Etude dynamique de la qualité de l'information et des données d'un système d'information complexe

Ion George Todoran

THÈSE / Télécom Bretagne
sous le sceau de l'Université européenne de Bretagne

pour obtenir le grade de Docteur de Télécom Bretagne
En accréditation conjointe avec l'Ecole doctorale Matisse
Mention : Traitement du signal et Télécommunications

présentée par
Ion-George Todoran
préparée dans le département Image et Traitement de l'Information (ITI)
Laboratoire Labsticc

Étude dynamique de la qualité de l'information et des données d'un système d'information complexe

Thèse soutenue le 11 décembre 2014
Devant le jury composé de :

Frédéric Dambreville
Expert (HDR), DGA-EORD – Bruz / président

Éloi Bossé
Chercheur, École Polytechnique de Montréal, Canada / rapporteur

Didier Dubois
Directeur de recherche, Université Paul Sabatier - Toulouse / rapporteur

Ali Khenchaf
Professeur, ENSTA Bretagne – Brest / examinateur

Marie-Véronique Serfaty
Ingénieur, DGA-MRIS – Bagneux / invitée

Jean-Marc Le Caillec
Professeur, Télécom Bretagne / invité

Basel Solaiman
Professeur, Télécom Bretagne / invité

Laurent Lecornu
Maître de Conférences, Télécom Bretagne / invité

N° d'ordre : 2014telb0343

Sous le sceau de l'Université européenne de Bretagne

Télécom Bretagne

En accréditation conjointe avec l'Ecole Doctorale Matisse

Ecole Doctorale – MATISSE

Étude dynamique de la qualité de l'information et des données d'un système d'information complexe

---

Thèse de Doctorat

Mention : Traitement du signal et Télécommunications

Présentée par **Ion-George Todoran**

Département : Image et Traitement de l'Information

Laboratoire : Lab-STICC UMR CNRS 6285

Directeurs de thèse : Jean-Marc Le Caillec et Ali Khenchaf

Soutenue le 11 décembre 2014

Jury :

M. Frédéric Dambreville : Expert DGA (HDR) à DGA-EORD (Président du jury)
M. Éloi Bossé, Chercheur à l'École Polytechnique de Montréal, Canada (Rapporteur)
M. Didier Dubois, Directeur de recherche CNRS à l'Université Paul Sabatier, Toulouse (Rapporteur)
M. Jean-Marc Le Caillec, Professeur à Télécom Bretagne (Directeur de thèse, invité)
M. Ali Khenchaf, Professeur à l'ENSTA Bretagne (Co-directeur de thèse, examinateur)
M. Laurent Lecornu, Maître de Conférences à Télécom Bretagne (Invité)
Mme. Marie-Véronique Serfaty : Ingénieur, RDS « Ingénierie de l'Information et Robotique » à DGA-MRIS (Invitée)
M. Basel Solaiman, Professeur à Télécom Bretagne (Invité)

Table des matières

Table des figures

Liste des tableaux

Introduction

Motivations

Si au III^{e} siècle avant Jésus-Christ, on considérait que la bibliothèque d'Alexandrie renfermait la totalité du savoir humain, la quantité gigantesque d'informations disponible aujourd'hui ne peut être que grossièrement quantifiée. Une estimation récente [Mayer-Schonberger 13] montre qu'une division de la quantité totale d'informations entre tous les Terriens donnerait à chaque personne une quantité d'informations 320 fois supérieure à celle d'Alexandrie, c'est-à-dire de 1200 exaoctets. Une autre estimation de cette même étude, cette fois-ci plus visuelle, décrit la masse d'informations actuelle comme l'équivalent de 5 piles de CD capable de relier la Terre à la Lune. Cette explosion de la quantité d'information a été générée par les progrès inimaginables au cours du 20ème siècle de l'électronique et de l'informatique. En effet, à partir des années 70, on a commencé à parler d'une ère de l'information [Mason 78]. Les estimations récentes montrent que la quantité d'informations générées dans le monde augmente de 30% chaque année [Dong 09]. De plus, aujourd'hui presque chaque personne a accès à un ordinateur avec une connexion Internet. Ainsi, il est connecté à un ensemble très grand et très varié de sources d'informations lui proposant de nombreux services en ligne, payants ou en accès libre. Avec ces ressources mises à sa disposition, il est très simple pour chacun de créer son propre système d'information. Mais cette facilité d'accéder aux informations cache une dimension peu visible de ces dernières : la confiance dans les sources d'information.

Le développement des réseaux sociaux, le déploiement d'un nombre de plus en plus élevé de capteurs (biologiques, d'observation de l'environnement : radar, sonar, lidar, électrique, etc.) font que la quantité des données à traiter continue d'augmenter. On vit une époque de « Big data », un déluge numérique. Ainsi, le télescope Sloan Digital Sky Survey installé en 2000 à New Mexico, États-Unis, a collecté lors des ces premières semaines de fonctionnement une plus grande quantité de données que celle de toute l'histoire de l'astronomie. Un autre exemple est celui du groupe américain Wal Mart, spécialisé dans la grande distribution qui réalisait en 2010 un million de transactions par heure, alimentant ainsi ses bases de données de plus de 2.5 pétaoctets [Economist 10]. Grâce aux technologies actuelles, ces données pourraient être utilisées pour extraire des informations sur le comportement des utilisateurs. De plus, presque chaque aspect de notre vie personnelle et professionnelle est très souvent numérisé, enregistré, traité et finalement échangé par l'intermédiaire des ordinateurs, des dispositifs intelligents (téléphones, tablettes, etc.), des caméras vidéo, des dispositifs RFID, de l'Internet, des réseaux sociaux, des courriels, des différentes applications (e-commerce, dans le cadre de l'entreprise, etc.). Et ces données sont, potentiellement, instantanément accessibles en tout moment et en tout lieu.

[Gates 99] dans son livre « *Business @ the Speed of Thought* » fait l'affirmation suivante :

> *The most meaningful way to differentiate your company from your competition, the best way to put distance between you and the crowd, is to do an outstanding job with information. How you gather, manage and use information will determine whether you win or lose.* [1]

Ainsi, les organisations se sont rendues compte que les données et les informations sont une ressource d'une valeur potentielle incroyable. En conséquence, afin de rester compétitives, les organisations ont commencé à beaucoup investir dans leurs systèmes informatiques. Malgré la crise

1. En traduction : « *La façon la plus signifiante de différencier ton entreprise de ta compétition, la meilleure manière de prendre de l'avance vis-à-vis des autres, est de faire un travail extraordinaire avec les informations. La façon avec laquelle tu les collectes, les gères et les utilises va déterminer si tu vas gagner ou si tu vas perdre.* »

économique, les organisations continuent à dépenser en 2014 de plus en plus dans les technologies de l'information et de la communication (TIC), selon une étude de Gartner [Kanaracus 14].

Cependant, ces données stockées sont inutiles si elles ne sont pas transformées en information utile. Ainsi, les algorithmes de traitement de données ont aussi connu un développement incroyable. Prenons l'exemple du décodage du génome humain, qui nécessite l'analyse de 3 milliards de paires de base. En 2003, quand ce processus a été pour la première fois implémenté, dix ans ont été nécessaires pour compléter le décodage. À l'heure actuelle, cela prendrait moins d'une semaine. Aujourd'hui, de volumes gigantesques de données peuvent être traitées dans des temps raisonnables.

Les systèmes de traitement de l'information continuent d'évoluer. Ils sont passés d'un système d'information utilisé par un seul utilisateur dans une seule tâche à un système d'information utilisant une multitude de sources de données et réalisant une intégration de services.

Pourtant, cette augmentation de la quantité de données disponibles, accompagnée par l'emploi de techniques de plus en plus complexes peut avoir dans beaucoup de cas un effet inverse sur la qualité des informations extraites. Même si les nouvelles technologies d'information et de communications continuent d'offrir de plus en plus de possibilités de traitement de l'information, leur simple utilisation ne va pas résoudre le problème de la qualité des informations fournies à l'utilisateur final.

Suite à ces changements (augmentation de la quantité de données, l'Internet, techniques de communication, etc.) les organisations sont confrontées à de nouveaux problèmes sur la quantité (**et la qualité**) des données/ des informations/ des connaissances disponibles. Très souvent dans les organisations, les personnes qui sont en charge de ces problèmes n'ont pas de réponse sur la quantité et la qualité des données et des informations dont l'organisation dispose. Cependant, les données et les informations sont indispensables dans le processus de prise de décisions.

Si on passe du côté des utilisateurs des systèmes d'informations, on peut voir qu'ils ont besoin des informations pertinentes, exactes, complètes, consistantes, actualisées et présentées d'une façon facilement compréhensible. Mais le passage des données vers des informations n'est pas évident et beaucoup d'organisations peuvent être caractérisées comme étant riches en données et pauvres en informations. Les principales causes de cette situation sont que [Pautke 02] :

- La plupart des organisations n'ont pas la culture de quantifier et de qualifier les données, les informations et les connaissances dont elles disposent ;
- Les managers prennent les décisions en utilisant leur intuition et dans beaucoup de cas n'ont pas confiance ou ont peur des décisions qui leurs sont proposées par un système d'information - ils n'ont pas confiance à la seule vue d'une valeur sans aucune autre explication ;
- Dans la plupart des organisations, les processus de prise de décisions sont ad-hoc et non-contrôlés.

Au cours de ces dernières décennies la plupart des travaux de recherche ont été orientés vers le développement et l'implémentation de nouveaux algorithmes de traitement de données. Ces algorithmes sont devenus de plus en plus spécialisés, avec de très bonnes performances dans le cas spécifique de leur application, malheureusement cela s'est accompagné d'une complexité de plus en plus grande. En même temps, très peu de travaux de recherche ont été menés sur la qualification de ces algorithmes, et du système de traitement de l'information en général, afin d'évaluer leurs capacités dans des situations réelles, dynamiques et complexes. Les algorithmes de traitement de données sont habituellement analysés et validés en utilisant des simulations et/ou des tests dans un environnement contrôlé. Malheureusement, la validation du système d'information dans un contexte réel et dynamique ne peut se faire qu'après sa mise en application. À l'heure actuelle, il n'existe pas une méthodologie permettant d'évaluer automatiquement la qualité des informations proposées par un système d'information.

Les problèmes de qualité de données et d'information ont eu comme conséquences des pertes énormes pour les organisations. Une statistique présentée dans [English 09] montre que 122 organisations ont accumulé des pertes totales d'une valeur de 1212 milliards de dollars. À lui seul, le système "Mars Climate Orbiter" développé par la NASA a gâché 125 millions de dollars (sans prendre en compte les pertes de mesures scientifiques) à cause d'une mauvaise transformation des

unités de mesure anglaise en unités métriques [2].

L'évaluation de la qualité de l'information ou plus généralement l'évaluation des performances d'un système d'information est un problème interdisciplinaire faisant référence aux domaines de recherche comme le design de systèmes (architecture, développement, etc.), l'analyse de données, les techniques de simulation ou encore l'inférence statistique. Définir une telle méthodologie reste un problème très difficile, surtout à cause de la nature intangible de l'information. Dans l'industrie, il existe depuis des décennies des mesures de performances et de productivité pour des produits tangibles, physiques, comme par exemple *le nombre d'automobiles produits par homme par heure*. Ces mesures sont dans la plupart de cas des indicateurs de productivité faisant la liaison entre la sortie et l'entrée de la chaîne de production.

La complexité des systèmes d'information continue à augmenter en ajoutant de plus en plus de fonctionnalités. Ainsi, l'ajout de nouveaux modules de traitement ou la mise à jour d'existants est devenu une pratique courante. En conséquence, il est nécessaire pour le processus d'évaluation de la qualité de l'information de pouvoir prendre en compte ces changements, de préférence d'une façon plus ou moins automatique. De plus, nous sommes persuadés que pour l'utilisateur d'un système d'information la qualité joue un rôle aussi important que la valeur en elle-même de l'information. Ainsi, il y a un vrai besoin d'équiper chaque système d'information d'une méthodologie d'évaluation de la qualité de l'information. Malheureusement, à l'heure actuelle une telle méthodologie n'existe pas. Les propositions existantes considèrent le système comme une « boîte noire » et analysent la qualité en entrée et en sortie du système. Malheureusement, cette vision restreinte du système impose l'utilisation de formulaires comme seule possibilité d'évaluation. Ce processus d'évaluation demande beaucoup de temps, est très subjectif et réalise une évaluation moyenne du système, c'est-à-dire l'utilisateur donne son avis sur l'ensemble de ses expériences d'utilisation du système et non sur une seule. En conséquence, une autre limitation est l'impossibilité de fournir la qualité individuelle de chaque information proposée.

Dans cette thèse, nous proposons une nouvelle méthodologie d'évaluation de la qualité de l'information. Afin de pouvoir couvrir un spectre large d'application, cette méthodologie est développée dans un contexte général d'un système d'information quelconque sans *a priori* sur une application précise. Le point de départ de cette méthodologie est l'exploitation de la structure interne du système d'information et de son architecture. Ainsi, une des plus importantes caractéristiques de cette méthodologie est sa flexibilité à s'adapter aux évolutions du système. Si dans la plupart des travaux de recherche les notions de données et d'information sont traitées de façon équivalente, cette méthodologie fait appel à leurs caractéristiques individuelles afin de pouvoir définir la notion de qualité. Ainsi, dans cette méthodologie les caractéristiques sémantiques des données et des informations servent à choisir les dimensions de qualité adaptées à l'évaluation de la qualité des données et respectivement, des informations.

Cette méthodologie se voit d'abord comme un outil permettant aux utilisateurs d'avoir « *un degré de confiance* » dans l'information que son système d'information lui propose. Et si l'utilisateur est circonspect sur cette valeur, alors la méthodologie va lui permettre d'avoir une explication sur sa provenance. En plus, cette méthodologie va pouvoir être utilisée par les analystes et les designers des systèmes afin d'optimiser les performances du système d'information en analysant les améliorations et les dégradations de la qualité à travers le système.

Ainsi, l'objectif final de ce travail de doctorat est l'évaluation de la confiance que l'utilisateur peut avoir dans le système d'information utilisé. Autrement dit, nous envisageons de faire la liaison entre l'évaluation des performances d'un système d'information (évaluation objective) et l'adaptation du système d'information aux besoins de(s) utilisateur(s).

2. source CNN : http ://edition.cnn.com/TECH/space/9909/23/mars.orbiter.03/

Organisation du mémoire

Ce mémoire de thèse est divisé en trois parties. Dans la première partie, l'état de l'art de l'évaluation de la qualité des données et de l'information ainsi que le contexte de ce travail sont présentés. Dans le premier chapitre de cette partie, nous proposons des définitions pour les notions de base : données, information et connaissance dans le contexte général d'un système d'information complexe utilisé dans le cadre d'un processus d'aide à la décision. Ensuite, dans le deuxième chapitre, l'évaluation de la qualité des données stockées dans de bases de données est présentée. Dans le troisième chapitre, nous réalisons un état de l'art sur la définition et l'évaluation de la qualité de l'information suite à un tour d'horizon dans trois domaines d'applications des systèmes d'informations : management de systèmes d'information, systèmes d'information sur le Web et systèmes de fusion d'informations. À la fin de ce chapitre, nous adressons un bilan global et des critiques sur les méthodologies précédemment présentées.

Dans la deuxième partie, nous proposons une nouvelle méthodologie d'évaluation de la qualité de l'information. Par rapport aux autres méthodologies proposées dans la littérature, le système d'information est décomposé en modules de traitement élémentaires afin de pouvoir suivre l'évolution de la qualité à travers le système. Pour cela nous définissons, dans le quatrième chapitre, deux visions différentes de la notion de qualité. La première, appelée qualité locale, exprime la qualité en sortie d'un module de traitement du système d'information. La deuxième, appelée qualité globale, exprime la qualité du système d'information dans sa globalité. L'idée de base de cette double vision de la notion de qualité est que celle locale, propre à un module de traitement, est beaucoup plus facile à estimer que la qualité globale. Le point central de cette méthodologie est la modélisation de l'influence d'un module de traitement sur la qualité. Dans le cinquième chapitre, nous introduisons un nouveau concept, celui de la fonction de transfert de la qualité. Celle-ci est inspirée de la notion de fonction de transfert en traitement du signal, qui fait la liaison entre le signal en sortie d'un module et celui en son entrée. Ainsi, la fonction de transfert de qualité a pour rôle d'évaluer la qualité en sortie d'un module de traitement en fonction de la valeur et de la qualité de l'information à son entrée. Dans le chapitre six et dernier de cette partie, nous réalisons la liaison entre la qualité locale et la qualité globale grâce à la notion de fonction de transfert de la qualité. Par conséquent, nous montrons qu'il est possible d'évaluer automatiquement la qualité globale du système entier.

La troisième partie de cette thèse est consacrée à la validation de notre méthodologie. Comme nous voulons que la méthodologie soit généralement applicable pour tout type de système d'information, nous avons choisi deux applications différentes. La première est un système de reconnaissance automatique de cibles radar. La deuxième est un système d'aide au codage des actes médicaux.

Dans la quatrième partie, nous présentons une conclusion générale qui dresse le bilan des travaux réalisés et les perspectives offertes par ce travail de recherche. Il faut mentionner qu'aujourd'hui le domaine d'évaluation de la qualité de l'information est encore naissant et que dans les prochaines années de plus en plus de projets de recherche y seront consacrés.

« Even though quality cannot be defined, you know what it is. »

Zen and the Art of Motorcycle Maintenance : An inquiry into Values
Robert M. Pirsing 1974

PREMIÈRE PARTIE : QUALITÉ DES DONNÉES ET DE L'INFORMATION

Cette partie est organisée autour de trois chapitres. Le premier présente le contexte général de ces travaux de recherche, c'est-à-dire systèmes d'information complexes d'aide à la décision. Dans le cadre de ce contexte et afin de bien définir la notion de qualité pour les données et pour les informations, nous proposons des définitions pour les notions de données, d'information et de connaissance. Le deuxième chapitre fait l'objet d'un état de l'art sur la qualité des données. Ensuite, le troisième et dernier chapitre de cette partie, présente un état de l'art sur la qualité de l'information en considérant trois domaines différents d'application des systèmes d'information.

TELECOM Bretagne

1 Contexte général

Dans ce chapitre, nous définissons le contexte de ce travail de recherche ainsi que les notions centrales qui sont les **données**, l'**information** et la **connaissance**. Dans le paragraphe 1.1, la notion de système d'information complexe (SIC) est présentée en parallèle avec d'autres types de systèmes afin de pouvoir tracer les frontières de notre problématique : **système d'information complexe pour l'aide à la décision**.

Ensuite, le paragraphe 1.2 définit ce qui est entendu par contexte d'application. C'est une notion très importante qui intervient dans la plupart des définitions énoncées. Puis, le paragraphe 1.3 présente les définitions pour les notions de données, information et connaissance dans le cadre d'un système d'information complexe d'aide à la décision. Quant au paragraphe 1.4, il introduit la notion de qualité afin de la mettre en correspondance avec les données et l'information.

1.1 Système d'information complexe (SIC)

Avant de présenter ce qu'est un système d'information, les acteurs qui interagissent avec celui-ci seront définis ce qui permettra de définir un système d'information. Enfin, le système d'information complexe, qui est le système qui nous intéresse, sera présenté.

1.1.1 Les acteurs d'un système d'information

Un système d'information est en contact permanent avec une multitude d'utilisateurs ayant différents rôles, responsabilités et/ou expertises dans le cadre d'une organisation. Il existe trois types principaux d'acteurs : l'utilisateur, l'expert et l'analyste du système.

> **Définition 1 : Un utilisateur** est une personne (un humain) qui fait appel à un système d'information afin d'obtenir des informations nécessaires pour produire quelque chose, pour résoudre un problème ou bien effectuer une tâche liée à son activité quotidienne [Turban 05].

Les utilisateurs sont les bénéficiaires directs du déploiement du système d'information. En plus d'être les initiateurs des requêtes vers le système d'information, les utilisateurs sont aussi responsables pour déterminer [1] :

- les problèmes à résoudre,
- les opportunités à exploiter,
- les besoins à satisfaire,
- les contraintes à surmonter par le système,
- si le système est facile / difficile à utiliser.

Très souvent il existe une confusion entre l'utilisateur d'un système d'information et l'expert du domaine. Bien sûr il est possible que l'utilisateur soit la même personne que l'expert. Mais pour

1. Selon le document *Introduction to information systems* de l'Open University of Malayesia : `http://www.oum.edu.my/oum/v3/download/CBAD2103.pdf`, consulté le 15/08/2014

éliminer toute confusion, une définition de l'expert est donnée. Dans la littérature, il n'y a pas une définition standard d'un expert. La performance de ces décisions et son niveau d'expertise dans son domaine sont les critères de base dans l'identification d'un expert.

> **Définition 2 : Un expert** est une personne (un humain) qui possède des connaissances des jugements, des expériences et des méthodes spécifiques ainsi que la capacité d'appliquer ceux-ci afin de donner des conseils et de résoudre des problèmes [Turban 05].

Ainsi, suite à cette définition, un expert est capable de résoudre un problème et d'atteindre un meilleur niveau de performance qu'une personne ad-hoc. Cependant, les experts sont relatifs à un domaine spécifique, c'est-à-dire ils sont spécialisés dans un domaine restreint.

Deux types particuliers d'experts peuvent être identifiés : le designer du système et le développeur du système [1]. Le designer du système est un expert dans le domaine technique qui a pour rôle de concevoir un système qui répond aux besoins des utilisateurs. Un développeur du système est un expert dans le domaine technique qui a la responsabilité de développer, tester et délivrer un système répondant aux spécifications fournies par les designers du système.

Donc les développeurs du système utilisent des outils technologiques (logiciels, langages de programmation, méthodologies, etc.) pour développer le système d'information. En fonction des ressources humaines allouées par l'organisation pour le déploiement du système d'information, il se peut que le designer du système soit également le développeur du système.

Une autre entité très importante dans le déploiement du système d'information est l'**analyste du système**.

> **Définition 3 : L'analyste du système** est une personne responsable de la planification, l'analyse et l'implémentation du système d'information. Il a l'expertise nécessaire pour pouvoir cordonner l'activité des autres acteurs : d'élargir la vision des utilisateurs du système d'information, de s'assurer que les designers et les développeurs du systèmes ont les connaissances techniques et technologiques nécessaires pour répondre aux besoins de l'organisation.

Maintenant, nous présenterons ce qu'est un système d'information complexe utilisé dans le processus d'aide à la décision. Premièrement, la notion de système d'information sera définie. Puis, les caractéristiques indiquant la présence d'un système complexe seront présentées.

1.1.2 Système d'information

Les premières notions qui doivent être définies sont celles de « système » et de « architecture d'un système ». Ci-dessous sont présentées leurs définitions selon [Xiao 09] :

> **Définition 4 : Un système** est un ensemble d'éléments interdépendants qui interagissent entre eux de façon organisée et formant un ensemble unique. Parmi les éléments constituant un système, se retrouvent des produits (hardware, software, firmware), des processus, des humains, des informations, des techniques, des facilités, des services et d'autres éléments de support [INCOSE 04].

> **Définition 5 : L'architecture d'un système** est un arrangement d'éléments et de sous-systèmes qui allouent des fonctions afin de respecter les spécificités techniques requises par le système.

Très souvent, il y a une confusions entre la notion de système d'information, celle de système informatique et celle de système expert. Nous présentons en parallèle ces trois types de systèmes, dans la figure 1.1.

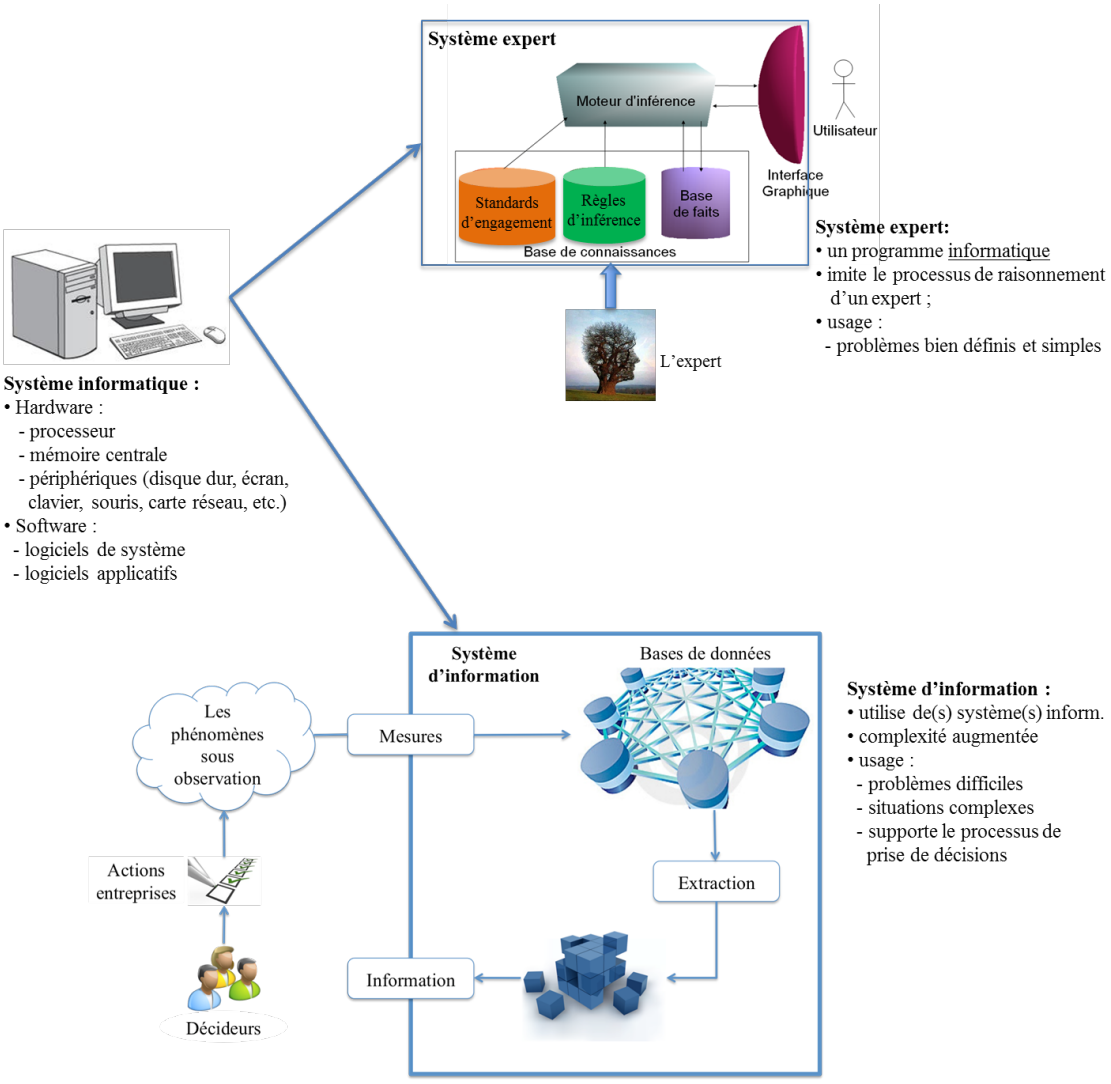


FIGURE 1.1: Différents types de systèmes : informatique, expert et d'information

L'objectif d'un système informatique est **d'automatiser le traitement des données / informations**. Il est composé de deux composants principaux : le matériel (le « hardware ») et les logiciels (le « software »). Les ordinateurs personnels (les PC), les tablettes, les téléphones intelligents (les « smartphones »), etc. sont quelques exemples de systèmes informatiques très communs dans notre vie quotidienne. Les systèmes informatiques servent comme éléments de base pour le développement des autres types de systèmes : expert ou d'information.

Un système expert est **un programme informatique** qui imite le processus de raisonnement d'un expert humain afin de répondre à un problème donné. Dans la plupart des cas, un système expert est composé d'au moins trois parties [Hall 92] :

1. *Un moteur d'inférence* : il utilise les connaissances du domaine et les informations acquises sur un problème donné afin de donner la solution expert ;
2. *Une base de connaissances* : elle contient les connaissances du domaine expert sous la forme des règles et des faits ;
3. *Une interface* : utilisée pour communiquer les informations aux utilisateurs ou vers d'autres systèmes.

Un système d'information a pour rôle de collecter, traiter, stocker, analyser et disséminer l'information pour un objectif spécifique [Turban 05]. De cette définition, on peut immédiatement observer qu'un système d'information est utilisé pour offrir le support nécessaire à une autre entité (un utilisateur ou même un autre système) qui ne fait pas partie du système d'information lui-même. Les fonctionnalités et le comportement du système d'information sont déterminés par rapport aux besoins de son environnement, c'est-à-dire des entités avec lesquelles il interagit.

