- `Sensor_timedeployment`, date de déploiement du capteur dans la zone d'observation.

- Interv_type, type d'intervention effectué sur le capteur (réparation, changement, vérification, calibration...)
- Interv_date, date et heure de l'intervention.
- Interv_contact, responsable de l'intervention.

▪ Métadonnées capteur dynamiques

Ces métadonnées représentent l'information concernant les caractéristiques des capteurs qui varient au cours du temps, pendant une observation. Ces informations sont importantes pour comprendre l'état actuel du capteur.

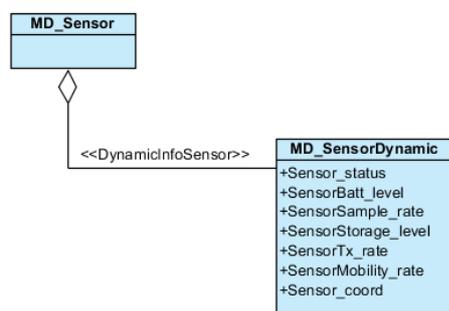


Figure 96 - Composition des Métadonnées Capteur Dynamique

```

.....
<xs:group name="MD_SensorDynamic">
<xs:annotation>
<xs:documentation>Métadonnées capteur dynamique </xs:documentation>
</xs:annotation>
<xs:sequence>
<xs:group ref="MD_SensorDynamic" minOccurs="0"/>
</xs:sequence>
</xs:group>
.....
  
```

Tableau 24 – Extrait du schéma de métadonnées capteur dynamique

- MD_SensorDynamic :

[Sensor_status, SensorBatt_level, SensorSample_rate, SensorStorage_level, SensorTx_rate, SensorMobility_rate, Sensor_coord]

- Sensor_status, représente l'état d'opération du capteur (on/off).
- SensorBatt_level, niveau actuel d'énergie disponible.
- SensorSample_rate, fréquence de prise de mesure actuelle du capteur.
- SensorStorage_level, capacité de stockage actuel au sein du capteur.
- SensorTx_rate, fréquence actuelle de transmission du capteur.
- SensorMobility_rate, taux de mobilité actuelle du capteur.
- Sensor_coord, position actuelle du capteur.