Une vision simplifiée d'un système d'information est présentée dans la figure 1.1 (adaptée de [Kwan 96]). Ainsi, un système d'information contient une première partie, appelée déclarative, qui est responsable de la description du monde réel (phénomènes sous observation). Ce sous-système réalise une numérisation de l'environnement sous observation et enregistre les données résultantes (suite à des opérations de mesure) dans des bases de données. Ces bases de données peuvent être locales ou distribuées. La deuxième partie d'un système d'information est sa composante opérationnelle, responsable de l'extraction de la partie descriptive des informations utiles qui est destinées aux utilisateurs finaux (les décideurs). Le processus d'extraction d'informations utiles fait appel à divers traitements numériques, modifiant et transformant les données pour les transformer en information. Ensuite, les décideurs sous l'influence des informations reçues vont entreprendre des actions, qui à leur tour vont impacter le monde sous observation.

Un système d'information est conçu pour répondre aux divers besoins de(s) utilisateur(s). Ainsi, si un système expert est conçu pour une tâche bien précise (par exemple le diagnostic d'une maladie) avec l'objectif de remplacer un expert, un système d'information répond à des problématiques beaucoup plus complexes (par exemple l'estimation et la visualisation de la propagation d'une maladie).

Les systèmes d'information sont devenus le cœur de toute organisation moderne dans tous les domaines d'activité. Par exemple, les systèmes d'information peuvent être utilisés dans le management d'une entreprise pour [O'Brien 11] :

- soutenir les processus et les opérations de l'entreprise ;
- aider les managers et les employés dans leur processus d'aide à la décision ;
- soutenir les stratégies afin d'obtenir des avantages compétitifs.

1.1.3 Système d'information complexe

Grâce aux développements technologiques, de plus en plus de données et de possibilités de traitements sont devenues accessibles. Ainsi, les systèmes d'information continuent d'évoluer, devenant de plus en plus *complexes*. Cependant, les organisations sont des entités dynamiques avec des besoins en terme d'information de plus en plus élevés. Ainsi, les nouvelles possibilités offertes

par l'évolution du système d'information se traduisent très souvent en nouveaux produits et/ou services.

Définir la complexité est un problème difficile. Un système complexe est caractérisé par des relations de causes à effets qui sont seulement *cohérentes en rétrospection* [Kurtz 03]. Cela veut dire que les sorties d'un système complexe peuvent être perçues mais pas prédites. Même si les mêmes résultats peuvent être susceptibles de réapparaitre après un certain temps dans un certain contexte, ce caractère reproductible ne peut pas être généralisé.

Ainsi, selon [Kurtz 03] les sorties d'un système complexe peuvent être caractérisées par des propriétés émergentes. Ces propriétés ne sont pas issues seulement des propriétés des composants du système, mais aussi des nombreuses relations entre ces composants. Cette vision est identique à celle de [Simon 62] qui présente le système complexe comme étant un système construit à partir d'un très grand nombre de parties qui interagissent d'une manière non-simple. Les propriétés des parties (des sous-systèmes) et les lois gouvernant leurs interactions font de l'évaluation des performances du système entier un problème non-évident.

L'étude de la complexité des systèmes d'information est dans la plupart des cas inspirée par le domaine de l'architecture logicielle [Godfrey 13]. Il existe plusieurs approches formelles capables d'analyser la complexité intrinsèque d'un système logiciel, comme par exemple l'utilisation du nombre cyclomatique associé à un graphe dérivé de la structure de l'architecture ou l'analyse de diagrammes de classe (en utilisant le formalisme UML). Malheureusement, ces méthodes s'appuient sur des descriptions d'architectures détaillées et sur des calculs complexes, qui les rendent difficilement applicable dans le cas d'un système d'information [Caseau 07].

Les systèmes complexes hiérarchiques et décomposables ont tendance d'évoluer plus rapidement que les systèmes intégrés [Ethiraj 04]. Une structure modulaire du système s'avère très utile quand le système tend à devenir très grand et que les interdépendances entre les éléments du système augmentent énormément. Ainsi, selon [Ethiraj 04], une modélisation modulaire du système est un moyen de gérer la complexité.

Une autre caractéristique d'un système complexe modulaire est la possibilité que ses modules soient utilisés dans plusieurs applications indépendantes c.à.d. fournissant des informations à plusieurs entités (utilisateurs ou d'autres systèmes).

Un cas particulier de systèmes complexes est représenté par le systèmes de systèmes [2]. Figure 1.2 présente l'exemple du système AGS Core de l'OTAN [3]. Ce genre de système, d'une complexité très élevée, contient un nombre important de sous-systèmes qui à leur tour peuvent être caractérisés comme étant complexes. Ainsi, les systèmes de systèmes sont un ensemble de systèmes qui partagent leurs ressources et capacités afin d'offrir des nouvelles fonctionnalités et des meilleures performances par rapport à la somme de contributions individuelles des systèmes.

Dans le cadre de cette thèse, le système complexe sera vu comme un système composé d'un nombre important de modules coopérant entre eux. De plus, du fait de l'indépendance des modules, ceux-ci peuvent être modifiés ou mise à jour en fonction des évolutions technologiques. En conséquence, le système et ses performances sont amenés à évoluer dans le temps.

Dans le chapitre 4 dédié à l'étude de la qualité d'un système d'information, le système d'information sera décomposé en modules élémentaires. Dans [Baldwin 00], un module est défini comme étant une unité composée des éléments ayant des connections très étroites entre eux et des connections faibles avec les éléments des autres unités. Notre définition d'un module élémentaire est la suivante :

> **Définition 6 : Le module élémentaire** est l'entité morpho-fonctionnelle d'un système d'information, constituant l'élément de base de l'architecture du système et ayant un comportement et un fonctionnement stable par rapport aux autres modules.

2. En anglais : System of Systems (SoS)
3. `http://www.nagsma.nato.int/Pages/AGS_General_Information.aspx`, consulté le 03 septembre 2014

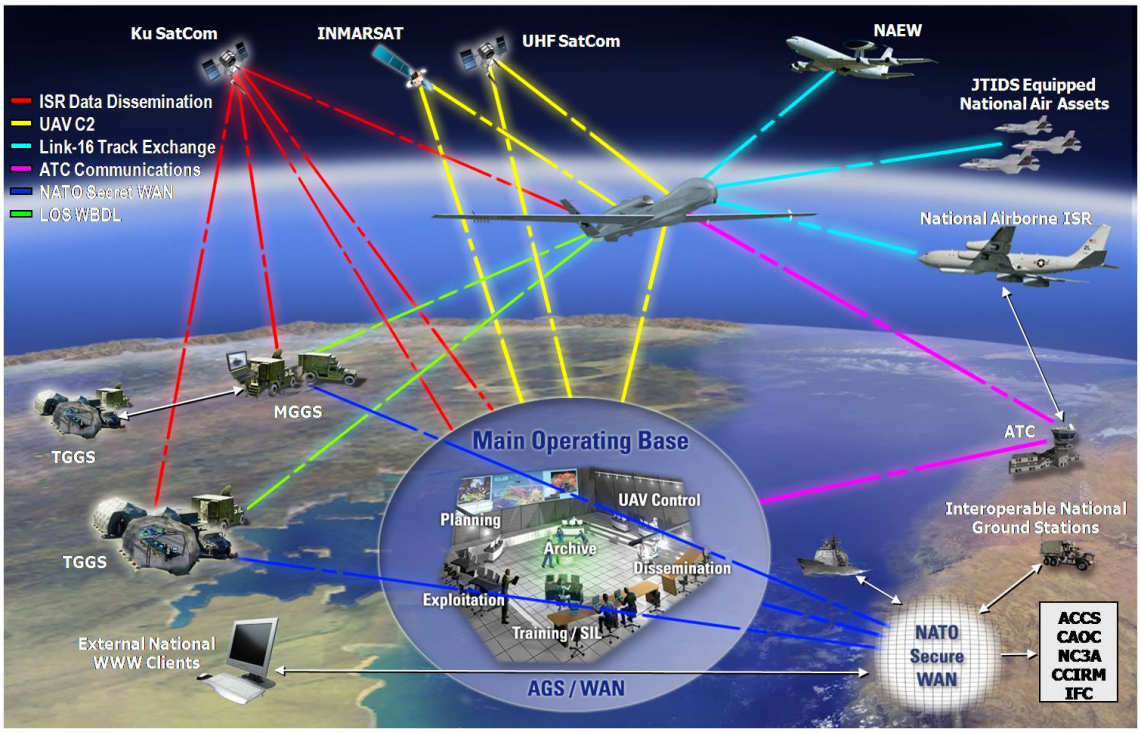


FIGURE 1.2: Exemple de système de systèmes : AGS Core de l'OTAN

1.2 MODÉLISATION DU CONTEXTE

Pour les êtres humains, le processus de transfert des informations et des connaissances par la communication orale ou écrite est favorisé par la compréhension, implicite ou explicite, des expériences ou des situations communes dans la vie quotidienne. La richesse du langage partagé par les interlocuteurs, les conventions sociales ou encore la compréhension mutuelle sur le fonctionnement des choses dans le monde sont parmi les plus importants catalyseurs de la communication. Toutes ces informations, complémentaires à la communication directe définissent *le contexte*. Pourtant, quand il s'agit d'une communication entre un humain et une machine, le contexte n'est pas partagé d'une manière explicite. Donc, une des directions de recherche actuellement très actives est la définition formelle et l'intégration du contexte dans les applications informatiques. Une définition générique du contexte a été proposée par [Dey 01] :

> **Définition 7 : Le contexte** est composé de toute information qui pourrait être utilisée pour caractériser les situations d'une entité. Par entité, est entendu une personne, un emplacement ou un objet considéré comme *pertinent* pour l'interaction entre un utilisateur et une application, en incluant l'utilisateur et l'application eux-même.

Dans cette définition, un des plus importants mots est le mot « pertinent ». En effet, le contexte est composé de tous les éléments complémentaires du message de communication directe et de toute nature qui sont pertinents à la communication.

Ce genre de définitions générales est très difficile à être implémenté dans la pratique, ainsi d'autres chercheurs ont proposé d'autres définitions du contexte. [Schilit 94] identifie trois catégories du contexte :

– *Le contexte informatique* : le réseau, le coût de communication, la bande passante, les périphériques, les postes de travail ;

– *Le contexte de l'utilisateur* : le profil, l'emplacement, les autres personnes de l'environnement, la situation sociale, etc. ;
– *Le contexte physique* : le niveau de bruit, les conditions du trafic, les conditions météo, etc.

La dimension temporelle du contexte est très importante pour toutes les applications et comme elle est difficilement intégrable dans les trois catégories de contexte, [Chen 00] en a proposée une quatrième :

– *Le contexte temporel* : le moment exact du temps : l'heure, le jour, le mois, l'année, etc.

En utilisant les définitions susmentionnées, le contexte peut être modélisé par un ensemble fini de *paramètres de contexte* (C_i). Chacun de ces paramètres (C_i) possède un domaine de définition noté $dom(C_i)$. Dans le cas le plus général, le domaine $dom(C_i)$ est un ensemble de valeurs infini dénombrable.

Si on considère une application particulière X, son contexte sera défini par [Stefanidis 11] :

> **Définition 8 : Le contexte d'une application** particulière X, noté $CA(X)$ est composé de l'ensemble de ces paramètres de contexte : $CA(X) = \{C_1, C_2, \ldots, C_n\}$, $n \geq 1$.

Représentation du contexte :

Pour que la modélisation du contexte soit suffisamment flexible pour permettre une spécification facile du contexte, les paramètres de contexte peuvent être définis à plusieurs niveaux, formant un schéma hiérarchique [Ye 08] and [Stefanidis 11], c'est-à-dire pour un même paramètre de contexte il existe plusieurs niveaux de détail. Soit un paramètre de contexte C pouvant être défini à $m \geq 1$ niveaux, son schéma hiérarchique sera défini par $N = \{N_1, N_2, \ldots, N_m\}$, avec N_1 le niveau le plus détaillé et N_m le niveau le plus général. Sur cette structure granulaire, on peut également définir un opérateur d'ordre partiel $\prec$ permettant d'ordonner les niveaux N_i par rapport à leur niveau de détail. Grâce à cet opérateur, pour chaque schéma hiérarchique N on a $N_1 \prec N_2 \prec \ldots \prec N_m$. Chaque niveau N_j de chaque paramètre de contexte C_i a un domaine propre de définition de valeurs $dom_{N_j}(C_i)$. De plus, comme indiqué dans [Stefanidis 11], il est préférable que le niveau le plus général ait une seule valeur : $dom_{N_m}(C_i) = \{\text{ALL}\}$. Cette valeur par défaut indique une incertitude totale sur la vraie valeur de ce paramètre de contexte. Maintenant, un exemple est présenté pour mieux visualiser les notions introduites jusqu'à présent.

Exemple 1 : Prenons l'exemple d'une application proposant la gestion du chauffage dans une institution. Dans ce cas les paramètres de contexte sont : *l'endroit*, *la température intérieure*, *la température extérieure*, *le nombre de personnes* et *le moment du temps*. La figure 1.3 présente un exemple possible du schéma hiérarchique caractérisant le paramètre de contexte {*température extérieure*}. De cette figure, il peut s'observer la structure granulaire des paramètres de contexte. Par exemple dans le cas des conditions météo, l'intervalle $[-10°\text{C}, -5°\text{C}]$ est d'une granularité plus fine que la valeur linguistique {Froid}.

Le schéma hiérarchique de chaque paramètre de contexte forme un arbre avec au moins un nœud à chaque niveau. Pour chaque niveau N_j, $1 \leq j \leq m$, est associé un domaine de valeurs $dom_{N_j}(C_i)$. Donc, le domaine de valeurs de chaque paramètre de contexte sera donné par : $dom(C_i) = \bigcup_{j=1}^{m} dom_{N_j}(C_i)$. U*n état du contexte* est un n-uplet $\{c_1, c_2, \ldots, c_n\}$ dans lequel $c_i \in dom(C_i)$, $1 \leq i \leq n$. Dans le cas de l'exemple actuel, un état du contexte pourrait être {Salle C03b, 18°C, Froid, Petit groupe, Après-midi}.

1.3 Données, information et connaissance dans un SIC

Dans ce paragraphe, nous définissons les notions suivantes : **les données**, **l'information** et **la connaissance**. Ces définitions sont importantes car elles ont une influence directe sur les notions

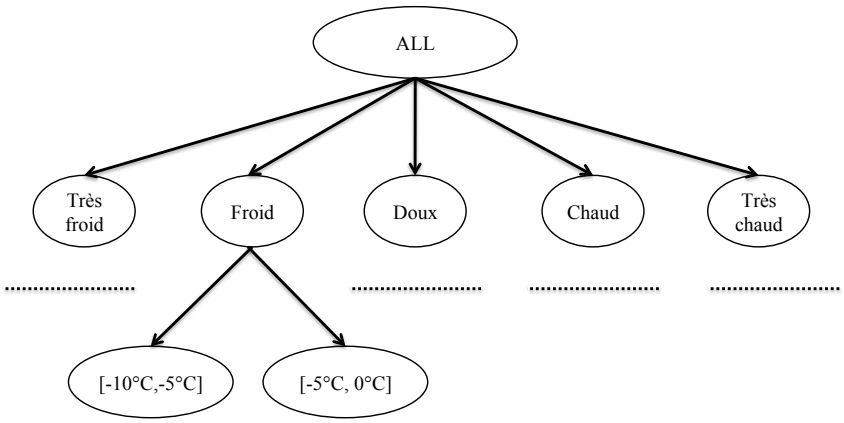


FIGURE 1.3: Exemple de schéma hiérarchique du paramètre de contexte *la température extérieure*

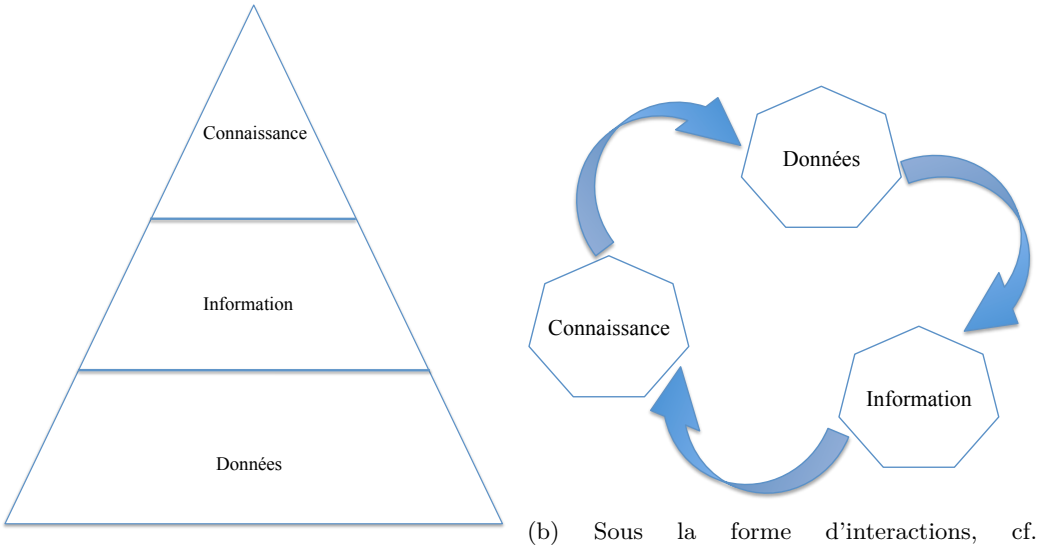
de qualité des données et de l'information. En fait, la qualité d'une entité est définie en fonction de ces propriétés. Malheureusement, la plupart des articles de recherche dans le domaine des systèmes d'information mélange les définitions de données et d'information, les considérant comme équivalentes. Il peut en être de même entre la notion d'information et celle de connaissance (cf. [Sølvberg 93] pages 366-372). Ainsi, très souvent ces notions sont vues comme étant synonymes. En conséquence, la qualité est définie et évaluée en employant les mêmes outils. Cela provoque, dans beaucoup de situations, des confusions sur ce qui doit être mesuré et sur ce qui est mesuré en réalité.

Essayer de définir ce qu'est l'information ou la connaissance reste un problème ouvert. Donc, il est impossible de proposer des définitions couvrant tous les aspects de ces notions. Ainsi, dans cette thèse, seront proposées des définitions uniquement dans *un contexte particulier*, celui des systèmes d'information utilisés dans le cadre d'un processus d'aide à la décision. Néanmoins, il est nécessaire de rester cohérent avec les définitions les plus globales, données dans des contextes plus généraux.

Une démarche très innovante pour la définition de ces trois notions, a été initiée par [Zins 07]. Dans son étude, Zins a demandé à quarante-cinq chercheurs (de 16 pays différents et couvrant tous les domaines de la science de l'information) de donner leurs propres définitions. Après avoir collecter toutes les réponses, sa conclusions a été que « apparemment, la communauté scientifique parle des langues différentes ».

La principale explication de cette difficulté de trouver une définition universellement acceptable pour ces notions est la nature *polymorphe* de celles-ci. Un concept polymorphe ne peut pas être défini d'une façon classique, c'est-à-dire en énumérant un ensemble de caractéristiques nécessaires et suffisantes qui soient universellement valides. Ce genre de concepts possède plusieurs définitions en fonction du contexte d'utilisation. Un exemple classique est le nom commun « voiture ». Sa définition varie si celle-ci est donnée par un mécanicien ou par un activiste de la protection de l'environnement. En même temps, une photo illustrant une voiture peut répondre à toutes les questions dans ce cas précis. Cependant, les notions mathématiques, comme par exemple les figures géométriques, possèdent des définitions très précises.

Pourtant, dans le cas des données, de l'information ou de la connaissance, la difficulté est encore plus grande, à cause de leur forte dépendance du contexte de définition. Actuellement, il n'y a pas de modalité reconnue capable de faire la différence entre une donnée, une information ou une connaissance, sur une base purement représentative. Ainsi, si un objet ou une structure, à l'intérieur d'un système est pris isolément, il n'est pas possible de discerner s'il s'agit d'une donnée, d'une information ou d'une connaissance. Néanmoins, la littérature présente des modélisations essayant de caractériser les relations existantes entre ces notions. La figure 1.4 présente deux modèles parmi les plus populaires. Le premier modèle à mettre toutes ces notions dans une seule formule est



(a) Sous la forme pyramidale, cf. [Zeleny 87]

(b) Sous la forme d'interactions, cf. [Gustavsson 05]

FIGURE 1.4: Modélisations illustratives des relations entre données, information et connaissance

celui de [Zeleny 87]. Il a proposé une structuration hiérarchique pyramidale, cf. figure 1.4a. Dans sa vision, la connaissance est extraite d'information, elles-mêmes issues de données. Une autre modélisation, très répandue dans le domaine de la fusion d'information, est celle présentée dans la figure 1.4b [Gustavsson 05]. Celle-ci présente les trois notions dans une boucle représentant la transformation continue d'une notion en une autre.

Le point de départ choisi pour définir ces notions consiste en l'identification de leurs rôles respectifs dans le cadre d'un système d'information d'aide à la décision.

1.3.1 État de l'art des définitions : contexte général

Avant de donner nos propres définitions, est présenté un état de l'art en analysant les définitions proposées dans des références incontournables : les dictionnaires, les standards et les travaux qui ont marqué le développement scientifique.

1.3.1.1 Les données

Selon *Le petit Robert* les données peuvent être définies comme :
- **Math.** : Ce qui est donné, connu, déterminé dans l'énoncé d'un problème, et qui sert à découvrir ce qui est inconnu ;
- **Inform.** Représentation conventionnelle d'une information (fait, notion, ordre d'exécution) sous une forme (analogique ou digitale) permettant d'en faire le traitement automatique.

Selon *Hachette* :
- Élément servant de base à un raisonnement, une recherche, etc. ;
- **Inform.** Information servant à effectuer des traitements ;
- **Math.** : Grandeur permettant de résoudre une équation, un problème ;

L'organisation internationale de normalisation [4] a également donné une définition des données dans la norme ISO 11179 [Sebastian-Coleman 13] :

4. ISO - International Standards Organization

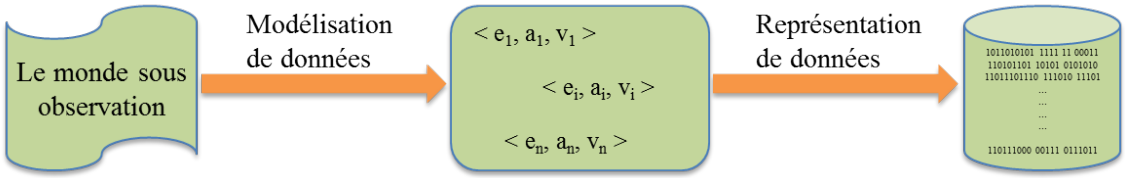


FIGURE 1.5: Les données étant décrites comme le résultat d'une étape de modélisation et d'une étape de représentation

– Représentations ré-explicables de l'information dans une manière formalisée et adaptée pour la communication, l'interprétation ou le traitement.

En analysant presque chaque monographie traitant du thème des bases de données et quelque soit le domaine d'application, le concept de données est décrit de la même façon. Ceci est illustré figure 1.5. Ainsi, dans ce domaine[5], les données sont supposées être le résultat d'une étape de modélisation et d'une étape de représentation. L'étape de modélisation a pour but d'obtenir des triplés $< e, a, v >$ permettant de décrire le domaine de l'observation. Dans ce triplé, v est la valeur prise (valeur faisant partie d'un domaine prédéfini) par l'attribut a correspondant à la description de l'entité e qui fait partie du domaine observé. La deuxième étape est la représentation et l'enregistrement de ces triplés, très souvent dans des bases de données.

Ainsi, en faisant une synthèse sur les divers définitions de la notion de données, il peut s'affirmer que :

> **Définition 9 : Les données (contexte général)** sont des représentations discrètes et objectives des événements et des entités qui ont été observés.

Les données peuvent être représentées selon un format structuré, semi-structuré ou encore non-structuré. Les données structurées sont numériquement enregistrées dans des tableaux faisant partie d'une base de données (par exemple une base de données SQL). Les données semi-structurées sont enregistrées dans un format moins bien défini, comme par exemple des fichiers XML ou HTML. Ce type de données est issu de l'Internet. Les données non structurées sont dans la plupart des cas issues de saisies textuelles des utilisateurs dans des formulaires, courriels ou autres documents écrits. Le traitement de ce dernier type de données est plus difficile à mettre en place et nécessite une étape de pré-traitement afin de pouvoir organiser le contenu, extraire des mots-clés, etc.

Les valeurs des données peuvent prendre une forme numérique, d'une chaîne de caractères ou encore de mots, de propositions ou d'autre objet (par exemple une images, une equation mathématique, ...), etc.

Observation : Même si dans la plupart des cas, les données sont collectées et stockées dans des bases de données pour un but bien précis, elles ont comme caractéristique principale : **l'indépendance du contexte**. Grâce à cette caractéristique elles peuvent être vues comme des représentation objectives. Cependant, lorsque les données sont mises dans le contexte (c'est-à-dire quelle est leur source, dans quelles conditions elles ont été collectées, etc.), elles cessent d'être purement objectives. Mais, dans ce cas, il s'agit plus de données, mais ... d'information.

1.3.1.2 L'information

Le paragraphe 1.1.2 indique qu'un système d'information ne peut pas exister sans l'information qu'il fournit à son utilisateur. Cependant, dans le cadre des systèmes d'information, il n'y a pas de définition générale acceptable par l'ensemble de la communauté. De plus, la nature même de l'information soulève des différentes opinions [Mingers 96].

5. Le domaine des bases de données est le seul domaine qui gravite autour de la notion de donnée. Elle est leur « matière première »

La notion d'information est souvent utilisée avec différents sens. Son utilisation est source de confusions importantes. L'utilisation habituelle du mot information, dans notre vie quotidienne, est liée à la notion de message : par exemple une lettre, une émission TV ou une page Internet. Tous ces exemples sont des sources d'information, elles communiquent de l'information.

Une consultation des dictionnaires fournit, en plus du sens introduit ci-dessus, les définitions suivantes (cf. Larousse et Hachette) :

- Élément de connaissance susceptible d'être codé pour être conservé, traité ou communiqué ;
- Une production sociale : pour qu'un fait devienne une information, il doit être communiqué à un public.

Donc, selon ces définitions la notion d'information se superpose beaucoup à celles de données et de connaissance. La seule différence majeure, par rapport aux données est la nécessité qu'elle soit communiquée à un auditoire.

Il existe également des définitions faisant appel à des modélisations mathématiques pour un plus grand formalisme. La plus connue des définitions est celle donnée par Claude Shannon, dans son article « A mathematical theory of communication » écrit en 1948. Cet article, considéré comme l'un des plus influents articles du XX-ème siècle, a posé les bases de la théorie mathématique appelée « théorie de l'information ».

[Weaver 49] dans son étude sur la nouvelle théorie développée par Shannon fait une analyse de la notion d'information et il identifie trois problèmes fondamentaux :

- **des problèmes techniques** : concernant la quantification et la transmission sans erreurs de l'information. Ceci a été couvert par la théorie mathématique de l'information de Shannon (l'entropie de Shannon) ;
- **des problèmes sémantiques** : liés à l'interprétation de l'information au niveau du sens et de la véracité de l'information ;
- **des problèmes d'impact et d'efficacité** : concernent le succès avec lequel le sens de l'information influence le comportement du destinataire (appelé par Weaver problème d'influence).

Selon [Shannon 93] p. 180 : « The word **information** has been given different meanings by various writers in the general field of information theory. It is likely that at least a number of these will prove sufficiently useful in certain applications to deserve further study and permanent recognition. *It is hardly to be expected that a single concept of information would satisfactorily account for the numerous possible applications of this general field.* » [6]

Pendant une longue période dans le domaine de l'ingénierie, le sens de l'information n'a pas été considéré comme une caractéristique importante de l'information, ce qui a généré une confusion générale entre la notion de donnée et celle d'information. Comme Shannon disait : « les aspects sémantiques d'une communication sont non-pertinents pour les aspects de l'ingénierie. » Donc, pour un ingénieur en télécommunications, le plus important était de pouvoir concevoir un système permettant de transmettre le plus grand nombre de bits possible avec un taux d'erreurs le plus faible possible.

Actuellement dans le domaine des systèmes d'information, il y a deux principales familles de définitions concernant la notion d'information :

1. La première famille définit l'information comme étant la résultante du traitement de données afin d'en extraire une utilité. Cette définition classique de l'information respecte le modèle pyramidale présenté dans la figure 1.4a. Même si elle ne couvre pas tous les aspects de l'information, elle est très intuitive et très facile pour son application pratique. Cette définition est celle adoptée dans le domaine du traitement du signal.
2. Au cours de ces dernières trois décennies, dans beaucoup de domaines de recherches, comme la Théorie des systèmes d'information, le Management des systèmes d'information, la Théorie

6. « Au cours du temps, divers écrivains du domaine général de la théorie de l'information ont associé le mot information à différents sens. Il est possible qu'au moins quelques uns de ceux-ci soient suffisamment utiles dans certaines applications pour mériter des études plus approfondies et une reconnaissance permanente. Il est peu probable qu'un seul concept de l'information soit satisfaisant pour couvrir les nombreuses applications possibles de ce domaine général. »

de la décision, il a été adopté *une définition générale de l'information* en considérant que [Floridi 09] :

$$Information = Données + Sens$$

Dans ce cadre, la définition générale de l'information précise que σ est une instance de l'information si est seulement si :

- σ est constitué de n données (D), avec $n \geq 1$. En d'autres mots l'information est issue de données et ne peut pas exister sans ;
- les données sont bien formatées, c'est-à-dire l'intégration des données pour former l'information respectent des règles de syntaxe imposées sur la forme, la construction, la composition ou encore la forme ;
- les données bien formatées ont un sens.

La définition donnée par [Floridi 09] est très intéressante parce qu'elle permet de répondre aux trois problèmes identifiés par Weaver. En plus, elle précise de façon explicite les éléments nécessaires pour la transformation de données en information. Cet aspect sera développé dans le paragraphe 1.3.2.1.

Observation : Par rapport aux données, les informations sont directement dépendantes du contexte : le formatage et le sens ajouté ne peuvent se faire que dans un contexte bien défini.

1.3.1.3 La connaissance

Dans le paragraphe précédent, la définition (du dictionnaire) de l'information faisait référence à la notion de connaissance. Malgré cette violation des règles de définition, la définition de la notion de connaissance est présentée ci-dessous (cf. Hachette) :

- Fait de connaître une chose, fait d'avoir une idée exacte de son sens, de ses caractères, de son fonctionnement ;
- Notions acquises ; ce que l'on a appris d'un sujet.

Ainsi, la connaissance a un caractère personnel et subjectif, étant interne à la personne qui détient la connaissance et donc, modélisée en fonction de ses perceptions et de ses expériences existantes.

Dans le domaine de la psychologie, il est identifié deux types de connaissances :

- **La connaissance explicite**, qui peut toujours être exprimée et expliquée aux autres. Ce type de connaissance peut être : encodée, acquise, créée, utilisée à l'effet de levier, enregistrée, transférée et partagée ;
- **La connaissance implicite**, qui représente les aptitudes personnelles. Elle est difficilement explicable aux autres (un exemple est l'aptitude à faire du vélo). Ces caractéristiques principales sont : le caractère personnel, la dépendance du contexte et la difficulté à être formalisée.

En fonction de la forme sous laquelle les connaissances sont exprimées, [Sølvberg 93] identifie cinq catégories :

La connaissance factuelle : qui s'exprime sous la forme : X est / a été / sera quelque chose.

La connaissance déontique : qui s'exprime sous la forme : X devrait être quelque chose. Elle exprime nos croyances sur quelque chose qui est ou qui devrait être.

La connaissance explicative : qui s'exprime sous la forme : X est quelque chose parce que Y est ... Ce type de connaissance est utilisé lorsqu'on cherche une solution à un problème.

La connaissance instrumentale : qui s'exprime sous la forme : X est réalisé, donc Y est la conséquence. Ce type de connaissance fournit les moyens avec lesquelles la personne qui détient la connaissance est capable de changer quelque chose.

La connaissance conceptuelle (appelée aussi méta-connaissance) : qui est une connaissance sur le sens des mots et sur les autres moyens utilisés dans le processus de communication afin de la rendre intelligible. Ce type de connaissance définit le protocole de communication.

Si on fait une comparaison entre la notion d'information et celle de connaissance, il peut s'observer que si l'information est représentée par le sens incorporé dans le message, la connaissance est l'information assimilée par les structures cognitives d'une personne. Réciproquement, il peut s'affirmer que l'information est une représentation des connaissances comprises [San Segundo 02].

1.3.2 Définitions de données, de l'information et de la connaissance dans le contexte d'un SIC

Les notions de données, information et connaissance, adaptées à notre étude sont maintenant définies dans le contexte d'un système d'information complexe d'aide à la décision. Ainsi, dans un premier temps, sont identifiées ces notions par rapport au système et puis, dans un deuxième temps, sont étudiées les interactions entre ces notions, toujours dans ce même contexte.

Afin de positionner ces notions dans le contexte d'un système d'information complexe d'aide à la décision, nous garderons la description d'un SI, présentée dans la figure 1.1, avec une modélisation des données, de l'information et de la connaissance sous forme pyramidale (figure 1.4a) et sous forme d'interactions (figure 1.4b).

En analysant la structure générique d'un système d'information, quatre domaines d'abstraction différents peuvent être identifiés. Ces domaines représentent respectivement l'environnement du système d'information, les données, l'information et les connaissances (figure 1.6). La première observation est que la structure de cette représentation suit le modèle pyramidal. Également ont été ajoutées les diverses interactions entre ces quatre domaines, identifiées dans [Aamodt 95]. Ces interactions sont représentées par des flèches numérotées.

Ce type de représentation générale d'un système d'information peut couvrir un large spectre de contextes d'application, des plus simples jusqu'aux plus complexes. Ainsi, si nous considérons l'exemple d'un système d'information utilisé pour la gestion d'un conflit militaire, ce schéma peut balayer toutes les entités du champ de bataille jusqu'à l'évaluation et la conscientisation de la situation [7]. Pourtant, dans ce type d'application les deux acteurs du conflit ont, généralement, deux visions différentes de la même réalité, qui est unique et qui est représentée par le terme *situation*. Cette différence de vision est due aux trois domaines (correspondant aux données, aux informations et aux connaissances) qui ne sont pas traités de la même façon par les deux côtés.

Maintenant, chaque domaine ainsi que ses interactions avec les autres domaines seront présentés, afin de donner les définitions des données, de l'information et de la connaissance.

Le domaine physique décrit les entités constituant l'environnement. Il contient également des entités abstraites comme par exemple les plans, les intentions, etc. des entités physiques. Comme ces entités abstraites peuvent être vues comme étant des représentations de haut niveau, elles peuvent être considérées comme étant des connaissances *a priori* et donc elles sont injectées directement (voir la flèche (1') dans la figure 1.6) au domaine cognitif (de plus haut niveau). Toutes ces entités abstraites ont, en général, un caractère dynamique et nécessitent des estimations pour une bonne évaluation de la situation.

Le deuxième domaine est celui de données, obtenues à partir du domaine physique par différents processus de mesure (par exemple en utilisant différents types de capteurs). Elles forment le domaine des données, contenant les données stockées dans des bases de données. Le passage vers les données est représenté par la flèche (1) dans la figure 1.6. Ces données sont par la suite utilisées pour l'extraction d'informations. Comme il est possible que les données collectées ne soient pas satisfaisantes (la situation ayant évolué), il est nécessaire d'en enregistrer d'autres. Cette commande vient de la part du domaine cognitif (représentée par la flèche (7)) et elle est en correspondance directe avec le besoin en information du processus de prise de décisions. À ce niveau, nous proposons une définition de données :

7. De l'anglais : *situation awareness*

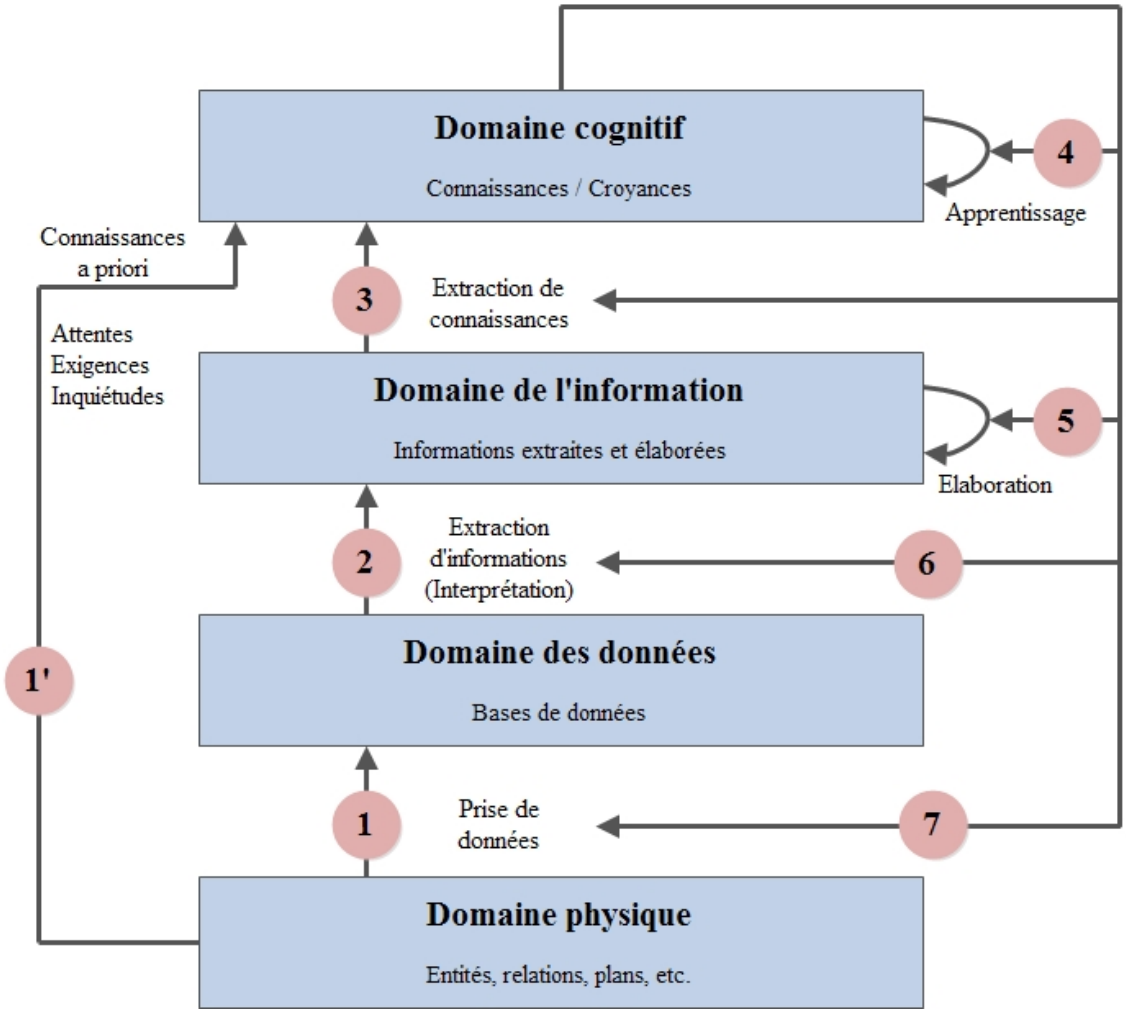


FIGURE 1.6: Les quatre domaines (niveau) : physique, données, informations et cognitif

Définition 10 : Les données (contexte SI) sont des représentations syntaxiques (sans aucun sens associé) qui n'ont pas été interprétées dans un contexte bien défini. Elles sont issues de l'environnement sous observation et peuvent avoir diverses formes de représentation (discrètes dans le cas d'un système d'information) : caractères numériques ou alphanumériques, signaux n-dimensionnels, etc.

Le domaine de l'information est caractérisé par un processus de transformation des données en informations, suite à une étape d'interprétation [Aamodt 95] (représentée par la flèche (2) dans la figure 1.6). Ce passage est contrôlé par le domaine cognitif (flèche numéro (6)) parce que la transformation des données en informations ne peut se faire qu'avec la contribution de la connaissance. En même temps, à ce niveau, d'autres informations peuvent être obtenues suite à une étape d'élaboration des informations existantes, toujours avec l'intervention des connaissances (cf. la flèche numéro (5)). Comme notre principal intérêt est la qualité des données et des informations, l'aspect de la transformation des données en information sera traité plus en détail dans le paragraphe 1.3.2.1. Néanmoins, ci-dessous on donne une définition de l'information :

Définition 11 : L'information est issue des données qui ont été interprétées en y ajoutant de la sémantique dans un contexte bien défini et par rapport aux besoins du processus d'aide à la décision.

Le domaine cognitif est caractérisé par les connaissances des utilisateurs. Ces connaissances peuvent être des connaissances *a priori*, des attentes, des exigences, des inquiétudes ou des connaissances acquises à partir des informations disponibles. C'est à ce niveau que l'utilisateur vient en contact avec les informations fournies par le système d'information, par une étape d'extraction de connaissances (cf. flèche (3)). Ainsi, au fur et à mesure du temps, les connaissances de l'utilisateur évoluent suite à son interaction à l'environnement. L'acquisition des nouvelles connaissances se fait en intégrant les informations reçues de la part du système d'information dans ses propres structures cognitives. Cette intégration se fait suite un processus d'apprentissage (cf. flèche (4)) qui peut être décrit par :

$$\text{Nouvelles informations} + \text{Connaissances existantes} \rightarrow \text{Nouvelles connaissances} \tag{1.1}$$

De plus, tout apprentissage se fait avec un objectif bien défini, c'est-à-dire il fait la liaison entre les connaissances et leur utilisation potentielle : c'est à ce niveau que le processus de prise de décisions a lieu.

Ainsi, si on reste dans le cas d'une application de gestion d'un conflit militaire, l'utilisateur peut avoir des connaissances a priori sur l'emplacement de ces forces et des forces ennemies, mais aussi des connaissances sous la forme de performances attendues de ces unités et des unités ennemies. Toutes ces connaissances ont été représentées comme provenant du domaine physique et ensemble avec les connaissances acquises, suite à l'utilisation du système, vont donner une estimation actuelle de la situation.

Sans connaissance, le système d'information ne peut pas fonctionner. Les connaissances sont indispensables pour l'interprétation des données, l'élaboration des informations et pour l'apprentissage.

Exemple 2 : prenons le cas d'un système d'information ressemblant à celui présenté dans la figure 1.2 et supposons qu'un des capteurs de ces plateformes a enregistré l'image [8] présentée dans la figure 1.7. Sans d'autres éléments, cette image représente une donnée. Si cette image est par la suite :

1. **mise dans le contexte** (provenance capteur ISAR surveillant la zone X) et
2. **interprétée** (la cible est identifiée comme étant un avion F4)

8. Cette image est issue d'une acquisition expérimentale de la chambre anéchoïque de L'ENSTA Bretagne en utilisant une maquette à l'échelle $1/48^{ème}$.

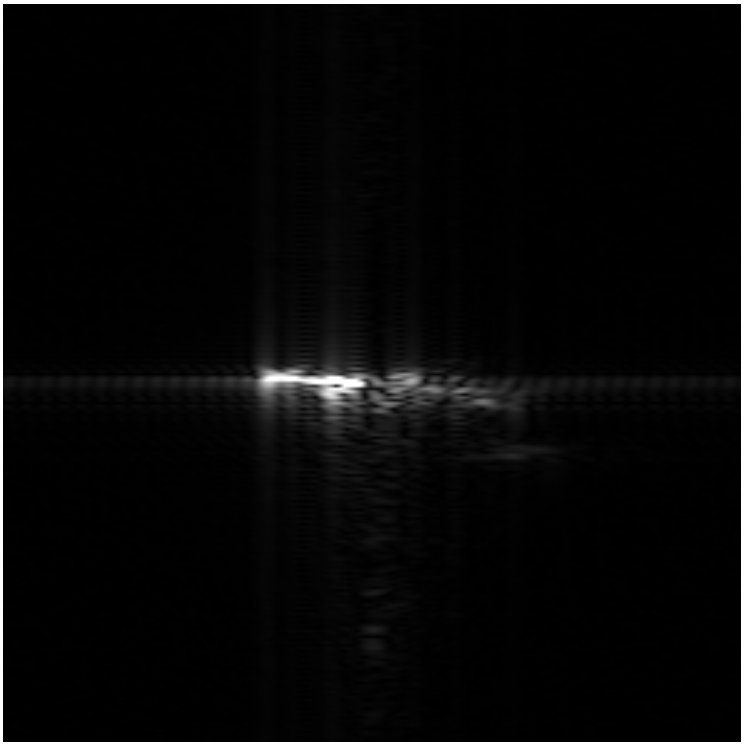


FIGURE 1.7: Exemple de signature d'un capteur ISAR

l'information suivante est extraite : un avion F4 a été identifié dans la zone X. Ensuite si cette information est fournie à un commandant, il va l'intégrer dans ses propres structures cognitive en réalisant une mise à jours des ces connaissances : un avion d'attaque F4 est dans la zone X. Afin de pouvoir se former une idée plus claire de cette situation, le commandant pourrait avoir besoin d'autres informations, comme par exemple l'allégeance de cette cible. Cette demande, représentée par la flèche (7) dans la figure 1.6, sera poursuivie par l'activation d'autres équipements (par exemple un système identification ami ou ennemi) afin d'obtenir l'information demandée.

1.3.2.1 Le transformation des données en informations

Dans le paragraphe précédent, l'information a été définie comme étant un ensemble de données organisées qui possèdent un sens (une sémantique) dans le cadre du contexte d'application. Cette organisation de données peut se faire suite à des opérations de clustering, d'indexation ou d'association en fonction de la syntaxe utilisée.

Le sens de l'information est acquis par la projection des données dans un autre domaine en correspondance directe avec le contexte d'application. Ce domaine contient les valeurs possibles de l'information. Si on considère le cas d'un système d'information utilisé dans le processus d'aide à la décision, le nombre de décisions possibles est fini. En conséquence, le domaine de l'information peut être supposé de dimension finie. Le passage du domaine des données vers le domaine de l'information est illustré dans la figure 1.8. Dans cette figure, à partir du domaine des données sont extraites et organisées [9] les données D_0, D_i et D_l. Associée à ces données se trouve l'information I_3, d'un total de N informations possibles. Cette association peut être par exemple la résultante

9. L'extraction et l'organisation de données se fait avec l'apport d'une base de connaissance comme a été défini dans le paragraphe précédent. Pour des raisons de simplicité on ne présente pas cette intervention et on la considère comme implicite

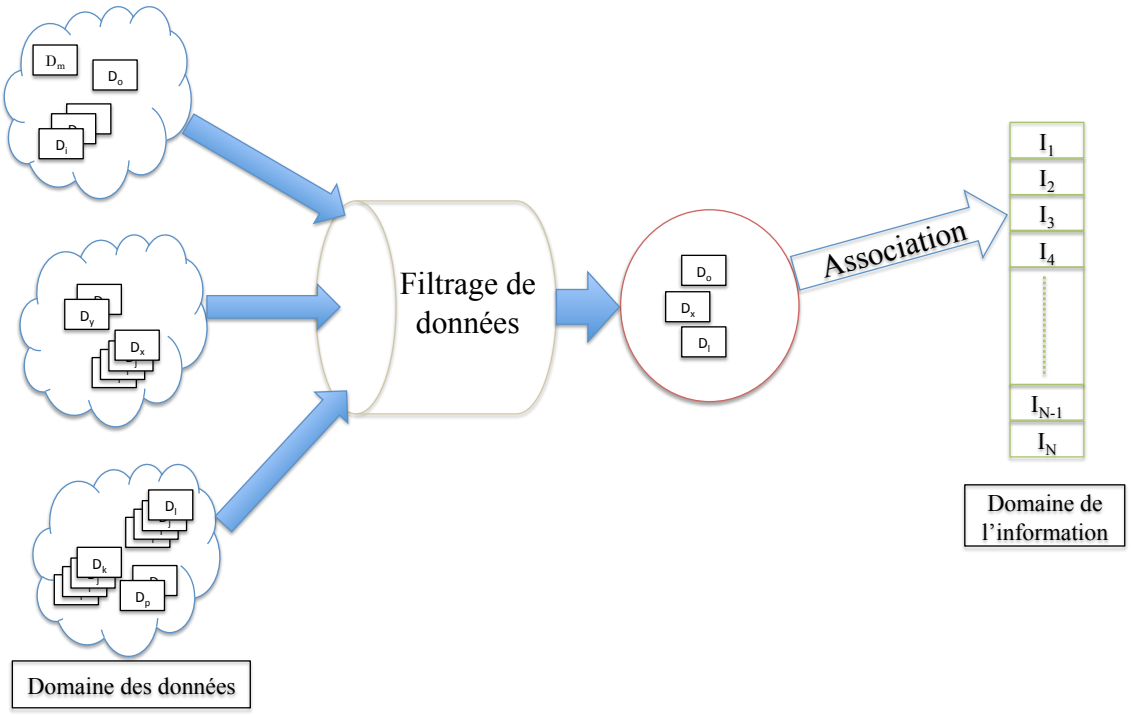


FIGURE 1.8: La transformation des données dans des informations

d'une règles du type « SI - ALORS » [10] [Berkan 97].

Cependant, cette cartographie du domaine de données vers le domaine de l'information n'est pas toujours parfait. Ainsi, il existe la possibilité que les données disponibles ne supportent pas l'extraction des informations attendues. Il y a plusieurs explications à ce phénomène. La plus fréquente est liée à l'indisponibilité et à l'imperfection des données : il n'est pas toujours possible d'avoir les données voulues. La deuxième est liée à l'imperfection de la transformation des données en information. Cette transformation est la fonction principale d'un système d'information et comme tout système réel il possède des limites.

Dans la figure 1.9 est présentée une nouvelle vision de la transformation des données dans des informations. Cette fois-ci, le système d'information joue le rôle d'extraction d'informations et il est modélisé comme un processus.

Dans cette nouvelle représentation, il peut s'observer que la principale caractéristique des données est d'englober les informations utiles (par exemple à la figure 1.9, il y en a N informations utiles). Ainsi si les données « contiennent » [11] toutes les informations dont l'utilisateur a besoin, ces données peuvent être qualifiées de très bonne qualité. Bien sûr les données peuvent contenir des « non-informations », c'est-à-dire des données qui ne sont pas pertinentes et donc de qualité nulle pour l'utilisateur. C'est le rôle du système d'informations de discerner entre informations utiles et non-informations. En fonction de l'adaptation des traitements, le fonctionnement du système peut être « qualifié » (dans la figure $Qual_{proc}$). Au final, à la sortie du système d'information, trois cas peuvent être identifiés en fonction du nombre M d'informations proposées et du nombre N d'informations utiles pour l'utilisateur. Le premier cas est quand le système propose exactement les N informations attendues par l'utilisateur. Dans ce cas, la satisfaction de l'utilisateur est maximale et donc le système d'informations est d'une très bonne qualité. Sinon, dans les cas quand $N < M$ ou quand $N > M$, la qualité du système d'informations est sous-optimale. Cet exemple montre que la

10. Le terme anglais est « IF-THEN rules »
11. Dans le sens que les données supportent l'extraction des informations

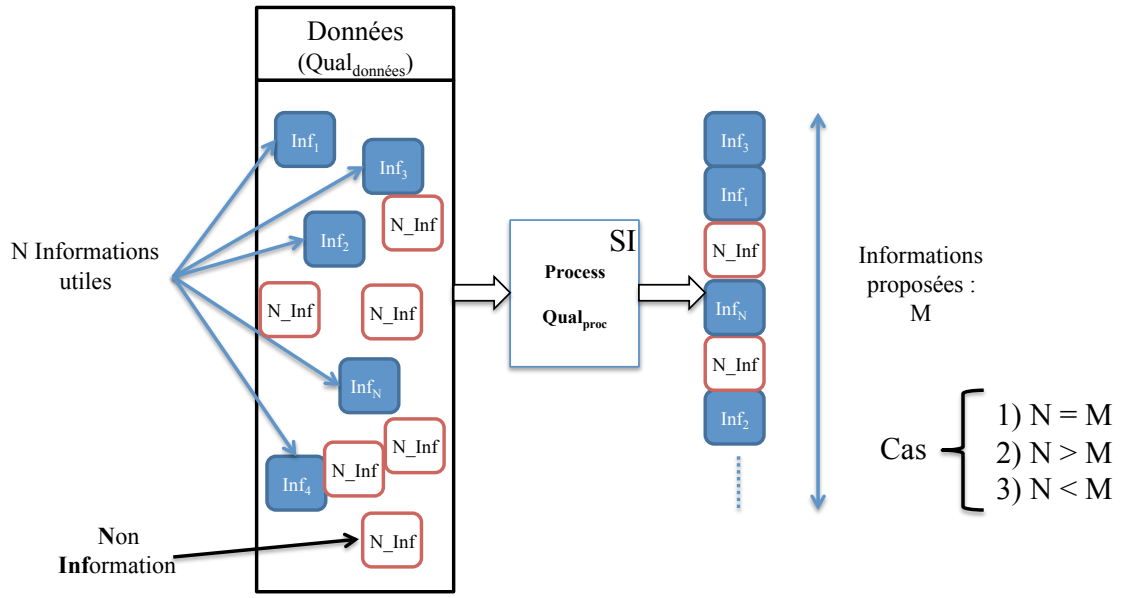


FIGURE 1.9: Le passage des données vers des informations

qualité des données ne peut s'évaluer que par rapport aux informations pouvant s'en extraire. Une présentation plus détaillée de la qualité des données et de l'information se trouve dans chapitre 2 et respectivement, chapitre 3.

Un exemple de transformation de données en information est présenté dans la figure 1.10. Il s'agit d'un traitement typique pour un système d'information géographique [12]. De cet exemple, nous observons qu'à partir des données, représentées par la première couche (en bas de la figure 1.10), différents types d'informations peuvent être extraites, en fonction du besoin de l'utilisateur final. Ainsi, par exemple, l'information extraite peut être :

- l'utilisation du terrain (forêt, fermes, maisons, etc.), la deuxième couche de bas en haut ;
- le relief de la zone sous observation, la troisième couche de bas en haut ;
- les parcelles de terrain, la quatrième couche de bas en haut ;
- les routes, la cinquième couche de bas en haut ;
- les entités identifiées dans cette aire géographique, la sixième couche en haut de la figure 1.10.

Cet exemple est une illustration pratique de notre modélisation des notions de données et d'information, présentée dans les figures 1.8 et 1.9 : l'information est obtenue en filtrant les données en ne gardant que les éléments informatifs.

Une autre question qui peut se poser ici est la façon dont la qualité de l'information obtenue peut être calculée sachant la qualité des données et la qualité du processus faisant la liaison entre les données et les informations. Dans la deuxième partie de ce mémoire de thèse, il est proposé une possible réponse à cette question sous la forme d'une nouvelle méthodologie d'évaluation de la qualité de l'information.

À la fin de ce paragraphe, nous proposons une formalisation mathématique de la notion d'information dans le contexte d'un système d'information. Une information peut être modélisée sous la forme d'une paire : $I = (E, A)$ dans laquelle :

- E est un ensemble fini non-nul d'entités ;
- A est un ensemble fini non-nul d'attributs ;
- pour chaque attribut $a \in A$ il existe une fonction $a : E \rightarrow V_a$ assignant des valeurs d'un

12. En anglais « Geographic Information Systems (GIS) »

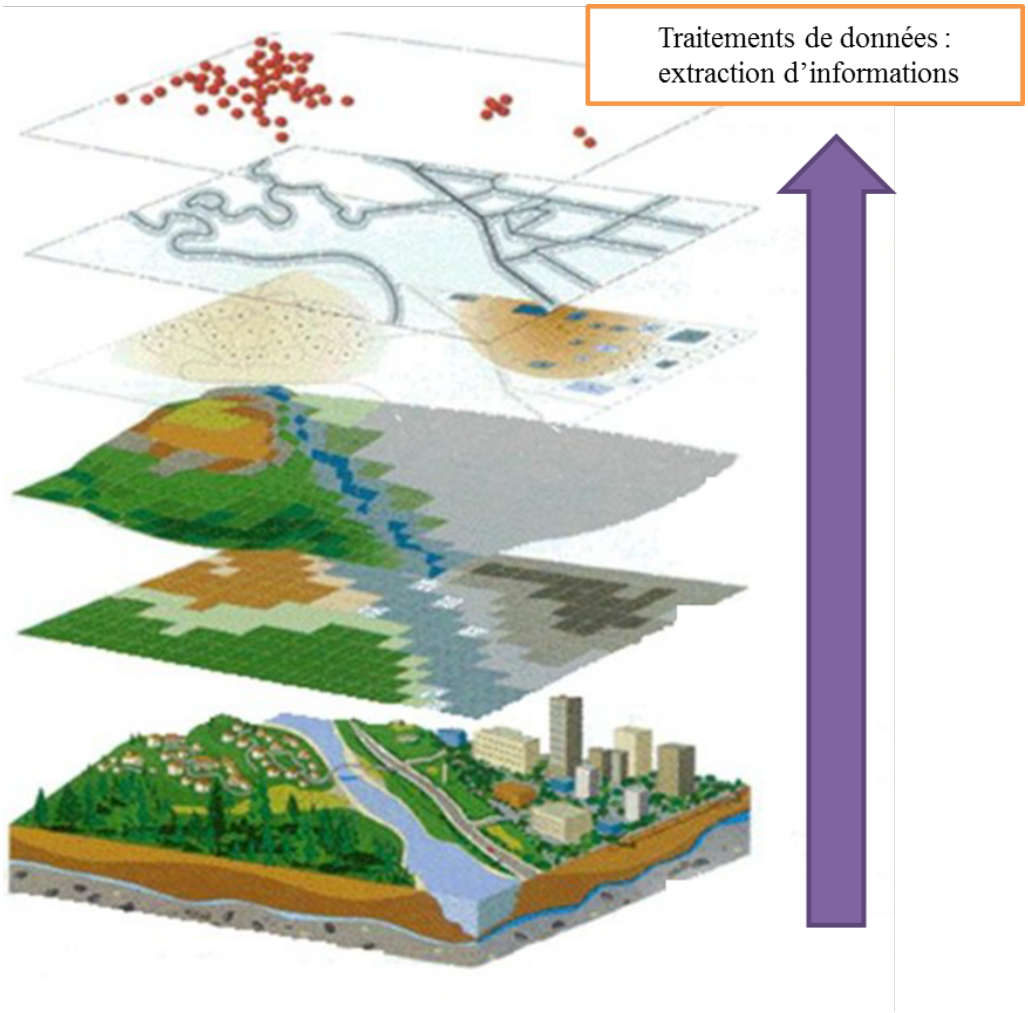


FIGURE 1.10: Exemple de transformation des données en informations, adapté de [Solano 12]

• caractéristiques ; • compétitivité ; • courtoisie ; • conformité aux standards et procédures ; • convivialité ; • fiabilité ; • durabilité ; • rendement du processus ; • temps du cycle de développement ; • coût ; • responsabilité sociale.	• performance ; • promptitude ; • potentiel du processus ; • manque d'erreurs ; • sécurité : employés, clients et environnent ; • maintenance ; • esthétique ; • coût de la faible qualité ; • prix ; • satisfaction des employés ;

TABLE 1.1: Liste de dimensions de qualité selon [Juran 89]

domaine V_a propre à l'attribut a aux entités E.

Si un système de reconnaissance de cibles radar est considéré, comme celui de l'exemple 2, les entités E correspondent à l'ensemble de cibles pouvant être détectées (F4, Rafale, A10, A320, etc.) et les attributs A aux caractéristiques de ces cibles (allégeance, position, vitesse, etc.).

Précédemment, la notion de qualité a été utilisée sans la définir. Par conséquent, dans le paragraphe suivant, nous passons rapidement en revue les définitions d'origine de la qualité (c'est-à-dire d'un produit physique), pour ensuite introduire la qualité d'une données / information.

1.4 LA QUALITÉ D'UN PRODUIT

Il n'y a pas une seule définition de la notion de qualité. Au fur et à mesure du temps plusieurs chercheurs (appelés très souvent « gourous de la qualité ») ont essayé de donner des définition génériques.

Ainsi, selon [Feigenbaum 91] la qualité est définie comme étant « *La composition totale de caractéristiques de produits et services de marketing, ingénierie, fabrication et maintenance à travers lesquels les produits et les services vont répondre aux attentes des clients.* »

À partir des études de Deming [Deming 82] et de Juran [Juran 89] la notion de qualité est vue comme étant directement liée au contexte. Juran introduit l'expression « fitness for use » pour décrire la qualité. Deming et Juran ont été les premiers à s'intéresser à l'augmentation de la productivité des entreprises et ils ont défini la qualité par rapport aux produits développés par ces dernières. Dans ce cas, la qualité d'un produit *physique* est généralement définie d'une façon unique. Cette forme unique de la qualité est donnée par les caractéristiques qui, grâce à la nature tangible du produit, sont facilement mesurables et comparées à des étalons (standard de qualité). [Juran 89] avance également une liste (non-exhaustive) de dimensions possibles de la qualité (Cf. tableau 1.1).

Selon [Juran 89], la qualité d'un produit devrait être analysée par rapport à quatre aspects :

1. Le processus de production
2. Les caractéristiques du produit
3. La concurrence : (c'est-à-dire par rapport aux autre produits sur le marché)
4. Les employés

au processus de production, aux caractéristiques du produit, à la concurrence et aux employés.

[Hunt 92] fait une classification des dimensions de la qualité d'un produit en mettant l'accent sur deux aspects : l'aspect factuel et la perception (Cf. tableau 1.2). La qualité factuelle décrit ce que le produit ou le service doit faire, tandis que la qualité de la perception concerne la vision du produit ou du service dans les yeux des clients, c'est-à-dire ce que les clients doivent croire sur le produit ou le service.

Qualité factuelle	Qualité de la perception
• Soit le bon • Fonctionne correctement • Fonctionne correctement dès la première foi • Soit délivré en temps	• Est le bon • Répond aux besoins • Satisfait les attentes • Est délivré avec intégrité et respect

TABLE 1.2: Liste de dimensions de qualité selon [Hunt 92]

Par opposition avec un produit physique, les données et les informations ont un caractère fortement *fongible*, c'est-à-dire en ayant des significations différentes en fonction du contexte de définition. Cette différence fait que la qualité des données et des informations et fortement dépendante de l'utilisateur et du contexte d'application. La même information pourrait être d'une très bonne qualité pour un utilisateur et mauvaise pour un autre. Prenons par exemple le cas d'un signal reçu de l'espace, un bruit cosmique. Pour un système de communication par satellite il est un simple bruit qui doit être filtré, mais pour un système d'observation de l'espace il porte toute l'information utile pour les astronomes.

1.5 CONCLUSION

Avant de pouvoir définir la qualité des données et de l'information, il est nécessaire de définir les notions de données et d'information. En effet, la qualité d'une entité exprime les propriétés voulues, désirées de celle-ci. Ainsi, il est impératif d'identifier les propriétés des données et de l'information. Malheureusement, dans la littérature il n'y a pas de consensus sur leurs définitions et très souvent, il existe des confusions entre les propriétés des données et celles de l'information ou encore de celles de la connaissance.

Par conséquent, ce premier chapitre de ce mémoire de thèse a été consacré à la définition aux notions de base : données, information et connaissance. Comme il s'agit de notions contextuelles, nous avons commencé par la définition d'un système d'information.

La définition d'un système d'information a été donnée par rapport à deux autres types de systèmes : informatique et expert. Ainsi, nous avons très clairement établi les frontières entre ces trois types de systèmes en présentant les caractéristiques de chacun. De plus, comme ce travail de recherche est fait dans le cadre d'un système d'information complexe, nous avons également traité la notion de complexité afin d'en déduire les caractéristiques qui rend complexe un système d'information.

La seule définition d'un système d'information n'est pas suffisante pour la définition des trois notions sus-mentionnées, car une autre notion, celle du contexte d'application, joue également un rôle important. En effet, un système d'information ne peut être développé que dans un contexte d'application très bien défini. Par conséquent, le contexte d'application et le système d'information constituent le contexte dans lequel les données, l'information et la connaissance devrait être définies.

Avec le contexte de définition établi, nous avons proposé des définitions pour les trois notions. Nous avons pris en compte les définitions existantes dans la littérature, en les adaptant au contexte de systèmes d'information complexes utilisés dans le processus d'aide à la décision.

À la fin de ce chapitre, un survol rapide de la littérature a été réalisé, afin de trouver les principales définitions de la notion de qualité. Ces définitions vont nous servir dans les deux prochains chapitres, lors de la définition de la qualité des données et, respectivement, de la qualité de l'information.

2

Qualité des données

Ce chapitre présente les techniques de calcul de la qualité des données les plus utilisées dans la pratique. La qualité des données est un sujet de recherche *in sine* depuis le développement des techniques modernes de traitement numérique.

Le paragraphe 1.3.1.1 a donné une définition de la notion de « données ». Sans perdre en généralité, dans ce chapitre nous traitons la qualité des données stockées dans une base de données. Ainsi, pour l'étude de la qualité des données, il est nécessaire de considérer deux aspects : la qualité individuelle des données (c'est-à-dire des valeurs constituant la donnée) et la qualité de la représentation des données dans les bases de données (c'est-à-dire par rapport aux règles à respecter de la base de données).

Le processus de mesure et enregistrement des données est limité par le monde observé, qui est la source de toutes les données. Et ces limitations sont en relation directe et implicite avec le niveau de qualité des données. Ainsi, par exemple, le principe d'incertitude de Heisenberg nous dit qu'il est impossible de mesurer à la fois la vitesse et l'emplacement d'une particule élémentaire au-dessus d'un certain niveau de précision. À une échelle plus grande, il existe un nombre très important de phénomènes avec des caractéristiques aléatoires ou peu connus, par exemple les conditions météo, le comportement humain (dont la prise de décisions), etc. Toutes ces caractéristiques non-déterministes imposent des contraintes dans le processus de prise de mesure et chaque violation partielle ou totale de ces contraintes affecte la qualité de données en résultant.

Prenons l'exemple d'un signal audio. Afin d'être numérisé et enregistré, il est nécessaire de l'échantillonner. Mais pour garder toute l'information portée par ce signal audio, il faut que la fréquence d'échantillonnage soit plus supérieure au double de la plus haute fréquence de ce signal. Cette dernière condition est un exemple de contrainte.

Une base de données est un modèle numérique du monde réel. Ainsi, comme tout autre modèle, il essaie de représenter une version abstraite de la réalité. Le niveau d'abstraction est donné dans la plupart de cas par le contexte d'application [Motro 95]. La modélisation la plus simple d'une base de données prend la forme de plusieurs tables relationnelles. Chacune de ces tables modélise un ensemble d'objets similaires du monde réel. Chaque ligne décrit un objet particulier et les colonnes, communes pour tous les objets, décrivent les attributs des objets.

Avant de rentrer dans l'étude de la qualité des données et des bases de données, on rappelle quelques notions et notations usuelles dans le domaine de base de données [Ramakrishnan 03], [Decker 09] :

- Un *atome* est une expression de la forme $p(a_1, a_2, \ldots, a_n)$, où p est un prédicat n-aire, $n \geq 0$;
- Les a_i s'appellent *attributs* (ou encore *arguments*) et peuvent prendre des valeurs constantes ou variables ;
- Un *fait* est un atome avec tous les attributs constants ;
- Une *clause de base de données* est soit un fait dont son prédicat correspond à une table relationnelle et dont les attributs correspondent aux valeurs des colonnes de cette table, soit une formule de la forme $A \leftarrow B$, où la tête A est un atome et le corps B est une conjonction de littéraux ;
- Une *base de données* est un ensemble fini de clauses de base de données ;

– Une *mise à jour* est un ensemble fini de clauses de base de données à insérer ou à supprimer. Ainsi pour une mise à jour $U = U^+ \cup U^-$ de la base de données D, la nouvelle base de données sera notée D^U et sera la résultante de l'ajout dans D de toutes les clauses à insérer de U^+ et de la suppression de toutes les clauses à supprimer de U^-. Donc, on peut aussi utiliser la formule $D^U = (D \cup U^+) \backslash U^-$, dans laquelle $\cup$ et $\backslash$ sont les opérateurs de réunion et respectivement de différence des ensembles.

L'étude de la qualité des données stockées dans une base de données peut se faire à deux niveaux. Le premier, au niveau syntaxique, concerne l'intégrité de la base de données, c'est-à-dire les contraintes que les données doivent respecter afin de pouvoir être enregistrées dans la base de données. Le deuxième, au niveau sémantique, concerne la conformité à la réalité des données enregistrées. Plus précisément sont étudiées les imperfections des données.

Le paragraphe 2.1 traite la consistance des bases de données, exprimée sous la forme de contraintes d'intégrité. Ensuite, le paragraphe 2.2 présente les imperfections les plus fréquentes de données. Le paragraphe 2.3 fait une récapitulation de la qualité des données et propose une représentation sous la forme d'une taxonomie. Suite à ces trois paragraphes, le paragraphe 2.4 présente différentes possibilités d'étendre le modèle relationnel afin de pouvoir incorporer la qualité des données stockées. Ce chapitre se termine par une conclusion qui présente une synthèse sur la notion dédiée à la qualité des données.

2.1 Consistance des bases de données

Ce paragraphe présente un état de l'art sur les différents types de contraintes d'intégrité existantes dans la littérature. La première classification de ce type de contraintes a été proposée par Codd dans les années 70 [Codd 70].

Dans la théorie des bases de données, la qualité de données peut être modélisée au niveau déclaratif sous la forme d'une série de conditions appelées *contraintes d'intégrité*. Ces contraintes ont pour rôle de garder une conformité sémantique des données enregistrées. Prenons le cas d'une donnée représentant l'âge d'une personne. Un exemple de contrainte d'intégrité pour ce type de donnée est la restriction de valeurs acceptable aux entiers positifs et inférieurs à 130.

Ainsi, à l'aide de ces contraintes d'intégrité, la qualité de données peut être contrôlée, leur violation indiquant une mauvaise qualité. Les contraintes d'intégrité sont des conditions spécifiées lors de la définition du schéma de la base de données et restreignent les données qui peuvent y être enregistrées.

Une base de données n'est pas une ressource statique ainsi, au fur et à mesures du temps, des nouvelles contraintes d'intégrité sont spécifiées et renforcées [Ramakrishnan 03] :

– Quand l'administrateur de la base de données ou l'utilisateur final définissent un nouveau schéma ;
– Quand une application sur la base de données est en train de s'exécuter, le système de management de la base de données vérifie pour les violations de contraintes d'intégrité et refuse les changements susceptibles de les violer.

Même si le contrôle de contraintes d'intégrité se fait d'une manière automatique, la consistance des données est très souvent compromise dans la pratique. Parmi les pratiques responsables de cette dégradation de la qualité de données, se retrouvent [Decker 09] :

– l'ajout sans vérification de nouvelles contraintes ;
– l'arrêt temporaire de la vérification des contraintes d'intégrité pour télécharger un backup ou pour augmenter l'accessibilité de données ;
– l'intégration (la fusion) de plusieurs bases de données.

2.1.1 Définition

Une Contrainte d'Intégrité (CI) est un prédicat logique de premier ordre. Plus précisément il s'agit d'une condition qui doit être respectée par les enregistrements d'une relation. Les contraintes sont exprimées sous la forme de négations, d'interdictions. Un ensemble fini de conditions d'intégrité forme une *Théorie d'Intégrité* (TI). Les contraintes d'intégrité dépendent du contexte d'application. Par exemple pour une base de données contenant les clients d'une banque, l'age de clients doit être supérieur à 18 ans, qui n'est pas le cas d'une base de données d'un hôpital. Les contraintes d'intégrité s'appliquent à la base de données entière et donc leurs vérifications demandent un temps qui est proportionnel à la taille de la base [Christiansen 04].

> **Définition 12 :** Un état d'une base de données D est **consistant** avec une théorie d'intégrité TI si et seulement si $D(TI) = vraie$, c'est-à-dire toutes les contraintes d'intégrités composant TI sont respectées par D.

Par la suite, les contraintes d'intégrité vont s'exprimer par la formule générique $\leftarrow B$. Le corps B représente les déclarations qui doivent être respectées.

Exemple 1 : Soit une table $p(NumSecu, Nom, NumTel)$ définie par le prédicat p décrivant les personne enregistrées dans une base de données quelconque. Les trois colonnes contiennent le numéro de sécurité sociale $NumSecu$, le nom Nom et le numéro de téléphone $NumTel$ des personnes. Supposons que la contraintes d'intégrité I est définie comme :

$$I = \leftarrow ((p(x, y_1, z_1)) \wedge (p(x, y_2, z_2)) \wedge (y_1 \neq y_2 \vee z_1 \neq z_2)) \tag{2.1}$$

La contrainte d'intégrité I interdit que deux personnes ayant le même numéro de sécurité sociale x aient des noms différents (y_1 et y_2) ou des numéros de téléphone différents (z_1 et z_2). En d'autres mots, I oblige que le numéro de sécurité soit unique pour toutes les personnes. Maintenant, si suite à une opération de mise à jour U, l'enregistrement $p(1111, Jean, 0611111100)$ est inséré, la contrainte I impose la vérification de :

$$I^{'} = \leftarrow ((p(1111, Jean, 0611111100)) \wedge (p(1111, y_2, z_2)) \wedge (Jean \neq y_2 \vee 0611111100 \neq z_2))$$

Donc, lors de la mise à jour, il est vérifié qu'il n'y a pas un autre enregistrement ayant le même numéro de sécurité mais avec un nom ou un numéro de téléphone différent. La contrainte $I^{'}$ peut encore s'exprimer sous une forme plus simple :

$$I_s^{'} = \leftarrow ((p(1111, Jean, 0611111100)) \wedge (Jean \neq y_2 \vee 0611111100 \neq z_2))$$

Cette dernière est beaucoup plus intéressante car elle permet de ne parcourir qu'une seule fois la base de données (complexité linéaire), tandis que la contrainte d'intégrité $I^{'}$ est plus complexe demandant à la limite une opération de $JOIN$ de p avec lui même (de complexité $N \cdot \log(N)$, avec N le nombre d'enregistrements).

Malheureusement très souvent dans la pratique les données ne sont pas consistantes avec les conditions d'intégrité précédemment définies. Ainsi, les systèmes de management des bases de données peuvent soit ne pas les prendre en compte, soit les tolérer. Les premières ont la propriété de produire en sortie des résultats crédibles dans la plupart de cas. Par contre les systèmes tolérant les inconsistances ne sont pas crédibles, sauf si on est capable de quantifier les inconsistances et de les prendre en compte.

2.1.2 Mesures de l'inconsistance dans les bases de données

Une fois les contraintes d'intégrité définies, des mesures peuvent être définies afin de pouvoir quantifier les éventuelles inconsistances.

Soit un opérateur d'ordre $\preceq$ ayant les propriétés d'antisymétrie, réflexivité et transitivité.

> **Définition 13 : Une mesure d'inconsistance** μ est définie sur des n-uplets (D, CI) et fournit une valeur dans un treillis partiellement ordonné par cet opérateur d'ordre $\preceq$ [Decker 09].

Ainsi, pour tout état d'une base de données D satisfaisant la théorie d'intégrité TI, $D(TI) = vrai$, tout autre état de cette base de données $D^{'}$ et toute autre théorie d'intégrité $TI^{'}$ on a :

$$\text{Si} \quad D^{'}(TI^{'}) = faux \quad \Rightarrow \quad \mu(D, TI) \prec \mu(D^{'}, TI^{'}) \tag{2.2}$$

De plus, pour chaque (D, TI) et pour tout autre $(D^{'}, TI^{'})$ on a :

$$\text{Si} \quad D(TI) = vrai \quad \Rightarrow \quad \mu(D, TI) \preceq \mu(D^{'}, TI^{'}) \tag{2.3}$$

Des exemples concrets de mesures d'inconsistance sont décrits dans [Decker 08] et [Grant 06].

La définition des mesures d'inconsistance n'est pas suffisante. Il faut également les utiliser pour prévenir des dégradations de qualité suite à des mises à jour. Ainsi, dans [Decker 09] il a été montré que chaque mesure d'inconsistance induit une méthode sensée vérifier l'intégrité. Si les méthodes de vérification de l'intégrité intolérantes aux inconsistances demandent que la condition d'intégrité $D(TI) = vrai$ soit totalement satisfaite, les méthodes tolérantes demandent une satisfaction partielle. Ainsi, au lieu d'utiliser des contraintes universelles, il est plus intéressant de pouvoir distinguer parmi plusieurs cas de violations d'une contraintes. En utilisant cette stratégie, des mesures d'inconsistance différenciées peuvent être définies, c'est-à-dire plus/moins le nombre de contraintes violées, pire/meilleure est la consistance des données.

2.1.3 Catégories de contraintes d'intégrité

Ce paragraphe présente les entités sur lesquelles des contraintes d'intégrité peuvent être imposées. Pour la simplicité d'exposition, il n'est considéré que le cas de bases de données relationnelles.

Dans la littérature, il existe trois catégories de contraintes d'intégrité : du domaine d'attributs, de l'entité et référentielle. Ci-dessous, pour chacune de ces catégories sera donnée sa définition, ainsi qu'un exemple pour indiquer ce qui est mesuré.

> **Contrainte 1 (Intégrité du domaine [Ramakrishnan 03]) :** Soit $\mathcal{R}(f_1 : D_1, f_2 : D_2, \ldots, f_n : D_n)$ un schéma relationnel pour lequel chaque attribut f_i, $1 \leq i \leq n$ prend des valeurs dans l'ensemble Dom_i correspondant au domaine D_i. Une instance de la relation $\mathcal{R}$ qui respecte les contraintes d'intégrité du domaine est un ensemble de n-uplets :
>
> $$\{(f_1 : d_1, f_2 : d_2, \ldots, f_n : d_n) \quad | \quad d_1 \in Dom_1, d_2 \in Dom_2, \ldots, d_n \in Dom_n\} \tag{2.4}$$

Un schéma relationnel spécifie les domaines de chaque champ ou colonne de l'instance de la relation. Ces contraintes d'intégrité du domaine ont pour rôle d'imposer à toutes les valeurs de la même colonne un seul domaine de définition. D'un point de vue informatique, du langage de programmation, le domaine d'un attribut se traduit par le *type* de cet attribut.

Exemple 2 : Soit une relation ETUDIANT décrivant les étudiants d'une université. Elle est définie par le prédicat $p(ID, NomPrenom, Age, NumTel, GPA)$. Les cinq attributs contiennent l'identifiant ID, le nom et le prénom $NomPrenom$, l'âge Age, le numéro de téléphone $NumTel$ et le GPA GPA de chaque étudiant enregistré. Dans le cas de cet exemple, les contraintes suivantes pourraient être déclarées :

- ID un String de 6 caractères ;
- $NomPrenom$ un String de maximum 40 caractères ;
- Age un nombre entier compris entre 16 et 100 ;
- $NumTel$ un nombre entier contenant exactement 10 chiffres ;

– GPA un nombre réel à deux décimales compris entre 0.00 et 4.00.

Pour résumer cette contrainte d'intégrité, les valeurs des attributs d'une relation doivent prendre des valeurs autorisées. Parmi les violations les plus rencontrées de cette contraintes il y a :

– des valeurs manquantes ;
– des valeurs par défaut NULL ;
– des valeurs non-uniques (s'il y a obligation) ;
– des valeurs qui ne font pas partie d'un ensemble de valeurs numériques ou alphanumériques valides ;
– des valeurs (continues) qui sont en dehors de l'intervalle numérique imposé.

En reprenant l'Exemple 2, on peut immédiatement observer qu'il faut absolument que deux étudiant n'aient pas le même ID. Il s'agit d'une contrainte d'intégrité sur les entités qui impose qu'un certain sous-ensemble minimal d'attributs (appelés *attributs primaires*) soit l'identifiant unique pour un n-uplet. Un tel sous-ensemble est appelé dans la théorie de base de données *clé primaire*, ou simplement *clé*, pour la relation respective.

Contrainte 2 (Intégrité de l'entité) : Si l'attribut a_i d'une relation $\mathcal{R}(\overrightarrow{a})$ est un attribut primaire, a_i ne peut accepter que des valeurs non-NULL.

Ainsi il y a deux cas de violation de la contrainte d'intégrité de l'entité : le premier se produit quand la valeur de la clé primaire est nulle et le deuxième quand les valeurs de la clé primaire ne sont pas uniques. Dans le cas de l'Exemple 2, l'attribut ID est la clé primaire de cette relation.

Très souvent les données contenues dans une relation sont reliées à d'autres données faisant partie d'une autre relation. Dans ce cas, lorsqu'une relation est modifiée, il est nécessaire de vérifier l'autre pour vérifier les éventuelles inconsistances engendrées. Ainsi, il faut définir des contraintes d'intégrité pour les deux relations ensembles. Dans la théorie de bases de données ce type de contraintes d'intégrité fait appel à une *clé étrangère*.

Une clé étrangère est un attribut ou un ensemble d'attributs d'une relation $\mathcal{R}_1(\overrightarrow{a_1})$, tel que la valeur de chaque attribut de cet ensemble est celui de la clé primaire de la relation $\mathcal{R}_2(\overrightarrow{a_2})$.

Si la relation $\mathcal{R}$ contient des références à une autre relation $\mathcal{S}$, on doit vérifier les liaisons entre les n-uplets de $\mathcal{R}$ et ceux de $\mathcal{S}$ sont valides. La contrainte d'intégrité suivante, agissant sur la clé étrangère, garantit que les références de n-uplets de $\mathcal{R}$ vers les n-uplets de $\mathcal{S}$ sont définies d'une manière non-ambigüe [Desai 90] :

Contrainte 3 (Intégrité référentielle) : Soit la relation $\mathcal{R}$ faisant référence à la relation $\mathcal{S}$ via un ensemble d'attributs formant la clé primaire de la relation $\mathcal{S}$ et que cet ensemble d'attributs forme une clé étrangère dans la relation $\mathcal{R}$. Dans ce cas, la valeur de la clé étrangère d'un n-uplet de $\mathcal{R}$ doit être, soit égale à la clé primaire d'un n-uplet de $\mathcal{S}$, soit entièrement nulle.

Ainsi, si un attribut a_i de la relation $\mathcal{R}$ est défini dans le domaine D_i et la clé primaire de la relation $\mathcal{S}$ est aussi définie dans le domaine D_i, alors les valeurs de a_i dans les n-uplets de $\mathcal{R}$ doivent être soit nulles, soit égales à la valeur v de la clé primaire d'un n-uplet de $\mathcal{S}$. La contrainte d'intégrité référentielle est très importante dans la pratique parce qu'elle impose l'existence d'un n-uplet de la relation qui correspond à l'instance de l'entité référencée. De plus, cette contrainte indique d'une manière implicite les actions pouvant être entreprises lors d'opérations de mise à jour. Ainsi, si un n-uplet référencé par une clé étrangère est supprimé, il existe trois possibilités pour garder l'intégrité de la base de données [Desai 90] :

– Tous les n-uplets ayant des références vers le n-uplet supprimé devraient être également supprimés. Ainsi, comme conséquence, une suppression détermine d'autres suppression en cascade ;
– Seulement les n-uplets non-référencés par d'autres n-uplets peuvent être supprimés ;
– Le n-uplet est supprimé, mais afin de ne pas avoir un effet de domino les attributs des clés primaires pertinentes de tout n-uplet référencé sont mis à valeur nulle.

2.2 Les imperfections des données

Dans le paragraphe précédent, les contraintes d'intégrité ont été introduites comme outil de préservation de la consistance des bases de données en identifiant et traitant les n-uplets avec des problèmes. Dans ce paragraphe, nous traitons les imperfections des **valeurs** des données et les techniques qui peuvent être mises en place afin de pouvoir les modéliser et les prendre en compte lors du traitement.

Soit l'exemple d'un attribut dont la valeur exacte n'est pas connue. Dans cette situation, un système traditionnel de base de données met une valeur par défaut ou la valeur par défaut NULL (voir le paragraphe 2.2.3). Les systèmes modernes essaient de prendre en compte les différentes imperfections. Ainsi, il est plus intéressant de mettre, au lieu d'une valeur par défaut, un intervalle qui contient la valeur exacte.

[Motro 96] présente parmi les imperfections les plus importantes des valeurs de données : l'erreur, l'incomplétude, l'imprécision et l'incertitude. Ces différents types d'imperfection sont détaillés ci-après.

2.2.1 Les données erronées

Ce type d'imperfection est le plus souvent rencontré. Une valeur erronée est une valeur différente de celle réelle. Il s'agit d'un cas très important et très fréquent. L'existence de contraintes d'intégrité trop strictes peut entrainer comme conséquence la saisie de valeurs erronées. Ainsi, certaines applications obligent le remplissage d'attributs. Soit l'exemple d'une base de données d'une compagnie qui loue des voitures. Chaque client doit fournir ses informations personnelles incluant le code postal de son adresse. Mais chaque pays a un système propre de définition du code postal. Ainsi, très souvent, un code postal étranger sera refusé, mais comme il est impératif d'en saisir un, un code aléatoire, donc erroné, sera choisi (par exemple le code postal de la compagnie louant les voitures).

Un cas particulier de données erronées est celui des valeurs aberrantes, appelées dans la littérature spécialisée *outliers*. Les outliers sont des erreurs grossières de mesure. Selon [Pearson 05] :

> **Définition 14 : Un outlier** est une valeur d'un attribut d'une base de données qui est anormale par rapport au comportement observé dans la majorité des autres valeurs de cet attribut.

Cette définition donne les éléments nécessaires pour détecter si une valeur correspond à un outlier. Pour cela, il faut :

- caractériser le comportement nominal, c'est-à-dire le non-anormal ;
- définir un critère quantitatif permettant de décider si une valeur est en conflit significatif avec cette caractérisation nominale.

En fonction du contexte d'application, il existe différents types de comportements normaux, d'où l'existence de différents types d'outliers et de différents algorithmes de détection. Le cas le plus simple et un des plus fréquent dans la pratique est celui des outliers uni-variés. Prenons le cas d'une séquence réelle $\{x_k\}$. Dans ce cas la technique de détection des outliers est basée sur l'hypothèse que les valeurs $\{x_k\}$ peuvent être décrites par une distribution de probabilité. Dans la plupart des situations le comportement nominal est décrit par une distribution gaussienne de moyenne μ et de variation σ^2. Sous cette hypothèse, une valeur $\{x_i\}$ est déclarée un outlier en appliquant la règle de 3σ : si la valeur $\{x_i\}$ n'appartient pas à l'intervalle $[\mu \pm 3\sigma]$ il s'agit d'un outlier. Une hypothèse moins restrictive est de considérer des distributions de probabilité symétriques de moyenne et de variance finie. Toute valeur qui est anormalement éloignée du centre du cluster des données est déclarée outlier.

En fonction du contexte d'application, l'apparition des outliers peut être la résultante de différentes sources, par exemple une défaillance technique ordinaire d'un équipement. Dans ce cas, les outliers n'apportent pas une information utile et donc, ces valeurs ne doivent pas être prises en

compte lors du traitement de données [1]. Cependant, dans d'autres situations, les outliers portent une information très intéressante caractérisant des phénomènes rares (phénomènes astronomiques, attaques cybernétiques, etc.).

[Pearson 05] décrit en détail les effets négatifs des outliers et les méthodes statistiques qui peuvent être employées afin de les détecter et d'appliquer le traitement correspondant.

2.2.2 Les données incomplètes

Les données incomplètes sont une anomalie qui affecte surtout les bases de données volumineuses. Ce type d'imperfection décrit 2 situations possibles :

- *Des attributs manquants* : dans cette situation, il existe des attributs pertinents (c'est-à-dire portant de l'information) qui n'ont pas été enregistrés. Il se peut également que l'ensemble des attributs pertinents n'est pas connu. Dans la littérature, il existe un nombre important d'algorithmes capables de traiter ce genre de données et particulièrement dans le domaine « business intelligence ». [Kim 13] présente des méthodes statistiques de traitement adaptées à ce domaine.
- *Des valeurs manquantes* : dans cette situation, un attribut pertinent et connu a été mesuré mais il existe des valeurs manquantes pour certains n-uplets. Deux cas peuvent être distingués :
  - l'absence d'une valeurs existante mais qui n'est pas enregistrée dans la base de données ;
  - la valeur n'existe pas.

  Un exemple classique illustrant ces deux cas est celui du domaine des assurances, en demandant l'état civil et le type d'assurance du conjoint(e). Chaque personne a un état civil et donc chaque n-uplet doit avoir une valeur de cet attribut (premier cas). Mais tout le monde n'a pas obligatoirement un(e) conjoint(e) (deuxième cas).

Une des causes les plus fréquentes est le non-fonctionnement des systèmes de mesure. Ainsi, il est possible que sur une période de temps, les valeurs des données ne soient pas enregistrées. Les données issues d'un système de collecte manuelle sont aussi susceptibles d'être incomplètes. [Adriaans 96] montre que la fréquence des valeurs laissées non-complétées est inversement proportionnelle à l'importance des attributs.

Les conséquences des données incomplètes dans la pratique dépendent de la *proportion* des valeurs manquantes et de leur *type*. Il peut y avoir des données manquantes qui n'affectent pas les traitements suivants. Par exemple si on a suffisamment données (échantillons), on peut filtrer les données manquantes et ne prendre en compte que les données complètes. Cependant, il peut y avoir des données manquantes qui, en les ignorants, introduisent des biais assez importants. C'est le cas des données qui sont systématiquement manquantes. Donc, il faut les détecter et les mesurer afin de pouvoir les prendre en compte dans les futurs traitements.

2.2.3 Les données imprécises

Une donnée imprécise représente la situation dans laquelle sa vraie valeur ne peut pas être déterminée qu'avec approximation. Par rapport aux données erronées, les données imprécises n'affectent pas la consistance de la base de données. Parmi les types d'imprécision les plus communs, [Motro 95] identifie :

- les valeurs *disjonctives* : la valeur réelle fait partie d'un ensemble fini $\{v_1, v_2, \ldots, v_n\}$;
- les valeurs *négatives* : la valeur réelle n'est pas v ;
- les valeurs *dans un intervalle* : la valeur réelle est dans l'intervalle $[v - \epsilon, v + \epsilon]$;
- les valeurs *NULL* : une valeur NULL signifie que la vraie valeur n'est pas disponible. Elle peut être vue comme une valeur imprécise avec l'ensemble de valeurs possible équivalent au domaine de définition de l'attribut (voir le paragraphe 2.1.3).

1. Sauf si ces données sont utilisées pour la détection d'une panne

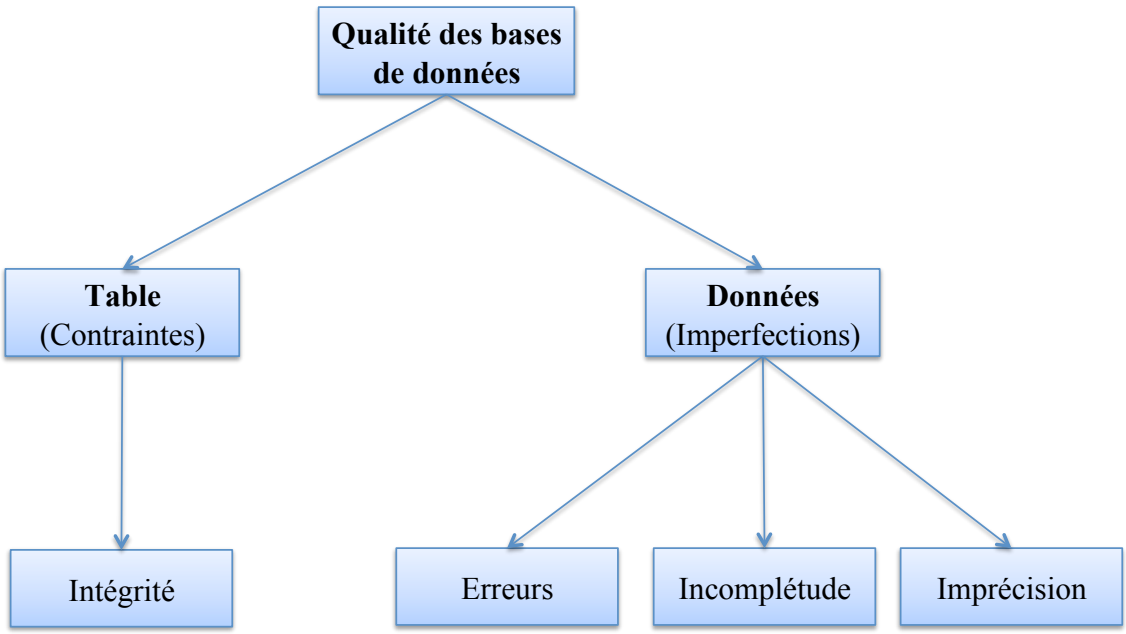


FIGURE 2.1: La qualité des données d'une base de données

Pour chaque type d'imprécision, il existe des méthodes mathématiques spécifiques capables de mesurer ce type d'imperfection. Un cadre mathématique généralisé dans lequel peuvent se définir de mesures d'imprécisions, est présenté dans l'annexe A.

2.3 TAXONOMIE DE LA QUALITÉ DES DONNÉES

Dans les paragraphes précédents, les caractéristiques les plus importantes de la qualité des données ont été présentées. Dans la figure 2.1, nous proposons une récapitulation des dimensions les plus importantes de la qualité des données sous la forme d'une taxonomie. De cette représentation, il peut s'observer que la qualité des données enregistrées dans des bases de données caractérise, d'une part comment elles sont enregistrées (qualité de la table) et d'autre part, ce qui est enregistré (qualité des valeurs de données). Dans cette étude, l'hypothèse de départ est que le système d'information est déjà développé et prêt à être mis en place. Ainsi, par la suite l'accent sera mis sur la qualité des valeurs de données, les bases de données étant considérées fonctionnelles avec toutes les contraintes d'intégrité définies et implémentées.

En ce qui concerne la qualité des valeurs de données, appelée par la suite simplement qualité des données, il peut s'observer qu'elle est multidimensionnelle, c'est-à-dire les données peuvent être affectées par plusieurs sources de défauts. En fait, chaque attribut est plus ou moins susceptible d'avoir un même type de problème de qualité pour tous les n-uplets, à cause de l'utilisation du même processus de mesure [2] (avec ces limitations).

Dans le paragraphe suivant, il est proposé d'étendre le modèle relationnel afin de pouvoir prendre en compte la qualité des données dès leur enregistrement.

2. Utilisation du même capteur, opérateur humain, etc.

2.4 MODÉLISATION DES IMPERFECTIONS DES DONNÉES DANS LE MODÈLE RELATIONNEL

Les systèmes de bases de données relationnelles commerciaux classiques (par exemple DB2 d'IBM, SQL Server de Microsoft ou encore Oracle DBMS d'Oracle) partent de l'hypothèse que les données enregistrées sont correctes et d'une qualité parfaite [Wang 02]. Mais, comme indiqué au paragraphe précédent, les données sont loin d'être parfaites à cause des sources multiples. Dans la pratique, les systèmes d'information utilisant ces données sont habituellement responsables de la prise en compte des éventuels problèmes de qualité. La plupart des chercheurs insistent que la responsabilité de la mesure et de la modélisation de la qualité des données doit revenir au système de management des données. Ainsi, les recherches menées récemment ont essayé d'intégrer au modèle relationnel une autre dimension, celle de la qualité des données : voir par exemple les projets MystiQ [Boulos 05], ORION [Cheng 05] et Trio-One [Mutsuzaki 07].

L'estimation la plus simple de la qualité des données peut se faire au niveau relationnel. Dans ce cas, les mesures de qualité sont définies en appliquant des modèles statistiques comme par exemple : la proportion de valeurs manquantes, la proportion de valeurs de NULL, etc. Le problème avec cette approche est que toutes les données (n-uplets) d'une même relation vont être caractérisées avec la même qualité. Par conséquent, ce type de modélisation de la qualité des données ne peut pas prendre en compte l'hétérogénéité au niveau des n-uplets. Une autre possibilité est d'essayer modéliser la qualité au niveau de n-uplet. Mais, dans ce cas tous les attributs d'un même n-uplet vont avoir la même qualité. Cependant, comme il a déjà été précisé dans le paragraphe 2.3 chaque valeur d'attribut peut être la résultante d'un autre processus de mesure ou d'une autre source. Ainsi, cette méthode est également insuffisante. En conséquence, la procédure la plus adaptée est d'essayer d'étiqueter chaque attribut avec un ensemble de paramètres de qualité.

Le modèle de qualité présenté à la figure 2.1 montre qu'une valeur d'un attribut peut être associée à un ensemble de paramètres de qualité. Cette liste de paramètres de qualité est non-exhaustive et peut être augmentée selon les besoin du contexte de l'application. Ainsi, il est possible d'ajouter des paramètres de qualité d'ordre supérieur ayant pour rôle de décrire la qualité avec laquelle les paramètres de qualité de premier ordre ont été évalués.

Exemple 3 : : Soit une source de données fournissant des valeurs avec une déviation standard $\pm\epsilon$, c'est-à-dire la valeur réelle v d'une donnée issue de cette source sera dans l'intervalle $[v - \epsilon, v - \epsilon]$. Néanmoins, il existe des situations où cette représentation sous la forme d'un intervalle de l'imprécision de données n'est pas suffisante. Ainsi, un autre paramètre de qualité est usuellement utilisé afin de compléter le précédent : le degré de *confiance* que la vraie valeur se situe dans cet intervalle. Ce dernier est un exemple de paramètre de qualité d'ordre supérieur (d'ordre deux dans ce cas).

Donc, *a priori*, un attribut peut être associé à un nombre arbitraire de niveaux de paramètres de qualité. Cette observation est illustrée dans la figure 2.2. Ainsi, un attribut de qualité est caractérisé par ces paramètres de qualité de premier ordre (de niveau 1 - N1). Si un de ces paramètres ne peut pas représenter tous les aspects de cette dimension de la qualité, des paramètres d'ordre deux peuvent lui être associés (de niveaux 2 - N2). Et si ces derniers ne sont encore pas suffisants, des paramètres d'ordre 3 peuvent lui être associés. Ce raisonnement en niveau de qualité peut être itéré jusqu'au moment où la dimension de qualité a été complètement modélisée.

Une telle modélisation de la qualité des données stockées dans des bases de données relationnelles pose des problèmes techniques. Une solution élégante a été proposée dans [Wang 95]. Dans cette étude, il est proposé d'étendre la notion d'attribut à une paire ordonnée, appelée *attribut de qualité*. Celui-ci est composé de l'attribut lui-même et d'une *clé de qualité*. Cette clé de qualité a pour rôle de référencer les paramètres de qualité des niveaux supérieurs. Ainsi, la clé de qualité sert comme clé étrangère. Cette idée est représentée à la figure 2.3. Chaque attribut a de la base de données relationnelle est étendu à la paire ordonnée $< a, \underline{Q(a)} >$, dans laquelle $\underline{Q(a)}$ est la clé de qualité. Celle-ci, ensuite, fait référence à la table de qualité correspondante. Dans cette table, les attributs consistent en paires $< q_i, \underline{Q(q_i)} >$, avec q_i un paramètre de qualité de $Q(a)$ et avec

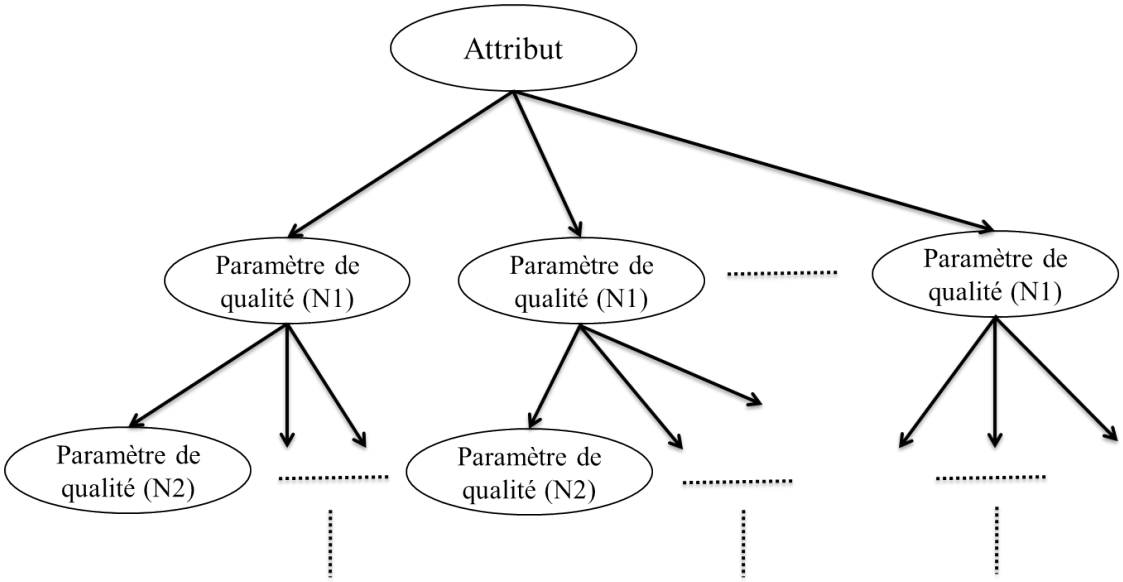


FIGURE 2.2: Association de différents niveaux de qualité pour un attribut

$\underline{Q(q_i)}$ la clé de qualité de q_i qui développe encore cette notion de qualité (un niveau en plus). Il est évident que cette arborescence ne peut pas aller à l'infini. Ainsi, lorsqu'une clé de qualité prend une valeur prédéterminée, par exemple *nill* dans [Wang 95], le référencement vers un autre niveau s'arrête.

Toujours dans la même étude, [Wang 95] a défini les conditions d'intégrité que le nouveau modèle relationnel doit respecter. Ainsi, la valeur d'un attribut et ses valeurs de qualité correspondantes (contenant tous les niveaux) doivent être traitées comme une entité atomique. En conséquence, chaque fois qu'une valeur d'un attribut est créée ou modifiée, son correspondant de qualité doit être aussi créé ou modifié en concordance avec ce changement. En plus des contraintes d'intégrité, les opérations algébriques de traitement de données doivent être également redéfinies [Wang 95] et [Wang 02].

Dans le paragraphe suivant, nous présentons deux types particuliers de modèles relationnels incorporant l'imprécision des données : le modèle probabiliste et le modèle possibiliste. Principa-

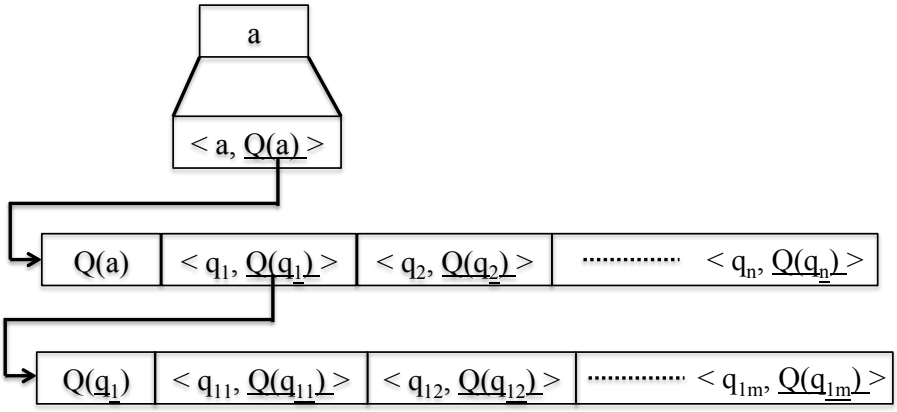


FIGURE 2.3: Extension du modèle relationnel : attributs de qualité

Clé	Indépendant	Interdépendants	Indépendant
	Déterministe	Stochastique	Stochastique
Nom du salarié	Département	Proba. [Qualif_Travail / Bonus]	Salaire
Jean	Maison	0.2 [Très bien / Oui] 0.6 [Bien / Oui] 0.2 [Satisfaisant / Non]	0.4 [40-50 k€] 0.6 [50-60 k€]

TABLE 2.1: Exemple d'une table probabiliste

lement, il s'agit de remplacer une valeur d'une donnée affectée d'imprécision par l'ensemble (fini) des valeurs possibles, accompagné d'indices de qualité.

2.4.1 Le modèle probabiliste de bases de données

Le modèle probabiliste suppose que la valeur d'un attribut a est une variable X_a et que cette variable est décrite par une distribution de probabilité P_{X_a}. Comme toute distribution de probabilité, P_{X_a} assigne des valeurs comprise dans l'intervalle unitaire $[0, 1]$ aux éléments du domaine de définition de l'attribut a. La condition de normalisation implique que la somme des valeurs assignées soit égale à 1.

Afin de respecter les contraintes d'intégrité sur les clés, voir le paragraphe 2.1.3, il est supposé que les attributs composant la clé de la relation ne sont pas affectés d'imprécision et ont un comportement déterministe. Les attributs n'appartenant pas à la clé peuvent avoir une nature déterministe ou stochastique. Les attributs déterministes ne sont pas affectés par des incertitudes, ainsi ils sont supposés avoir une qualité parfaite. Cependant, les attributs stochastiques sont affectés par différents types d'incertitudes et ils sont sensibles aux imperfections comme celles décrites dans le paragraphe 2.2.

Ci-dessous on présente une définition d'une relation probabiliste selon [Barbara 92] :

> **Définition 15 :** Soit K un ensemble d'attributs $K_1, \ldots, K_n$ formant la clé primaires de la relation $\mathcal{R}$ et A un ensemble d'attributs $A_1, \ldots, A_m$. Les domaines de ces attributs sont donnés par $dom(K_i)$ et $dom(A_j)$. Une relation probabiliste de $\mathcal{R}$ est une fonction de $dom(K_1) \times \ldots \times dom(K_n)$ vers **PF**, une famille de distributions de probabilité pour A. Chaque $p \in \mathbf{PF}$ est une fonction de $dom(A_1) \times \ldots \times dom(A_m)$ dans l'intervalle unitaire, vérifiant la propriété de normalisation :
>
> $$\sum_{a \in dom(A_1) \times \ldots \times dom(A_m)} p(a) = 1 \tag{2.5}$$

Dans le tableau 2.1 est présenté un exemple adapté de [Barbara 92]. Il s'agit d'une relation probabiliste contenant les salariés d'une entreprise. La clé primaire ***Clé*** contient les noms des salariés. L'attribut *Département* est déterministe et indépendant des autres. Les attributs *Qualif_Travail* et *Bonus* sont stochastiques et interdépendants, donc il a été choisi de les représenter dans une seule colonne. En effet, la valeur de l'attribut *Qualif_Travail* influence la valeur de l'attribut *Bonus*. Donc, pour ces deux attributs, il s'agit d'une probabilité conjointe, conditionnée par le salarié : $Pr\{\text{Qualif\_Travail}, \text{Bonus} | \text{Nom du Salarié}\}$. Par exemple :

$$Pr\{\text{Qualif\_Travail} = \text{Très bien}, \text{Bonus} = \text{Oui} \mid \text{Nom du Salarié} = \text{Jean}\} = 0.2 \tag{2.6}$$

Le dernier attribut, *Salaire* est stochastique et indépendant des autres. Sa distribution de probabilité est conditionnée par le nom du salarié.

Le principal problème avec ce modèle probabiliste est la construction de la distribution de probabilité. En conséquence, il se peut que l'on ait pas toutes les valeurs de probabilité pour

l'ensemble du domaine de l'attribut ou qu'il existe une incertitude vis-à-vis de l'exactitude de la distribution de probabilité. Dans ce cas, un autre niveau de qualité peut être ajouté : la *confiance* dans le modèle de la distribution de probabilité, situation illustrée dans le schéma de qualité figure 2.2.

Par définition, les distributions des probabilités ont pour rôle la modélisation l'imprécision. Par conséquent, les bases de données probabilistes sont bien adaptées pour modéliser les données imprécises, voir le paragraphe 2.2. L'exemple du tableau 2.1 contient trois attributs dont des valeurs de probabilité ont été associées à chaque valeur d'attribut disjonctive.

Nous ne rentrons pas dans plus des détails concernant ce type de modèle probabiliste. Néanmoins, nous recommandons l'article [Barbara 92] comme étant une très bonne et complète référence. Dans cette étude, l'accent est mis sur le développement des techniques de modélisation dans le cas de distributions de probabilité partiellement connues. Aussi, ils proposent une algèbre relationnelle afin de pouvoir faire des requêtes à ce type de base de données.

2

2.4.2 Le modèle possibiliste de bases de données

La théorie des possibilités est fondée sur la théorie des ensembles flous [3]. Un ensemble flou F est un ensemble d'éléments dont à chaque élément il est associé une valeur dans l'intervalle unitaire $[0, 1]$. Cette valeur représente le degré d'appartenance à l'ensemble.

Les domaines de définition des attributs des bases de données relationnelles floues peuvent être représentés soit par des scalaires discrets (ex. $\{\text{petit}, \text{moyen}, \text{grand}\}$) soit par des nombres discrets d'un ensemble fini ou infini. Dans une base de données relationnelle floue, il n'est pas nécessaire que les valeurs des attributs soient atomiques [Buckles 82]. Ainsi, les attributs a_i prennent des valeurs dans le sous-ensemble du domaine de définition de cet attribut $dom(a_i)$. Donc, tout élément de l'ensemble des parties de l'ensemble $2^{dom(a_i)}$ pourrait être une valeur de l'attribut a_i. Ci-dessous, les définitions d'une relation floue et d'un n-uplet flou sont données selon [Buckles 82].

Définition 16 : Une relation floue $\mathcal{R}$ est un sous-ensemble du produit croisé des ensembles $2^{dom(a_1)} \times 2^{dom(a_2)} \times \ldots \times 2^{dom(a_n)}$.

Définition 17 : Un n-uplet flou est un membre de $\mathcal{R}$ et du produit croisé des ensembles $2^{dom(a_1)} \times 2^{dom(a_2)} \times \ldots \times 2^{dom(a_n)}$.

Sachant qu'un objet (valeur d'un attribut) est un membre d'un ensemble flou, l'objectif de la théorie des possibilités est de déterminer la *possibilité* qu'une valeur spécifique s'applique [Klir 88]. Ainsi, pour une valeur particulière d'un attribut d'un n-uplet, a_i, une distribution de possibilité π_{a_i} lui est associée.

Dans le tableau 2.2 est présentée une relation possibiliste. Elle contient trois attributs : le *Nom*, l'*Âge* et le *Salaire* d'un employé. Comme dans le cas probabiliste, la clé primaire de cette relation, *Nom*, est considérée comme étant déterministe et sans incertitude. Deux n-uplets (entités) sont identifiés afin d'illustrer la notion de distribution de possibilité. L'attribut *Salaire* du premier n-uplet est modélisé par une distribution de probabilité de la forme :

$$\pi_{Salaire(Jean)} = \begin{cases} 39 & : 0.8 \\ 40 & : 1.0 \\ 41 & : 1.0 \\ 42 & : 0.5 \end{cases} \tag{2.7}$$

L'interprétation de la distribution de possibilité de l'équation 2.7 est que la valeur du salaire de Jean est « totalement » possible d'être 40 ou 41 k€, « fortement » possible d'être 39 k€, « moyen-

3. Pour une introduction mathématique dans la théorie des ensembles flous et dans la théorie des possibilités voir l'annexe A

Nom	Âge	Salaire [k€]
Jean	28	{39/0.8 , 40/1.0 , 41/1.0 , 42/0.5 }
Marie	{29/0.7 , 30/1.0 , 31/0.8 }	42

TABLE 2.2: Exemple d'une table possibiliste

nement » possible d'être 42 k€et complètement impossible d'avoir une autre valeur. Les mêmes observations peuvent être faites pour le cas de l'attribut *Âge* du deuxième n-uplet.

Une observation importante est maintenant faite. Très souvent, ce n'est pas possible de quantifier numériquement une valeur d'un attribut. Dans ce cas, il est préférable d'utiliser des notions linguistiques, par exemple {*bas* , *moyen* , *haut* }. Un des avantages de l'utilisation de distributions de possibilités est leur adaptation à la modélisation [Motro 95] :

- des termes linguistiques ambigus (exemple : Âge = *jeune*) ;
- des valeurs disjonctives (exemple Département = {*Avant vente* , *Après vente* }) ;
- des valeurs de NULL : toutes les valeurs du domaine sont possibles ;
- des valeurs simples : une seule valeur du domaine est possible.

Afin de pouvoir manipuler ce genre de bases de données il faut, comme dans les autres cas précédemment présentés, définir une algèbre relationnelle. Ceci est hors sujet dans ce mémoire de thèse et donc, sont recommandées les références [Buckles 84], [Bosc 96] et [Ma 06].

2.5 CONCLUSION

Ce chapitre a été consacré à la définition de la qualité des données stockées dans des bases de données relationnelles. Deux aspects de la notion de qualité ont été présentés :

- le premier représente la qualité des relations composant la base de données ;
- le deuxième représente les imperfections des données stockées.

Par chacun de ces deux aspects, les principales dimensions de qualité ont été présentées et un exemple a été donné. Ainsi, la principale dimension de qualité pour les relations est la consistance. Afin de garder la consistance des bases de données, plusieurs critères d'intégrité, exprimés sous la forme des contraintes logiques, ont été définis.

La consistance des bases de données peut être définie et contrôlée par l'administrateur de la base de données, par contre en ce qui concerne les imperfections des données stockées, elles sont implicitement issues des limitations techniques, temporelles, etc. des sources/processus les générant. Ainsi, la procédure que nous recommandons de suivre est d'essayer de prendre en compte ces imperfections dès leur enregistrement en ajoutant, pour chaque valeur, son niveau de qualité. Deux solutions pratiques de représentation des imperfections ont été présentées dans ce sens, la première sous la forme d'une distribution probabiliste et la deuxième sous la forme d'une distribution de possibilité.

3 Qualité de l'information

3.1 L'importance de la qualité de l'information

Dans la littérature, il y a beaucoup d'exemples qui présentent les effets désastreux de l'utilisation des données et des informations de mauvaise qualité. Ainsi, le 28 janvier 1986 la navette spatiale Challenger lancée par la NASA a explosé après quelques secondes. La commission d'investigation a conclu que cette tragédie a été la conséquence d'un processus d'aide à la décision basé sur des informations incomplètes et confuses. Une autre tragédie ayant à la base l'utilisation d'information de mauvaise qualité est la destruction d'un avion commercial iranien par l'USS Vincennes en juillet 1988. Ces deux accidents ont été pris comme études de cas par [Fisher 01]. Un autre exemple est celui de la fusée Ariane 5, développée par ESA et lancée en 1996. Elle devait transporter des charges utiles sur l'orbite terrestre sans aucun humain à bord. Après 40 secondes, la procédure d'auto-destruction de la fusée a été initialisée et l'Ariane 5 a été complètement détruite. L'analyse après cet incident [Lions 96] a montré que le système inertiel de référence (IRS) a cessé de fonctionner. Ce système IRS utilisait 7 variables pour l'enregistrement de données en provenance de différents capteurs. Les données envoyées par les capteurs était codées sur 64 bits en point flottant. Le système IRS travaillait sur des nombres entiers signés codés sur 16 bits. Ainsi, les données en entrées du système IRS n'étaient pas compatibles à son fonctionnement et une conversion aurait du être effectuée avant de procéder à l'exécution du traitement. Malheureusement, chaque fois que la conversion n'était pas possible à réaliser, une erreur était levée et le système cessait de fonctionner. Ce problème est apparu parce que les ingénieurs ont voulu utiliser le software présent dans le projet Ariane 4 et en réalisant le minimum de changements dans le système IRS. De plus, si le système IRS avait été testé avec les autres composants, ce type d'erreurs aurait été facilement repéré et réparé. La conséquence est :

- La perte de 500 millions dollars investis dans la construction de la fusée ;
- La perte de la charge transportée ;
- Une mauvaise image du programme spatiale européen ;
- Une perte de temps.

Les systèmes utilisés dans la défense et dans l'aéronautique ne sont pas les seuls à être affectés par la qualité des données et des informations. Il en est de même pour ceux utilisés dans notre vie quotidienne. Ainsi, une étude faite en 2006 par l' « Institute of Medicine » [Weise 06] montre qu'aux États-Unis plus de 1,5 millions de patients reçoivent une fausse médication. Ce même rapport montre aussi que les surcoûts de traitements provoqués par une mauvaise médication dans les hôpitaux s'élèvent à plus de 3,5 milliards de dollars. Une étude faite en 1999 [Charatan 99b] montre que presque 100000 patients meurent chaque année aux États-Unis à cause d'un dysfonctionnement du système médical. Une autre [James 13], plus récente, présente des chiffres beaucoup plus alarmants avec une limite inférieure de 210000 décès annuels. Selon ces deux dernières études, une des causes principales est la mauvaise implémentation du système d'information médical. Ainsi, dans beaucoup de cas, les informations dont le médecin a besoin ne lui sont pas accessibles. De plus, si le système d'information mis en place n'accompagne pas les informations par un niveau de qualité, le médecin est dans l'impossibilité de savoir la confiance qu'il peut avoir dans ces informations.

Toujours dans le cadre médical, un autre exemple typique concerné par des problèmes de qualité est l'utilisation des prescriptions pharmaceutiques écrites à la main par les médecin. Comme il y a des médicaments avec des noms très proches, mais avec une composition chimique très différente, le pharmacien peut se tromper et donner une mauvaise médication au patient. Les conséquences d'une telle confusion peuvent être tragiques dans certains cas [Charatan 99a].

Les problèmes auxquels sont confrontés les organisations, dus à un manque d'information sur la qualité des données et des informations, ont été largement présentés dans la littérature. Ainsi, dans [English 09] est avancé le chiffre de 1212 milliards de dollars de pertes ou des coûts provoqués par 122 des plus grandes entreprises du monde. Cette estimation est beaucoup plus étonnante si on considère le fait que ces entreprises ont alloué des ressources importantes et qu'elles étaient conscientes de cette situation.

Ci-dessous sont présentées d'autres conclusions issues de différentes études :

- Plus de 60% des 500 entreprises de taille moyenne se sont retrouvées avec des problèmes de qualité des données et d'information [Wand 96] ;
- Entre 1% et 10% des données enregistrées dans les bases de données des organisations sont imprécises, selon [Klein 97] ;
- Un taux d'erreur sur les données allant jusqu'à 30% est souvent considéré comme normal dans l'industrie et ce taux peut aller dans certains situation jusqu'à 70% [Redman 96] ;
- L'étude de [Wang 96], proposant un des premières méthodologies d'évaluation de la qualité de l'information, a pris comme point de départ la situation d'une banque new-yorkaise identifiée comme ayant 40% de données incomplètes sur des données de management des crédits à risque ;
- Entre 50% et 80% des casiers judiciaires américains contiennent des données imprécises, incomplètes ou ambigües [Strong 97] ;
- En moyenne, la faible qualité des données et des informations implique des pertes de revenus entre 8% et 12% pour une entreprise. Elle est également responsable de 40% à 60% de dépenses supplémentaires pour une entreprise de services [Redman 98].

En analysant tous ces exemples, on peut déduire que les problèmes de qualité concernent toutes les organisations, qu'ils sont très coûteux et qu'ils peuvent avoir des conséquences catastrophiques. Également, on peut conclure que la procédure traditionnelle d'inspection et de correction des défauts des données [1] et des informations, bien que très utile et importante, a des limitations. Ainsi, il est nécessaire d'avoir une méthodologie capable d'évaluer la qualité et de la faire parvenir au divers modules de traitement, pour augmenter leurs performances, et au final, informer l'utilisateur sur la qualité des informations qui lui sont proposées. De plus, il est intéressant de pouvoir expliquer ce niveau de qualité afin d'offrir à l'utilisateur une image de l'évolution de la qualité de l'information dans le système, c'est-à-dire de l'informer sur sa provenance.

Comme les principaux effets de la qualité de l'information impactent la qualité du processus de prise de décisions, le prochain sous-paragraphe va introduire la corrélation entre la qualité de l'information et le résultat de la prise de décisions.

3.1.1 L'utilisateur et la qualité de l'information

Si dans le passé les utilisateurs avaient des difficultés à mener à bien leurs tâches à cause d'un manque d'information, de nos jours les utilisateurs sont bombardés d'informations de différents types et provenant de différentes sources. Ce basculement a été réalisé grâce aux progrès technologiques du domaine de l'informatique, software et hardware, et aux déploiements de réseaux très haut débit.

De nos jours, les responsabilités d'un opérateur humain sont le monitoring et l'accès à l'information, la prise de conscience de la situation actuelle, l'inférence des futures conséquences et en final la prise de décision [Atoyan 10]. L'implémentation des systèmes d'information comme systèmes

1. En anglais : « data profiling » et « data cleansing »

d'aide à la prise de décisions par les utilisateurs a beaucoup amélioré la capacité des utilisateurs à discerner les informations utiles dans cet amalgame d'informations dont ils ont accès. Dans le cas de situations critiques, l'utilisateur qui doit pouvoir prendre des décisions, pouvant avoir des conséquences catastrophiques, a besoin de vite comprendre les informations qui lui sont proposées par le système. Cette compréhension s'exprime non seulement par la façon de présenter les informations et la précision de celles-ci, mais également en indiquant la confiance qu'il peut avoir dans le système et dans les informations fournies. Ces caractéristiques traduisent la qualité de l'information et elles sont en directe corrélation avec les attentes de l'utilisateur vis-à-vis du système qu'il utilise.

L'utilisation des systèmes d'information peut se faire selon différents niveaux d'automatisation. Ainsi en fonction de la tâche à réaliser un système d'information peut aider l'utilisateur humain ou il peut le remplacer et faire la tâche lui même. En fonction du niveau d'automatisation, les systèmes d'information peuvent être classifiés [Sheridan 05] :

- l'utilisateur humain fait tout, c'est-à-dire le système d'information n'est pas utilisé ;
- le système d'information offre des alternatives ;
- le système d'information élimine une grande partie des alternatives possibles et propose à l'utilisateur un ensemble restreint d'alternatives ;
- le système d'information propose une seule alternative ;
- le système d'information exécute l'alternative si l'humain est d'accord ;
- le système d'information exécute l'alternative, mais l'humain a un droit de véto ;
- le système d'information exécute l'alternative et il informe l'humain ;
- le système d'information exécute un ensemble d'alternatives sélectionnées et il informe l'humain seulement s'il est demandé ;
- le système d'information exécute un ensemble d'alternatives sélectionnées et il informe l'humain seulement s'il le décide ;
- le système d'information fonctionne d'une façon complètement autonome.

A l'exception du premier et des deux derniers, dans tous les autres cas d'utilisation il est nécessaire d'évaluer les performances du système d'information, plus précisément la qualité de chaque information proposée par le système afin de permettre à l'utilisateur d'avoir à sa disposition l'information et le degré de confiance qu'il peut y avoir.

Pour que l'utilisateur soit confiant dans un système d'aide à la décision il est nécessaire qu'il comprenne l'ensemble des opérations subies par les données et les informations dans le système. Cependant, du fait d'un nombre important de traitements, la représentation visuelle n'est pas une solution car elle est très complexe. Dans ce but, [Shankaranarayanan 06] propose d'utiliser des métadonnées, comme indicateurs de qualité, à chaque niveau de traitement du système d'information. Cette étude montre que l'utilisation de ces méta-données rend l'utilisateur plus conscient de la qualité des données et des informations et donc finalement augmente sa confiance dans le système d'aide à la décision. La théorie cognitive sociale dit que dans les situations où l'utilisateur est confiant dans ces actions, il est en moyenne plus performant.

Par conséquent, la qualité de l'information joue un rôle très important dans l'adoption d'un système d'information par un utilisateur, dans son utilisation et dans son réutilisation [Laudon 11]. Quand les informations proposées à l'utilisateur ne sont pas argumentées, l'utilisateur peut rapidement perdre la confiance dans le système d'information. Ainsi, le système d'information est en danger de ne plus être utilisé. Mais lorsque les informations sont expliquées, c'est-à-dire accompagnées par de coefficients de confiance, l'utilisateur sera intéressé par les processus qui les ont produites. Dans ce cas, la qualité finale des décisions est enrichie et l'utilisateur aura plus de confiance dans les décisions proposées, c'est-à-dire dans le système d'aide à la décision.

3.2 Introduction sur la qualité de l'information

Une recherche bibliographique spécialisée nous donne une quantité impressionnante d'articles traitant le sujet de la qualité de l'information. Dans la figure 3.1 l'évolution de la recherche sur le sujet de la qualité des données et de l'information est présentée en nombre de publications

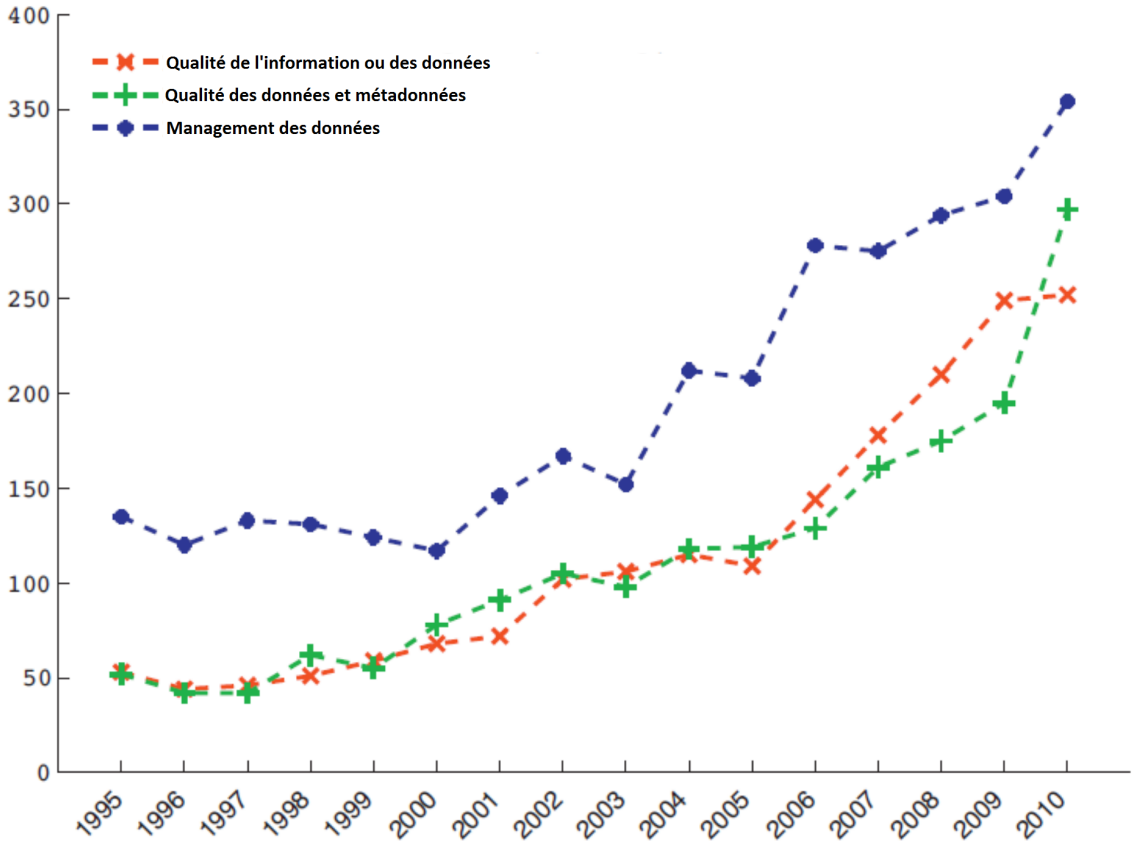


FIGURE 3.1: L'évolution du nombre de publications durant les dernières deux décennies de la recherche dans le domaine de la qualité des données et de l'information

annuelles [2]. Comme la notion de qualité de données et de l'information est très générale, en fonction du domaine de recherche/activité d'autres termes peuvent être employés. Ainsi, dans la figure 3.1, est présentée l'évolution pour trois catégories *différentes* : la qualité de l'information ou des données, la qualité de données et métadonnées et le management des données.

Comme observé dans le paragraphe précédent, la qualité de données et de l'information concerne tous les domaines de recherche. Afin de mieux illustrer l'évolution des recherches dans la modélisation et l'évaluation de la qualité de l'information, trois domaines sont considérés : les systèmes d'information d'aide au management des organisations (MIS [3]), les systèmes d'information adoptant les technologies Web (WIS [4]) et les systèmes avec fusion d'informations (IFS [5]). Dans la figure 3.2 sont présentés les plus importants articles de ces trois domaines, en terme d'impact c'est-à-dire de nombre de citations, ainsi que leurs interactions. De cette analyse, il peut s'observer que les premières recherches ont été faites dans le cadre du management des organisations. En effet à la fin des années 80, les organisations se sont rendues compte qu'elles investissaient beaucoup d'argent dans les nouvelles technologies de l'information et que ces dernières sont soit inutilisées soit d'une confiance incertaine. Ainsi, au début des années 90, beaucoup de recherches sur ce sujet ont été menées dans ce domaine. En analysant les interactions entre les citations trans-domaine, on peut observer que les deux autres domaines, WIS et IFS, ont adoptés des méthodologies développées

2. Selon ISI Web of Knowledge http ://apps.isiknowledge.com
3. de l'anglais Management Information System
4. de l'anglais Web Information System
5. de l'anglais Information Fusion System

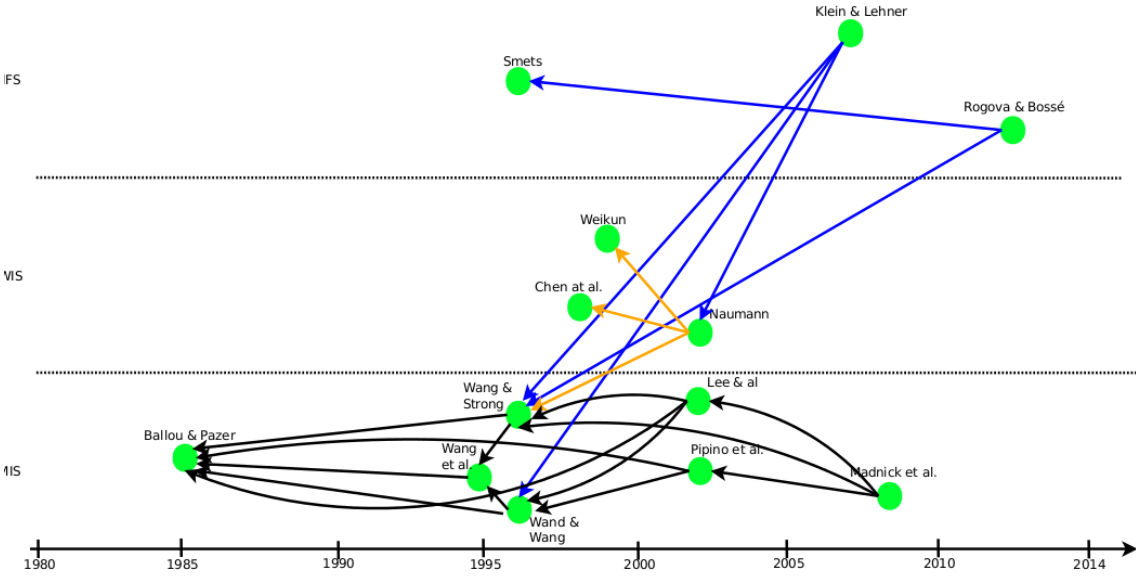


FIGURE 3.2: Les interactions entre les articles de référence dans la littérature concernant la modélisation et l'évaluation de la qualité de l'information

dans le cadre de MIS, plus précisément celles présentées dans les travaux de Richard Wang. Néanmoins, comme le contexte diffère, une adaptation particulière de ces méthodologies a été effectuée dans chacun des domaines.

Dans les trois paragraphes suivants, les définitions de la qualité des données et de l'information dans chacun de ces domaines seront présentées. De plus, l'influence du domaine des MIS sur les deux autres domaines sera détaillée afin de pouvoir, dans le paragraphe 3.7, tirer les conclusions sur l'état de l'art actuel de la qualité de l'information et mettre en évidence les points faibles méritant d'être étudiés et résolus.

3.3 LA QUALITÉ DE L'INFORMATION DANS LES MIS

À la fin des années 70, [Matlin 77] et [King 78] ont commencé les premiers travaux de recherche ayant comme sujet l'influence des systèmes d'information sur le fonctionnement de l'organisation. Ces recherches ont été développées pour des intérêts économiques particuliers et ils ont montré la nécessité d'utiliser plusieurs critères de qualité. Dans les années 80, les organisations (les entreprises) ont commencé à investir de plus en plus dans des systèmes d'information afin d'augmenter leur productivité et d'accélérer leur processus de prise de décisions. Cette adoption de nouvelles technologies a impliqué (et continue d'impliquer) des problèmes de confiance des utilisateurs envers les systèmes qu'ils utilisent.

Afin de répondre aux besoins des entreprises d'avoir des informations de très bonne qualité et suite à la réussite du programme de management de la qualité dans les processus industriels [Deming 82], le Dr. S. Madnick professeur à MIT Sloan School of Business, a initié au début des années 90 un programme de recherche en partenariat avec les entreprises appelé « Total Data Quality Management » (TDQM) [6]. L'objectif majeur de ce partenariat a été de créer un centre d'excellence responsable du développement d'une théorie sur la qualité des données et de l'information. Le plus grand succès de ce partenariat est le programme MIT Information Quality Program (MITIQ) [7] sous la direction de Dr. R. Wang. Ce programme a été le promoteur et le sponsor de la première

6. http ://web.mit.edu/tdqm/
7. http ://mitiq.mit.edu/

Acronyme	Nom	Référence	Flexible
TDQM	Total Data Quality Management	[Wang 96]	Non
TIQM	Total Information Quality Management	[English 99]	Non
AIMQ	A meth. for Information Quality Assessment	[Lee 02]	Non
DQA	Data Quality Assessment	[Pipino 02]	Oui
COLDQ	Cost-effect Of Low Data Quality	[Loshin 04]	Non
DaQuinCIS	Data Quality in Cooperative IS	[Scannapieco 04]	Oui
QAFD	Quality Assessment on Financial Data	[De Amicis 04]	Non
CDQ	Comprehensive meth. for Data Quality Management	[Batini 06]	Oui

TABLE 3.1: Huit méthodologies parmi les plus citées de la qualité de l'information

conférence internationale « International Conference in Information Quality », ainsi que du premier journal de ce domaine : « ACM Journal of Data and Information Quality ».

À l'heure actuelle, il existe plus de vingt méthodologies définissant la qualité de l'information [Knight 08]. Dans le tableau 3.1, nous présentons huit méthodologies parmi les plus citées dans la littérature. Chacune de ces méthodologies est fondée sur la définition d'un ensemble de dimensions de qualité. La dernière colonne du tableau 3.1 indique les méthodologies qui sont extensibles à d'autres dimensions de la qualité. Comme il s'agit de modèles théoriques de modélisation de la qualité, la grande majorité de ces méthodologies essaient de proposer une liste exhaustive afin de couvrir tous les aspects de la qualité. Dans le tableau 3.2 sont présentées les dimensions de qualité de ces méthodologies [8]. De ce tableau, on peut observer qu'une bonne partie des dimensions sont identiques pour plusieurs méthodologies : par exemple la *précision*, l'*obsolescence*, la *confiance*, etc. Cependant, il existe des dimensions avec un nom différent mais qui expriment la même chose. Ainsi, dans les méthodologies AIMQ et DQA la dimensions *sans erreur* est équivalente avec la *précision* des autres méthodologies. Une autre observation est, qu'en fonction du contexte dans lequel ces méthodologies ont été développées, certaines dimensions sont explicitement développées en plusieurs sous-dimensions. Par exemple, la dimension *précision* est évaluée au niveau sémantique et syntaxique dans la méthodologie QAFD.

La méthodologie TDQM est la plus utilisée dans la pratique et elle a servi de référence pour presque toutes les autres méthodologies. Ainsi, c'est celle-ci qui sera présentée par la suite. La méthodologie développée dans le cadre du programme TDQM est fondée sur une amélioration continue de la qualité des données et de l'information par un cycle du type *Définir*, *Mesurer*, *Analyser* et *Améliorer*, illustré dans la figure 3.3. Cette stratégie est inspirée de la méthodologie « Six-Sigma » [9]. Chacune de ces quatre étapes du processus TDQM est maintenant brièvement introduite.

Définir : Cette étape définit la qualité de l'information. Wang et Strong, dans leur étude [Wang 96], ont défini les dimensions de la qualité de l'information en interrogeant les utilisateurs et en ne gardant que les dimensions les plus communes. Comme cette définition de la qualité de l'information est la plus citée de la littérature, elle sera présentée plus en détail dans le paragraphe 3.3.1.

Mesurer : Après avoir défini la qualité de l'information, celle-ci est mesurée afin de pouvoir la quantifier et l'utiliser.

Analyser : Les mesures de l'étape précédente sont interprétées. Cette étape est responsable de l'identification des dimensions de la qualité déficitaires (ayant un faible niveau de qualité). De plus, sont également identifiées les sources de ces problèmes de qualité.

8. Les dimensions des méthodologies COLDQ et CDQ n'ont pas été illustrées parce qu'elles sont plutôt adaptées à la qualité des données

9. La méthode DMIC : « Define, Measure, Analyse, Improve, Control » soit « Définir, mesurer, analyser, améliorer, contrôler »

TDQM	TIQM	AIMQ	DQA	DaQuinCIS	QAFD
précision	*Dim. inhérents :*	accessibilité	accessibilité	précision	précision sémantique
	précision (source)	justesse	volume d'info.	complétude	précision syntactique
	précision (réalité)	sans erreurs	sans erreurs	obsolescence	consistance interne
confiance	confiance	confiance	confiance	confiance	consistance externe
objectivité	consistance	objectivité	objectivité	consistance	complétude
réputation	complétude	réputation	réputation		fraicheur
plus-value	conf. au business	facile d'utilisation	facile d'utilisation		unicité
pertinence	non-duplication	pertinence	pertinence		
obsolescence	redondance	obsolescence	obsolescence		
complétude	*Dim. pragmatiques :*	complétude	complétude		
volume	accessibilité	sécurité	sécurité		
facile d'interprét.	obsolescence	facile d'interprét.	facile d'interprét.		
compréhensibilité	clarté contextuelle	compréhensibilité	compréhensibilité		
consistance	intégrité	consistance	consistance		
concision	facile d'utilisation	concision	concision		
manipulabilité	exactitude		valeur ajoutée		
accès	coût				
sécurité					

TABLE 3.2: Les dimensions de qualité présentes dans diverses méthodologies

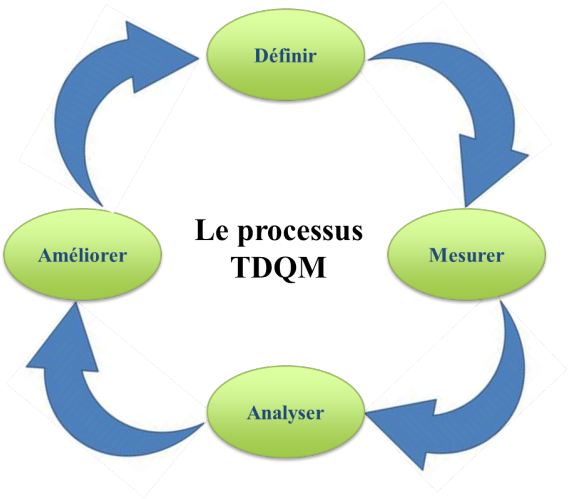


FIGURE 3.3: Le processus TDQM d'amélioration continue de la qualité des données et de l'information

Intrinsèque	Contextuelle	Représentation	Accessibilité
précision	plus-value	interprétabilité	accès
confiance	pertinence	compréhensibilité	sécurité
objectivité	obsolescence	consistance	
réputation	complétude	concision	
	quantité	manipulabilité	

TABLE 3.3: Les attributs de la qualité de données/information selon [Wang 96]

Améliorer : À cette étape sont entreprises les actions nécessaires pour diminuer les problèmes de qualité. D'habitude, l'intervention est faite au niveau des processus dont les problèmes sont issus.

Dans le paragraphe suivant, le modèle de qualité de Wang et Strong sera présenté plus en détail. Il est considéré comme étant le modèle de référence dans le domaine des systèmes d'information d'aide au management des organisations.

3.3.1 Le modèle de Wang et Strong

Dans [Wang 96] un cadre a été proposé pour l'évaluation de la qualité des données et des informations. Malheureusement, dans cette étude les deux notions : données et informations sont considérées comme étant équivalentes. Un point très important de cette étude est que les attributs définissant la qualité (des données et de l'information) ont été définis par les utilisateurs et non pas par les chercheurs. En première conclusion la notion de qualité, c'est-à-dire les dimensions employées pour sa définition, diffère en fonction du contexte d'utilisation : la tâche à réaliser, l'expérience de l'utilisateur, etc. Deuxièmement, la plupart de dimensions de qualité gravite autour de la notion de *précision*. Après une première analyse de la notion de qualité, Wang et Strong sont arrivés à un ensemble de 179 dimensions définissant la qualité des données et de l'information. Comme le nombre de paramètres est trop important, ils ont raffiné leur analyse à l'aide des autres groupes d'analyse et suite à une analyse factorielle, Wang et Strong ont réussi restreindre à 16 le nombre de dimensions. Pour une plus simple illustration, ces dimensions ont été regroupées en quatre catégories, tableau 3.3 :

- la 1ère catégorie correspond aux dimensions **intrinsèques** de la qualité des données et de l'information, c'est-à-dire indépendantes du contexte de l'application ;
- la 2ère catégorie contient les dimensions **contextuelles** de la qualité, c'est-à-dire les dimensions dépendantes du contexte de l'application ;
- la 3ère catégorie contient les dimensions de la qualité décrivant la **représentation** des données et des informations en vue de leur facilité d'utilisation ;
- la 4ère catégorie correspond aux dimensions de la qualité traduisant l'**accessibilité** des données et des informations.

Par la suite les dimensions de chaque catégorie de la qualité seront détaillées.

La qualité intrinsèque

- *La précision* correspond à l'erreur entre la valeur enregistrée et la valeur d'une base de référence (correspondante à la réalité physique). C'est la dimension de qualité la plus usuelle et elle est mesurée très souvent en utilisant une distance entre la valeur de référence et la valeur de facto. La distance euclidienne est parmi les mesures de précision les plus utilisées dans la pratique. Pour une information composée de plusieurs valeurs numérique on peut employer comme mesure de précision :
  - l'erreur quadratique moyenne :

$$EQM = \sqrt{\frac{1}{N}\sum_{1}^{N}(\omega_i - \hat{\omega}_i)^2} \tag{3.1}$$

    Dans cette équation les ω_i représentent les valeurs correctes (de référence) et les $\hat{\omega}_i$ les valeurs réelles (enregistrées) des données ou des informations.
  - le rapport entre le nombre de valeurs correctes sur le nombre total de valeurs.
- *La confiance*, comme le nom l'indique, correspond au degré de crédibilité dans une donnée/information. Cette dimension est d'une importance très grande et elle peut être incluse aussi dans la catégorie de qualité contextuelle car en fonction du contexte d'application certaines données/informations peuvent avoir différents niveaux de crédibilité. La notion de crédibilité peut faire référence à deux entités : la crédibilité de données/informations et la crédibilité des sources qui les ont produit.
- *L'objectivité* : décrit le degré d'objectivité de la source dont les données/informations sont issues. Lorsque des données ou des informations subjectives, c'est-à-dire issues des humains, sont utilisées, des biais peuvent être introduits. Cette dimension peut influencer la dimension confiance.
- *La réputation* : est une caractéristique de la qualité se rapprochant de la notion de confiance, mais qui est déterminée au fur et à mesure des expériences avec certaines sources de données/information. Ainsi, si une source d'information a fourni que des informations crédibles, sa réputation, ainsi construite, fait que les futures informations seront également considérées crédibles.

La qualité contextuelle

- *La plus-value* : représente le bénéfice d'utilisation de ces informations en terme monétaires. Cette dimension doit toujours être prise en compte lorsqu'il y a des coûts associés et quand la nature des informations n'est pas encore certaine ;
- *La pertinence* : correspond à l'utilité des données/informations pour la tâche en cours ;
- *L'obsolescence*[10] : correspond à une dimension de qualité temporelle représentant la nouveauté, la fraîcheur des données/informations. Dans le contexte des systèmes dynamiques

10. Le terme anglais est « timeliness »

c'est une des dimensions de qualité les plus importantes. Cette dimension de qualité peut être encore divisée en deux autres [Peralta 06] : *l'obsolescence de l'information*, par rapport aux sources d'information [11] et *l'obsolescence de données*, comme montré dans la figure 3.4. Pour la dimension d'obsolescence de l'information, dans la littérature les mesures suivantes sont utilisées :

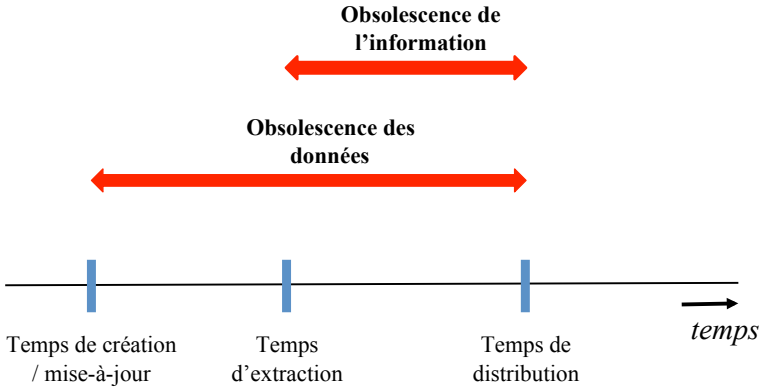


FIGURE 3.4: L'obsolescence des données et de l'information

- la différence entre le moment de l'extraction de l'information et le moment dont elle est reçu par l'utilisateur ;
- le nombre de mises-à-jour d'une source d'information par rapport au moment d'extraction de l'information ;
- le rapport d'obsolescence : le pourcentage d'informations extraites qui sont à jour, c'est-à-dire leurs valeurs sont actuelles (identiques) par rapport à leurs sources.

En ce qui concerne la dimension *d'obsolescence de données*, la mesure la plus utilisée est :

- le temps écoulé par rapport à la dernière mise-à-jour des données (utilisation des estampilles temporelles, « timestamps »).

À partir de ces mesures, d'autres indicateurs de l'obsolescence des données et de l'information peuvent être construits, comme par exemple [Ballou 98] :

$$max\left\{1 - \frac{obsolescence}{volatilité}, 0\right\}^{s} \tag{3.2}$$

où la volatilité exprime la durée d'actualité de l'information et s est un coefficient de contrôle de la sensibilité du rapport. L'avantage de cet indicateur est qu'il est défini dans l'intervalle unitaire le rendant facilement compréhensible et utilisable.

– *La complétude* (ou son équivalent inverse *l'incomplétude*) : mesure la présence de toutes les valeurs pour toutes les variables. Cette dimension peut être étudiée en la divisant en complétude de la structure et complétude du contenu [Ballou 03] :

$$\text{Complétude}_{\text{Struct}} = \frac{\text{valeurs enregistrées}}{\text{valeurs qui pourraient être enregistrées}}$$

$$\text{Complétude}_{\text{Contenu}} = \frac{\text{contenu transmis}}{\text{contenu qui pourraient être transmis}}$$

C'est une dimension très importante dans le cas où plusieurs sources d'information sont utilisées, ces sources pouvant fournir des informations complémentaires et donc de réduire l'incomplétude.

11. Le terme anglais est « currency »

- *La quantité de données/informations* : Un très grand volume de données a un impact important sur le temps de traitements nécessaires pour en extraire des informations. De plus, il n'est pas envisageable de présenter à l'utilisateur un très grand volume d'informations parce qu'il aura des problèmes avec leur compréhension. Les mesures classiques pour mesurer la quantité (le volume) de données sont le nombre de n-uplets et d'attributs d'une base de données, le volume en bits, le nombre de lignes et de colonnes d'un enregistrement d'une image, etc.

La qualité de représentation

- *L'interprétabilité* : les données/informations doivent être représentées en utilisant un langage et des unités de mesure claires pour l'utilisateur (tous les champs des données/informations doivent avoir des explications).
- *La compréhensibilité* : les données/informations doivent être facilement présentées sous une forme claire, non-ambiguë pour être facilement compréhensibles.
- *La consistance* : cette dimension de qualité a été traitée lors du paragraphe 2.1. Pour que la consistance soit préservée il est nécessaire d'utiliser le même format de représentation des données et des informations d'un système à un autre ou d'une application à une autre [Fisher 13]. Des problèmes de consistance peuvent apparaître lors de la combinaison de plusieurs données/informations (ex. la combinaison de deux tables utilisant un codage différent pour les valeurs des attributs). Pour augmenter le niveau de consistance on peut filtrer des données ou des informations [Ballou 03]. Ce filtrage doit se faire au niveau de la structure/du format de l'information et non pas au niveau du contenu de l'information.
- *La concision* : les données/informations doivent être présentées sous un format de petite taille, mais sans perdre en complétude.
- *La manipulabilité* : représente la facilité de traitement (modifier, mettre à jour, supprimer, copier, etc.) des données/informations.

La qualité d'accessibilité

- *L'accessibilité* : les données et les informations doivent être facilement et rapidement accessibles.
- *La sécurité* : les données et les informations doivent être protégées en limitant l'accès. Donc cette dimension est opposée à la précédente, l'accessibilité.

3.3.2 Les relations entre les dimensions de la qualité

Après avoir analysé individuellement les dimensions de qualité proposées par Wang et Strong, il est nécessaire de regarder s'il existe des influences entre les différentes dimensions. Déjà, lors de la définition de la dimension de qualité *l'objectivité*, il a été précisé qu'elle peut avoir un impact sur une autre dimension de qualité, *la confiance*. Un autre exemple d'influence, plus direct et plus évident, est la relation inverse entre *l'accessibilité* et *la sécurité* : une augmentation du niveau de sécurité va diminuer l'accessibilité et vice-versa. Dans [Ballou 03], il a été montré qu'il existe, dans beaucoup de cas, une relation entre *la consistance* et *la complétude*. Ainsi, très souvent il faut choisir entre ces deux variantes : soit d'avoir des informations complètes mais inconsistantes, soit d'avoir des informations incomplètes mais consistantes. Une autre dépendance qui peut être observée est entre *la précision* et *l'obsolescence*, car plus on veut avoir des informations précises plus on a besoin de temps pour collecter des données et les traiter [Ballou 95]. Ainsi il faut prendre des précautions quand on veut se concentrer sur une seule dimension de la qualité parce qu'à cause des relations existantes, d'autres dimensions peuvent influencer cette dernière.

Le modèle de [Zeng 04]	Le modèle de [Ran 03]	Le modèle UML de [OMG 08]
débit	débit	débit
temps de réponse	temps de réponse	latence
coût	coût	capacité
accessibilité	accessibilité	accessibilité
fiabilité	fiabilité	fiabilité
réputation	sécurité	sécurité
		confidentialité
		chargement
		probabilité d'erreur

TABLE 3.4: Les dimensions de la QoS

3.4 LA QUALITÉ DE L'INFORMATION DANS LES WIS

Le développement de l'Internet, plus particulièrement du World Wide Web (www), a ouvert la possibilité d'accéder à une multitude de sources d'information sur presque tous les sujets d'intérêt [Naumann 01]. Actuellement, presque tous les documents sont numérisés et mis sur un serveur afin d'être disponibles à distance grâce aux services Web.

Les systèmes les plus utilisés dans le vaste domaine du Web sont les moteurs de recherche. Ils nous aident à retrouver des sources d'information dont certaines caractéristiques (des « mots-clé ») sont connues ou même à découvrir des nouvelles sources d'information. Très souvent, grâce à l'existence de plusieurs alternatives, l'utilisateur choisit une source particulière ou combine plusieurs sources afin d'obtenir l'information dont il a besoin. Bien sûr, c'est une situation bien avantageuse mais qui demande du temps. Chaque site Web peut être considéré comme une source d'information. Le nombre de sites Web a continué augmenter au fur et à mesure du temps, voir figure 3.5, et actuellement ils viennent de dépasser 1 milliard.

Les systèmes d'information qui utilisent les technologies du Web ont commencé à être de plus en plus présents dans notre vie quotidienne. Ils ont un rôle primordial dans nos activités de communication, d'information ou de divertissement. Par rapport aux autres types de systèmes d'information, les WIS utilisent des technologies qui sont complètement indépendantes de l'utilisateur (du client) : le réseau, les serveurs, les sites Web, etc. Ainsi, l'utilisateur n'a aucune possibilité de contrôler le système. En conséquence, l'évaluation de la qualité des informations fournies par les WIS est très importante parce que les informations d'une faible qualité peuvent avoir des influences indésirables.

Les performances d'un tel système sont traditionnellement évaluées par la qualité de service (le QoS [12]). Les problèmes généraux couverts par la QoS sont liés à la transmission et aux taux d'erreur des sources Web. Dans le tableau 3.4, trois modèles proposant les dimensions couvertes par la QoS sont présentés. En les analysant, il peut s'observer que ces dimensions caractérisent le bon fonctionnement d'un système par : le débit maximal sur le réseau, le temps de réponse d'un service Web, la fiabilité d'un serveur, etc. Même si les fournisseurs d'Internet garantissent par contrat une certaine QoS pour chaque client, ce type de mesure de la qualité n'est que partiellement satisfaisant pour l'utilisateur. En plus de la QoS, l'utilisateur a besoin de connaître la qualité des informations qui lui sont fournies par le WIS. En conséquence, d'autres dimensions sont nécessaires afin de pouvoir caractériser l'utilité et l'adaptation des informations pour l'utilisateur.

[Naumann 02] a essayé de répondre à cette question en proposant vingt-deux dimensions de qualité compilées à partir de six méthodologies différentes. Comme dans le cas de Wang et Strong, il a regroupé les dimensions en quatre catégories, cf. tableau 3.5.

12. Acronyme du terme anglais « Quality of Service »

Année (juin)	Sites Web	Changement	Utilisateurs Internet	Utilisateur per site Web	Sites Web lancés
2013	**672,985,183**	-3%	2,756,198,420	4	
2012	**697,089,489**	101%	2,518,453,530	4	
2011	**346,004,403**	67%	2,282,955,130	7	
2010	**206,956,723**	-13%	2,045,865,660	10	Pinterest
2009	**238,027,855**	38%	1,766,206,240	7	
2008	**172,338,726**	41%	1,571,601,630	9	Dropbox
2007	**121,892,559**	43%	1,373,327,790	11	Tumblr
2006	**85,507,314**	32%	1,160,335,280	14	Twttr
2005	**64,780,617**	26%	1,027,580,990	16	YouTube, Reddit
2004	**51,611,646**	26%	910,060,180	18	Thefacebook, Flickr
2003	**40,912,332**	6%	778,555,680	19	WordPress, Linkedin
2002	**38,760,373**	32%	662,663,600	17	
2001	**29,254,370**	71%	500,609,240	17	Wikipedia
2000	**17,087,182**	438%	413,425,190	24	
1999	**3,177,453**	32%	280,866,670	88	Baidu
1998	**2,410,067**	116%	188,023,930	78	PayPal
1997	**1,117,255**	334%	120,758,310	108	Google
1996	**257,601**	996%	77,433,860	301	Yandex
1995	**23,500**	758%	44,838,900	1,908	
1994	**2,738**	2006%	25,454,590	9,297	Altavista, Amazon, AuctionWeb
1993	**130**	1200%	14,161,570	108,935	Yahoo
1992	**10**	900%			
Août 1991	**1**				World Wide Web Projet

FIGURE 3.5: L'évolution du nombre de sites Web et d'utilisateurs, selon [NetCraft 14]

Contenu	Technique	Intellectuel	Présentation
précision	accessibilité	confiance	volume
complétude	latence	objectivité	concision
support client	coût	réputation	consistance
documentation	QoS		compréhensibilité
interprétabilité	temps de réponse		vérifiabilité
pertinence	sécurité		
valeur ajoutée	fraicheur		

TABLE 3.5: La qualité de l'information dans le cas des systèmes d'information utilisant les services Web, selon [Naumann 02]

La qualité du contenu

La qualité du contenu décrit les aspects intrinsèques de l'information. La précision, la complétude, la pertinence, l'interprétabilité et la valeur ajoutée sont des dimensions identiques à celles du modèle de Wang et Strong. La seule différence est leur regroupement dans la même classe. Cependant, il existe des dimensions particulières :

- *Le support client* : ensemble avec l'autre dimension, *la documentation* décrivent la possibilité d'offrir le support nécessaire pour aider l'utilisateur dans la compréhension et dans l'utilisation des informations fournies. Le support client se fait habituellement par courriel ou par téléphone et il peut être quantifié par le temps moyen d'attente pour la réponse. Une autre quantification de cette dimension peut se faire en évaluant l'utilité de la réponse (c'est une dimension de qualité d'ordre supérieur) ;
- *La documentation* : consiste dans l'existence de méta-données qui renvoient l'utilisateur vers une page Web contenant des explications sur l'information fournie. Cette dimension peut être quantifiée par une variable booléenne (1 - documentation existante, 0 - documentation inexistante), par le nombre de mots / lignes / pages de la documentation ou par une évaluation subjective des utilisateur vis-à-vis de son utilité et de son compréhensibilité (dimension couverte par une autre dimension).

La qualité technique

Les dimensions techniques de la qualité caractérisent le software, le hardware et le réseau par rapport à la satisfaction de l'utilisateur.

- *L'accessibilité* : est la probabilité qu'une source d'information réponde à une requête dans un intervalle de temps prédéfini. Dans le cas des WIS, cette dimension concerne le bon fonctionnement du software, du hardware et du réseau entre l'utilisateur, le système d'information et les sources d'information. L'accessibilité est une des dimensions les plus critiques d'un WIS et une des plus difficile à paramétrer parce qu'elle est fortement dépendante du trafic dans le réseau, de la distribution de serveurs dans le monde, des attaques informatiques, des opérations de maintenance, etc. ;
- *La latence* : est le temps écoulé entre la requête de l'utilisateur et la réception du premier élément d'information. Dans le cas où un seul élément d'information est demandé, la latence est équivalente au *temps de réponse* ;
- *Le coût* : est le montant d'argent que l'utilisateur doit payer pour obtenir l'information ;
- *Le QoS* : il a été présenté précédemment ;
- *La sécurité* : décrit le degré de confidentialité de la requête envoyée par l'utilisateur et de l'information fournie par le WIS ;
- *La fraîcheur* : identique avec celle du modèle de Wang et Strong.

La qualité intellectuelle

La qualité intellectuelle concerne les aspects subjectifs des utilisateurs vis-à-vis de la qualité des sources d'information. Toutes ces dimensions se retrouvent également dans le modèle de Wang et Strong. La seule différence majeure est que dans ce cas, la qualité de l'information est évaluée par rapport aux sources d'information. Si dans le cas d'un système d'information d'une organisation, le nombre de sources d'information est restreint, dans le cas des WIS l'utilisateur dispose d'un nombre impressionnant des sources.

L'évaluation des sources d'information se fait habituellement en fonction de leur affiliation et de la possibilité de la vérifier (une autre dimension de qualité de ce modèle). Ainsi, les utilisateurs ont tendance à préférer les sources internes dans leur organisation ou les sources qui sont généralement connues comme étant crédibles : CNN pour les actualités, Yahoo Finance pour la bourse, IEEE et ScienceDirect pour des articles scientifiques, etc.

La qualité de présentation

La qualité de présentation caractérise la façon de visualiser les informations pour qu'elles soient facilement utilisables par l'utilisateur. À part la *vérifiabilité*, toutes les autres dimensions ont été décrites par le modèle de Wang et Strong.

– *La vérifiabilité* : décrit le degré avec lequel l'information présentée peut être vérifiée afin d'établir sa conformité. Le niveau de vérifiabilité peut être amélioré en présentant la source, c'est-à-dire la traçabilité de l'information ou en indiquant une source tierce (de préférence ayant un niveau de crédibilité élevé) la confirmant.

3.4.1 Conclusion

Les seules vraies différences entre le modèle de Naumman et celui de Wang et Strong se situent au niveau des dimensions techniques de la qualité. Même si la liste de dimensions de qualité est plus grande que celle des MIS, Naumann recommande d'utiliser dans la pratique un nombre réduit de dimensions. L'utilisation d'un nombre élevé de dimensions a pour conséquence une augmentation de la difficulté d'évaluation d'une source par l'utilisateur. Néanmoins, elle permet de couvrir toutes les aspects de l'information et donc, d'avoir une évaluation rigoureuse.

Comme pour le modèle de Wang et Strong, il existe de très fortes corrélations entre les diverses dimensions de qualité. Ainsi, certaines dimensions peuvent être dérivées par la combinaison de plusieurs dimensions. Par exemple, du point de vue de l'utilisateur l'accessibilité d'une source S_i pourrait s'exprimer comme étant :

$$Accessibilité(S_i) = Latence(S_i) \times Coût(S_i) \tag{3.3}$$

Un autre exemple est donné dans [Singh 13] pour la réputation d'une source. Dans leur vision elle est déterminée par l'accessibilité et la fiabilité de la source, donc :

$$Réputation(S_i) = Accessibilité(S_i) \times Fiabilité(S_i) \tag{3.4}$$

3.5 La qualité de l'information dans les IFS

Comme la plupart des chercheurs considèrent les notions de données et d'information comme équivalentes, il existe plusieurs termes qui sont utilisés avec le même sens : fusion d'informations, fusion de décisions, fusion de données, fusion de capteurs, etc. Dans ce mémoire de thèse est choisi le terme fusion d'informations [13] pour désigner l'ensemble de fusions possibles. Néanmoins, en fonction de la sémantique des entités à fusionner, différents types de fusion seront identifiées : de données, d'informations, de connaissances. Cette différentiation est très importante dans notre cas car l'objectif est la modélisation de la qualité, qui est directement dépendante de la sémantique, cet aspect sera détaillé dans le paragraphe 4.2.

Traditionnellement, un système de fusion d'informations utilise plusieurs sources de données : différents types de capteurs, experts humains, etc. Ainsi, très souvent dans la littérature ces systèmes sont appelés systèmes multi-sources ou encore systèmes multi-capteurs.

[Bossé 06] définit la fusion d'information comme étant un processus d'acquisition, de filtrage, de corrélation et d'intégration d'informations pertinentes, issues de différentes sources hétérogènes (différents types de capteurs, bases de données, experts humains, etc.) dans un format de représentation unique et adapté à la prise de décisions.

La fusion d'information est représentée par un module de traitement de l'information. Deux cas d'utilisation peuvent être identifiés :

13. Dans la communauté scientifique ce terme est presque exclusivement utilisé : la conférence principale s'appelle « International Conference on Information Fusion » et les deux principaux journaux : « Information Fusion » (Elsevier) et « Journal of Advances in Information Fusion » (ISIF).

- Le premier lorsque le module de fusion est intégré dans le système dès le début (par exemple pour la fusion de données distribuées dans plusieurs bases de données) ;
- Le deuxième lorsque le module de fusion est intégré afin d'améliorer les performances d'un système (en général en exploitant la redondance et la complémentarité des sources d'information).

Dans ces deux cas, la fusion d'information n'est qu'un module faisant partie du système de traitement de l'information. Ces deux cas d'utilisation jouent un rôle important dans l'évaluation des performances d'un tel module. Ainsi, pour le deuxième cas, l'évaluation des performances du module de fusion va devoir se faire par rapport à la situation où cette fusion n'est pas appliquée.

Pour l'analyse de situations complexes, il est nécessaire d'utiliser plusieurs capteurs afin de pouvoir tirer avantage de leurs comportements complémentaires. Selon [Appriou 01] les principaux avantages d'utilisation d'un système multi-capteurs sont :

- L'augmentation du nombre de situations pouvant être analysées : quand un des capteurs devient inefficace à cause des conditions externes (contre-mesures, conditions atmosphériques, etc.) ou internes (défections techniques), les autres capteurs vont continuer à prendre des mesures ;
- La réduction du temps nécessaire pour la prise de mesures grâce à la coopération des capteurs et au partage des fonctions ;
- L'augmentation de la capacité de discrimination grâce aux observations complémentaires qui sont localement partielles.

Selon d'autres auteurs [Blasch 07], les avantages d'utilisation de la fusion d'information sont :

- La réduction de l'incertitude ;
- La réduction de la dimension du problème (plusieurs valeurs sont agrégées en une seule) ;
- L'amélioration de la réactivité du système (le temps de transmission d'une seule information est plus rapide que la transmission de toutes les informations).

Un système de fusion d'information est par sa nature un système complexe qui nécessite, pour sa construction et pour son utilisation, des réponses à des questions comme [Waltz 90] :

- Quelle est la meilleure combinaison de capteurs et de sources de données pour que les besoins en terme d'information des utilisateurs finaux soient respectés ?
- Avec un ensemble particulier de capteurs et de sources de données quelles sont les performances qui peuvent être atteintes ? Comment les performances vont s'améliorer si d'autres capteurs et/ou sources de données sont ajoutés ? Comment les performances vont s'améliorer suite à une amélioration de la qualité (des performances) des capteurs ou des sources de données ?

Une notion très importante dans les systèmes avec fusion d'information est celle d'*incertitude*. Très souvent l'implémentation d'un module de fusion d'information a pour objectif la diminution de l'incertitude [Blasch 07]. Ainsi, dans le paragraphe 3.5.1, une courte introduction dans la modélisation de l'incertitude sera présentée. Ensuite, dans le paragraphe 3.5.2, la notion d'incertitude sera mise en correspondance avec celle de la qualité de l'information.

3.5.1 Modélisation des incertitudes

Les incertitudes peuvent être associées à l'acquisition des données, aux traitements de données, aux informations extraites et à la fiabilité du système [Atoyan 10]. Ces sources d'incertitude peuvent être caractérisés par un comportement objectif parce qu'elles ne sont pas dépendantes de l'utilisateur. En même temps, il existe des incertitudes liées à l'utilisateur, des incertitudes subjectives. Ces incertitudes dépendent de la perception des informations par les utilisateurs et par la méthode de traitement de ces informations. Ces différences de perceptions et d'intégration des informations influencent les décisions prises et les actions des utilisateurs. Parmi les causes de ces différences, [Atoyan 10] identifie les attentes des utilisateurs, leurs modèles mentaux et leurs confiances dans le système.

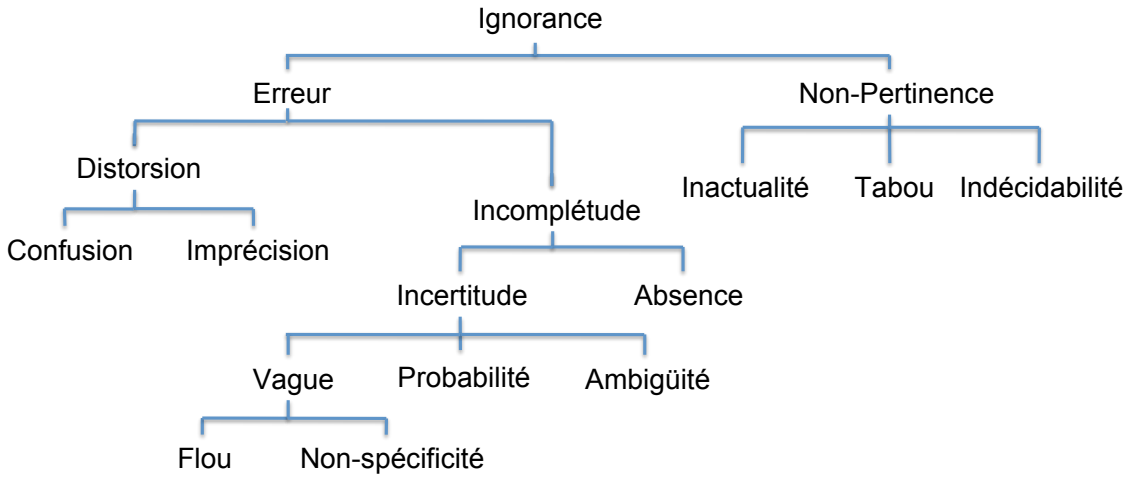


FIGURE 3.6: La taxonomie d'incertitude de Smithson

Les incertitudes sont directement dépendantes du contexte d'application. Dans la littérature, il existe plusieurs modélisations de l'incertitude sous la forme de taxonomies. Ces taxonomies, développées en grande majorité à la fin des années 80 et au début des années 90, décomposent la notion d'incertitude en plusieurs dimensions pour une meilleure compréhension. Trois taxonomies parmi les plus connues seront présentées par la suite. Une observation intéressante concernant tous ces modèles est qu'ils utilisent comme point de départ la notion d'*ignorance*.

3.5.1.1 La taxonomie de Smithson

La taxonomie proposée par [Smithson 89] s'appuie sur la notion d'ignorance ou celle de la non-connaissance. Smithson fait la distinction entre deux états différents : *état d'ignorance* et *état d'ignorer*. Quand une personne est ignorante à cause d'un manque de connaissance, elle est dans un état *d'erreur* par rapport à l'état de connaissance parfaite. Tandis que, dans le cas où une décision consciente est prise en déconsidérant quelque chose sous le motif de non-pertinence, cette information est ignorée. Ainsi, la notion d'ignorance est décomposée en deux grandes catégories : *l'erreur* et *la non-pertinence* (comme présenté dans la figure 3.6). Smithson divise la non-pertinence en 3 catégories : l'inactualité, l'indécidabilité et le tabou. L'inactualité représente les choses qui n'ont plus de connexion avec la situation présente. L'indécidabilité est issue de problèmes qui n'ont pas de solution. Le tabou représente les situations dans lesquelles les utilisateurs n'ont pas le droit de poser de questions et ils doivent accepter ce qu'ils ont reçu.

La grande majorité des études menées dans la modélisation de l'incertitude ont pris en compte l'erreur. Smithson identifie deux causes d'erreurs : *l'incomplétude* et *la distorsion* des données ou des informations. La distorsion des données/informations a pour cause principale l'imprécision, mais de possibles confusions peuvent également jouer un rôle important. L'incomplétude selon Smithson est due soit à l'absence de l'information, soit à l'incertitude de l'information. Afin de définir l'incertitude, Smithson propose de la subdiviser en trois catégories : la probabilité (l'incertitude des événements qui ne sont pas totalement prédictives) ; l'ambiguïté (l'incertitude provenant de l'incapacité à distinguer deux événements) et le flou (l'incertitude provenant des distinctions incorrectement observées).

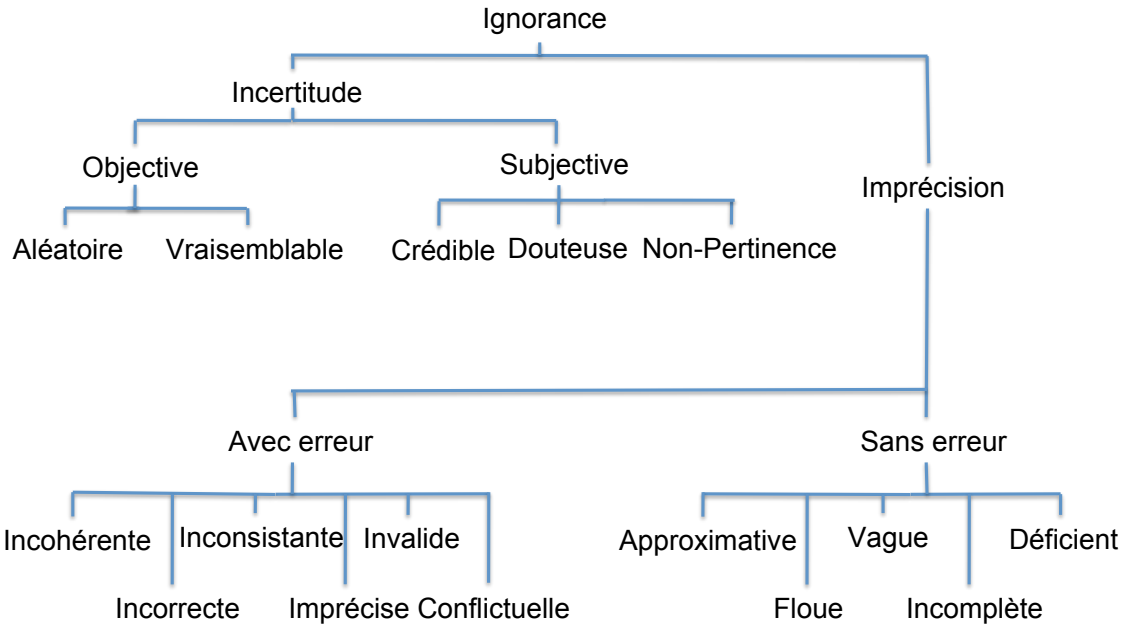


FIGURE 3.7: La taxonomie d'incertitude de Smets

3.5.1.2 La taxonomie de Smets

[Smets 96] a proposé une nouvelle taxonomie traitant le problème de l'incertitude. Smets fait la distinction entre *imprécision* et *incertitude*, voir la figure 3.7. Dans la vision de Smets, l'imprécision est une caractéristique du contenu de l'information, tandis que l'incertitude caractérise un manque d'information pour pouvoir décider si une affirmation est vraie ou fausse. Dans son étude Smets donne l'exemple de deux affirmations :

- *John a au moins deux enfants et j'en suis sûr.*
- *John a trois enfants, mais je n'en suis pas sûr.*

Dans le premier cas, il s'agit d'une information imprécise mais certaine, tandis que dans le deuxième cas, l'information est précise mais incertaine. De cet exemple, on peut déduire que les deux concepts peuvent coexister, mais qu'ils ont des sens différents. Ainsi, il est possible d'avoir une information moins précise mais avec un niveau élevé de certitude ou d'avoir une information plus précise mais d'un niveau plus faible de certitude. Smets va plus loin dans son raisonnement et affirme qu'il est possible d'énoncer un *principe de maximum d'information* qui dit que « le produit » entre le niveau de précision et le niveau de certitude ne peut pas dépasser un certain niveau critique, comme dans le cas du principe d'incertitude de Heisenberg-Gabor utilisé dans le domaine du traitement du signal.

L'incertitude peut être soit une propriété objective de l'information, soit une propriété subjective de l'observateur. L'incertitude objective peut être soit aléatoire, soit vraisemblable, si elle est connue statistiquement. L'incertitude subjective est dépendante de l'utilisateur : informations crédibles mais pas en totalité, informations de provenance douteuse ou informations non-pertinentes et dans ce dernier cas, elles ne sont pas utiles.

Smets divise l'imprécision en deux catégories caractérisant : les informations sans erreurs et les informations avec erreurs. Les informations sans erreurs peuvent être *vagues* : pas bien définies et étant ambigües ; *approximatives* : bien définies et connues d'être proches de la vraie valeur ; *floues* : pas bien définies ; *incomplètes* : des valeurs manquantes ; *déficientes* : quand le manque de valeurs est important. Les informations avec erreurs peuvent être *incorrectes* : complètement erronées ;

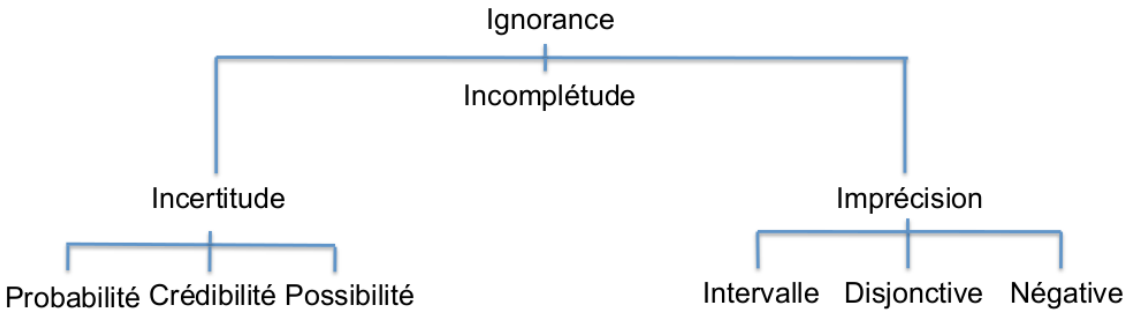


FIGURE 3.8: La taxonomie d'incertitude de Bonissone et Tong

imprécises : la vraie valeur étant comprise dans un intervalle ; *invalides* : générant de fausses informations ; certaines informations peuvent être *conflictuelles*, *incohérentes* ou *inconsistantes* entres elles.

3.5.1.3 La taxonomie de Bonissone et Tong

Dans [Bonnisone 85], Bonissone et Tong ont proposé, dans le cadre d'un système expert, un modèle avec trois types d'ignorances : l'incertitude, l'incomplétude et l'imprécision, voir la figure 3.8. Dans leur vision l'incomplétude représente l'absence d'une valeur et l'imprécision caractérise une valeur qui n'est pas suffisamment précise. L'incertitude apparait lorsqu'un agent doit construire une opinion subjective sur une hypothèse, sans avoir une connaissance certaine.

L'incomplétude, selon Bonissone et Tong, est définie lorsque l'inférence est réalisée en utilisant des informations partielles. Les éléments d'information sont ensuite regroupées en deux catégories : nécessaires et possibles.

L'imprécision peut être représentée soit par un intervalle (ou par une valeur floue), soit par une disjonction (X est x_1 ou x_2), soit encore par une négation (X n'est pas x_1). Malheureusement, Bonissone et Tong ne proposent pas de représentation mathématique particulière pour l'imprécision. Par contre, ils proposent de représenter l'incertitude soit par une probabilité (théorie des probabilités), soit par un degré de crédibilité (théorie de Dempster-Shafer), soit par une possibilité (théorie des possibilités), voir l'annexe A pour une description de ces trois théories mathématiques.

3.5.1.4 Les taxonomies de l'incertitude et la notion de qualité

Les taxonomies de Smithson et de Smets donnent une définition parmi les plus complètes des imperfections de données et des informations. Les deux utilisent quasi les mêmes termes pour caractériser l'incertitude. Ainsi, il peut s'affirmer que ces deux taxonomies sont en concordance, avec quelques différences impliquant que certaines notions ont des significations différentes [Parsons 01]. Ainsi, selon Smithson la notion d'incertitude est une caractéristique de la notion d'erreur, parce que les incertitudes provoquent des erreurs dans les processus de raisonnent. Par contre, pour Smets l'incertitude est une notion d'un niveau plus élevé caractérisant directement l'information. De plus, selon Smets les informations incertaines ne sont pas forcement le résultat d'utilisation de données erronées, elles peuvent être issues par exemple de données incomplètes.

Comme les attributs de ces taxonomies caractérisent des aspects spécifiques des données et des informations, des concepts mathématiques particuliers ont été suggérés pour la quantification de diverses imperfections. Ainsi, les aspects aléatoires sont mieux représentés par la théories des probabilités, les aspects flous et ambigus par la théorie des ensembles flous. Néanmoins, il existe des aspects qui peuvent être quantifiés dans plusieurs théories mathématiques. Un exemple dans ce sens est le conflit entre deux sources d'information (identifié dans la taxonomie de Smets). Celui-ci

peut être représenté dans la théorie des probabilités (par exemple en utilisant la mesure d'entropie de Shannon A.48) et aussi dans la théorie de Dempster-Shafer (par exemple en utilisant la mesure de dissonance A.53).

En faisant une comparaison entre ces taxonomies et la définition de la qualité donnée dans les domaines de MIS et de WIS, plusieurs points en commun peuvent être identifiés :

1. Les taxonomies de l'incertitude et la qualité essaient de représenter la même chose : les caractéristiques des données et de l'information.
2. La qualité et l'incertitude sont définies en utilisant une liste (voulue exhaustive) d'attributs, regroupée dans plusieurs catégories. Ces catégories correspondent aux caractéristiques primordiales des données ou de l'information.
3. Les attributs utilisés sont définis en faisant appel au langage naturel, sans (formellement) définir a priori les entités étudiées, c'est-à-dire les données et les informations.

   À cause du point précédent, les définitions restent très générales et elles ne sont pas accompagnées par des indications sur comment elles pourraient être adaptées dans la pratique.

Néanmoins, il existe aussi des différences :

1. La définition de qualité donnée par Wang et Strong, considère l'information en sortie d'un système d'information et l'analyse par rapport à l'utilisateur final à qui elle est destinée. Cependant, les taxonomies d'incertitude considèrent l'information dans un cadre générale, indépendamment du contexte d'utilisation. Ainsi, ces dernières rapportent les caractéristiques de l'information à la connaissance parfaite.

3.5.2 L'évaluation de la qualité d'un module de fusion

La grande majorité des systèmes d'information « multi-sources » est développée d'une manière optimiste en supposant que les sources sont indépendantes et que les informations issues de ces sources sont de très bonne qualité. Malheureusement, dans la réalité, ces deux hypothèses ne sont pas toujours vérifiées et les performances d'un processus de fusion d'informations sont directement dépendantes de la qualité des informations extraites des données (de la part des multiples sources d'information) et traitées par le module de fusion [Rogova 04].

Afin de pouvoir évaluer un IFS, il faut prendre en compte non seulement ses performances intrinsèques mais aussi la satisfaction des objectifs pour lesquels le système a été développé (en satisfaisant les contraintes de l'environnement réel). Dans les paragraphes suivants, trois modèles d'évaluation de la qualité de l'information en sortie d'un module de fusion sont présentés.

3.5.2.1 Le modèle de Lefebvre, Hadzagic et Bossé

Selon [Lefebvre 07], les éléments principaux d'un système de fusion d'informations sont **les sources d'information**, **le module de fusion** et **le résultat de la fusion**. Pour chacun de ces trois éléments, des mesures de qualité peuvent être associées, voir la figure 3.9. La bonne détection des événements par les capteurs (les sources d'information) est influencée par des *incertitudes*. Pour caractériser la qualité des informations fournies par les sources d'information, [Lefebvre 07] a utilisé la mesure de *confiance*. En fonction de cette mesure le module de fusion peut être construit. En ce qui concerne la qualité de l'information résultante du processus de fusion les mesures de *complétude* et de *pertinence* sont proposées. Par la suite, ces quatre dimensions de la qualité de l'information sont présentées selon [Lefebvre 07] :

1. **L'incertitude** : Cette notion a été présentée au paragraphe précédent, 3.5.1. Dans la littérature, il existe plusieurs théories mathématiques capables d'exprimer différentes aspects de l'incertitude : la théorie des probabilités, la théorie de Dempster-Shafer, la théorie des possibilités, la théorie de l'information généralisée. Toutes ces théories sont présentées dans l'annexe

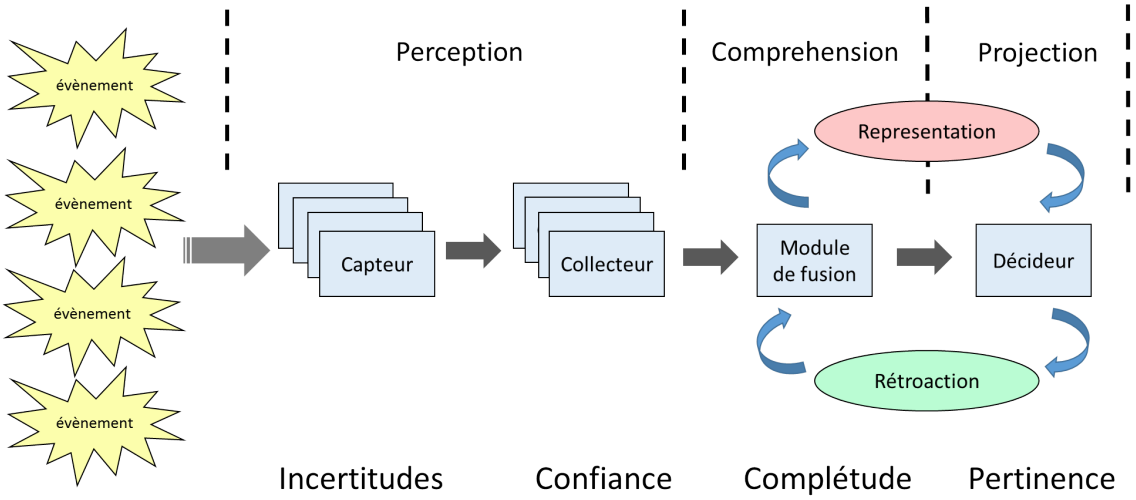


FIGURE 3.9: Modèle d'un système opérationnel de fusion avec les différentes caractéristiques de l'information [Lefebvre 07]

A. En plus, dans la figure 3.12, chaque aspect de l'incertitude est mis en correspondance avec la ou les théories mathématiques adaptées pour son traitement.

2. **La confiance** : caractérise la *précision* avec laquelle les sources d'information peuvent représenter la réalité. En fonction du contexte d'application, les sources d'information peuvent avoir des degrés de confiance différents. En conséquence, il est très important d'évaluer la confiance de chaque source pour avoir une meilleure représentation de la réalité. Comme les sources d'information peuvent être considérées comme étant incertaines, le niveau de confiance de chaque source est directement lié à la modélisation de l'incertitude. En gardant cette approche, [Rogova 10] présente la notion de confiance comme étant une incertitude d'ordre supérieur, c'est-à-dire représentant l'incertitude de l'évaluation de l'incertitude. Ainsi la confiance est dépendante, en plus du contexte d'application et des caractéristiques intrinsèques des sources, de la façon de représenter l'incertitude. L'évaluation des coefficients de confiance de chaque source peut se faire soit par un expert, soit par un apprentissage automatique [Kim 02]. En fonction du niveau de connaissance disponible [14], l'évaluation de coefficients de confiance des sources d'information peut se faire sous la forme [Lefebvre 07] :
   - D'un niveau de confiance numérique qui peut être désigné pour chaque source. Dans ce cas, le plus souvent le niveau de confiance s'exprime par un nombre dans l'intervalle unitaire $[0, 1]$, avec ou sans condition de normalisation du type $\sum_i Conf_i = 1$.
   - De sources d'information qui peuvent être ordonnées en fonction de leur niveau de confiance, mais sans connaître la valeur exacte du niveau de confiance.
   - D'un sous-ensemble de sources d'information qui ont un très bon niveau de confiance, mais sans savoir quelles sont les sources exactes de ce sous-ensemble.

   En conclusion, le module de fusion peut être décrit par une fonction F dépendante des informations reçues à son entrée I_i, $1 \leq i \leq N$ et des niveaux de confiance r_i, $1 \leq i \leq N$ de chaque sources d'information : $F(I_1, I_2, \ldots, I_N, r_1, r_2, \ldots, r_N)$.
3. **La complétude** : est une caractéristique de l'information dépendante directement du module de fusion. Dans la littérature, dans la plupart de cas, elle est décrite en utilisant son

14. Les valeurs de coefficient de confiance peuvent être vues comme des informations utiles d'un point de vue du module de fusion. Ainsi, comme présenté dans le paragraphe 1.3.2, l'évaluation de coefficients de confiance doit se faire avec l'apport d'une base de connaissances

correspondant antagonique, l'*incomplétude*, qui exprime une déficience d'information. La référence la plus citée pour décrire cette dimension de la qualité de l'information est le livre édité en 1997 par Motro et Smets [Motro 96]. La modélisation de l'incomplétude reste un problème difficile même à l'heure actuelle. Dans beaucoup de cas, l'utilisateur peut prendre des décisions même s'il ne dispose pas d'une information complète. Une définition plus récente, dans le domaine de la prise de conscience de la situation, est donnée par [Perry 04]. Selon eux, la complétude est une caractéristique de l'information décrivant le degré avec lequel tous les aspects d'intérêt de l'entité de l'étude sont exprimés. Ainsi, comme observation générale, la complétude est dépendante du contexte.

4. **La pertinence** : est une caractéristique fortement subjective qui peut encore se décomposer en deux autres propriétés : *la valeur ajoutée* et *l'obsolescence* de l'information [Lefebvre 07]. Pour l'évaluer, il faut connaître les besoins et les attentes de l'utilisateur. Ainsi cette dimension de la qualité ne peut pas s'exprimer sous une forme absolue, seulement par rapport au contexte : la situation, le choix des composantes du système, l'expérience de l'utilisateur, etc. Il est possible d'être dans une situation où la quantité d'informations disponibles est très grande et où elle continue à augmenter. Une quantité très grande d'informations est difficile à traiter par le système et à intégrer par l'utilisateur. Dans [Perry 04], la pertinence a été définie comme étant la proportion d'informations collectées qui sont reliées aux besoins de l'utilisateur. Un système d'information travaillant dans un environnement dynamique peut présenter deux types d'information non-pertinentes (selon [Lefebvre 07]) :

   (a) les croyances mutuellement indépendantes et conditionnellement indépendantes (définies dans la théorie de Dempster-Shafer) : considérées comme informations indépendantes et traitées comme telles ;

   (b) les informations obsolètes : le degré de pertinence de ces informations se dégradant en temps (modélisé dans un cadre probabiliste).

   Une autre modalité pour exprimer la pertinence de l'information est l'utilisation des mesures d'entropie [15] : si la quantité d'incertitude de l'information est grande, l'information est non-pertinente. La quantité d'incertitude qui reste après l'observation de y peut s'exprimer par l'entropie conditionnelle : $S(x|y) = S(x, y) - S(y)$. Ce problème a été beaucoup étudié dans le domaine de la recherche d'information.

Le modèle d'évaluation de la qualité de [Lefebvre 07] est intéressant parce qu'il prend en compte les différences sémantiques des informations tout au long de la chaîne de traitement. Dans la figure 3.9 est présenté ce modèle de système opérationnel de fusion avec les différentes caractéristiques de l'information : l'incertitude, la confiance, la complétude et la pertinence.

3.5.2.2 Le modèle de Rogova et Bossé

[Rogova 10] a proposé une première démarche pour définir une ontologie de la qualité de l'information dans le cadre d'un processus de fusion d'information. Dans cette étude, trois catégories de qualité ont été définies : la qualité de la source d'information, la qualité du contenu de l'information et la qualité de la présentation de l'information. Ensuite, pour chaque catégories [Rogova 10] a proposé plusieurs dimensions. L'ensemble de la modélisation de la qualité selon Rogova et Bossé est présenté dans la figure 3.10. Ci-après, chaque catégorie de qualité est présentée en détail.

1. **La qualité des sources d'information** : il existe deux types de sources d'information : des sources objectives et des sources subjectives. Les sources objectives correspondent aux capteurs, modèles, bases de données, processus automatiques, services d'intelligence [16], etc.

15. voir l'annexe A.8.4 pour une présentation des différents types d'entropies
16. Même si les services d'intelligence sont formés par des humains, ils sont considérés comme objectifs parce qu'a priori les informations fournies par eux sont non-biaisées par le facteur humain

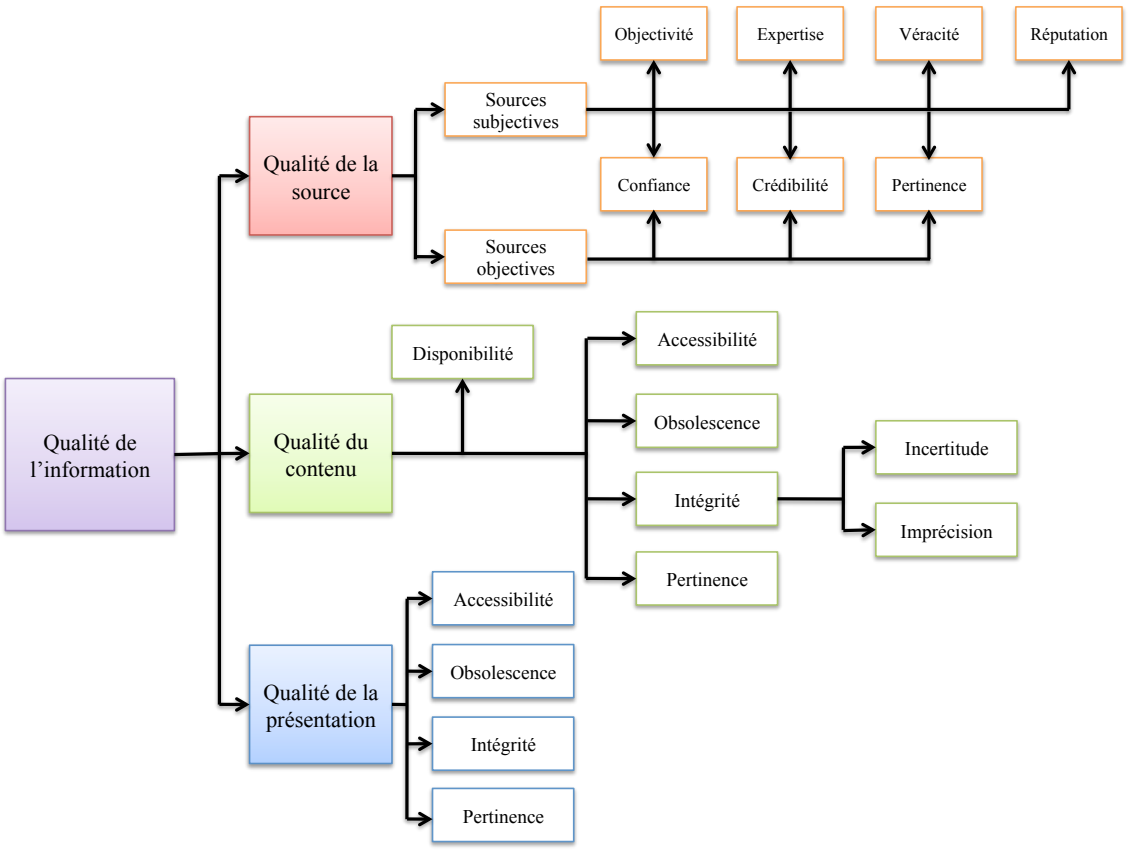


FIGURE 3.10: La qualité de l'information d'un IFS, selon [Rogova 10]

Leurs principales caractéristiques sont l'indépendance vis-à-vis des biais issus des jugements humains, les seuls facteurs affectant la qualité de ces sources étant leur calibration et l'adéquation des modèles à la réalité. Les dimensions de qualité proposées dans [Rogova 10] pour évaluer ce type de source sont : la *confiance*, la *crédibilité* et la *pertinence*. Les définitions de ces dimensions sont identiques à celles des domaines MIS et WIS, présentées précédemment.

Les sources subjectives peuvent être représentées par des hypothèses, opinions et croyances exprimées par des experts, des sources ouvertes (journaux, pages Web, chaînes de télévision, radios), les réseaux sociaux (Twitter, Facebook, etc.), etc. Ce type de source est affecté par une certaine subjectivité issue du raisonnement humain. L'évaluation de la qualité des sources subjectives doit prendre en compte, en plus des dimensions de qualité des sources objectives, l'*objectivité*, le *niveau d'expertise*, la *véracité* et la *réputation* de ces sources. Toutes ces dimensions ont déjà été définies précédemment. Cependant, pour l'évaluation de la réputation d'une source, il est nécessaire de prendre en compte les interactions et les expériences antérieures.

Comme les dimensions de qualité caractérisant les sources d'information sont fortement dépendantes du contexte d'application, leur quantification est habituellement faite par des experts. Ainsi, une source peut être objective pour un cas d'utilisation et subjective pour un autre. Cette différentiation peut se faire soit directement par un expert, soit en utilisant une base de connaissances construite sur l'expertise d'un ou des plusieurs experts du domaine.

2. **La qualité du contenu de l'information** : décrit les caractéristiques intrinsèques de l'information. Rogova et Bossé identifient cinq dimensions. La *disponibilité* et l'*accessibilité* concerne la possibilité de l'utilisateur d'avoir accès à l'information. La disponibilité est caractérisée par une variable binaire : information disponible ou indisponible et dans ce dernier cas, toutes les autres dimensions sont non-pertinentes. En même temps, l'accessibilité décrit les coûts nécessaires (temps, argent, etc.) pour obtenir l'information utile. L'*obsolescence* est mesurée par le niveau d'utilité de l'information au moment où elle devient disponible. L'information est considérée *pertinente* si les résultats des décisions ou des actions sont influencés par celle-ci. La dernière dimension de cette catégorie est l'*intégrité*. Elle décrit la manque d'imperfection du contenu de l'information. Pour la caractériser, Rogova et Bossé l'ont décomposé en incertitude et imperfection, puis ils ont utilisé la taxonomie de Smets, voir figure 3.7.
3. **La qualité de la présentation de l'information** : caractérise la façon dont l'information est perçue par les utilisateurs et son influence sur les actions, décisions et/ou jugements entrepris. Ces dimensions ont déjà été traitées dans les paragraphes précédents.

3.5.3 Les performances d'un IFS complexe

Dans le paragraphe précédent, deux modèles d'évaluation de la qualité de l'information ont été présentées pour le cas d'un module de fusion d'information faisant partie d'un système d'information. Les systèmes de fusion d'informations sont des systèmes complexes qui sont utilisés pour répondre à des tâches difficiles. Une des applications les plus fréquentes d'un IFS est l'évaluation et la prise de conscience d'une situation complexe. Un exemple très fréquemment utilisé dans la communauté est un scénario d'un conflit armé majeur, comme celui présenté dans la figure 3.11.

Dans le cas d'un système complexe, la fusion d'informations se fait à deux niveaux : fusion bas niveau (« low-level fusion » LLF) et fusion haut niveau (« high-level fusion »- HLF). La fusion bas niveau concerne la combinaison de mesures de capteurs, de données stockées dans des bases de données, de rapports, etc. pour la détection et l'identification des entités d'intérêt. Par contre, la fusion haut niveau s'occupe de la combinaison d'informations concernant plusieurs entités, ainsi que d'informations contextuelles, afin de caractériser une situation complexe, c'est-à-dire d'en déduire à la fois les relations entre les entités et leur évolution possible et de soutenir le management des ressources [Waltz 90].

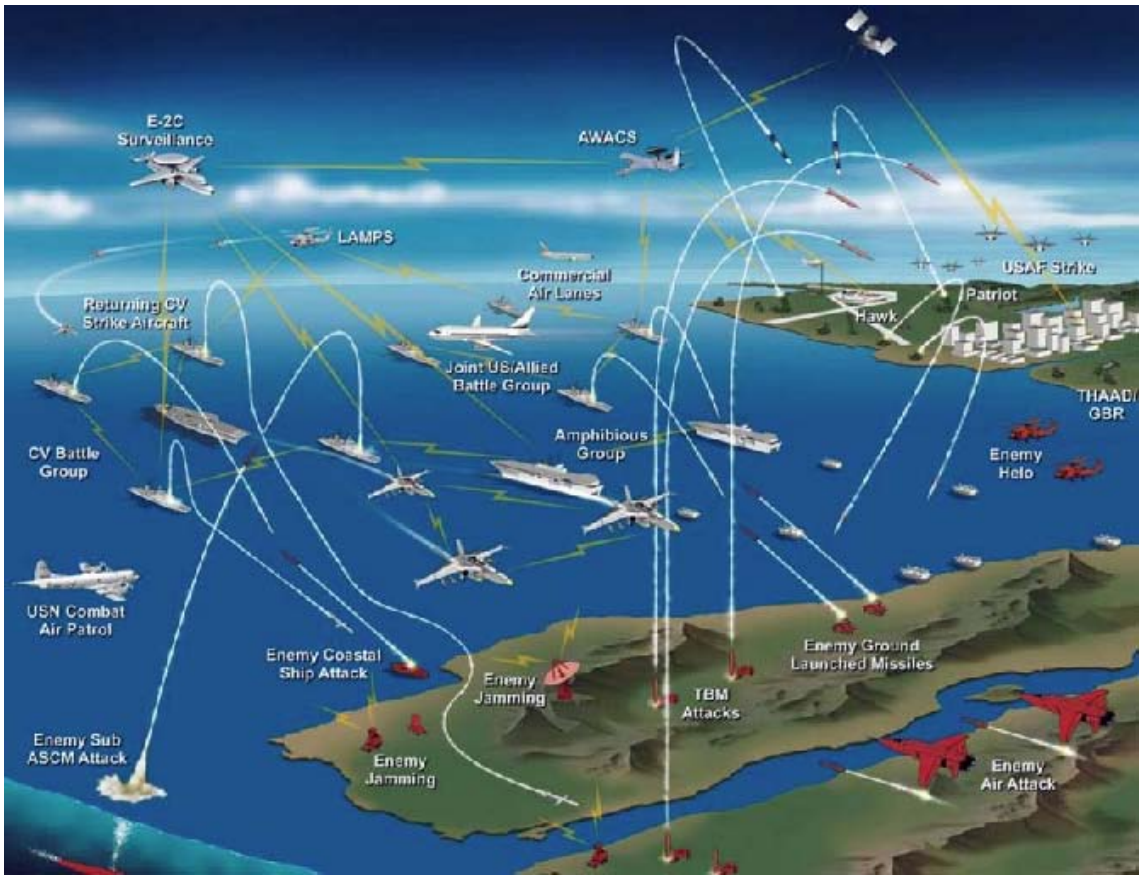


FIGURE 3.11: Scénario complexe d'un conflit armé, selon [Rempt 01]

[Costa 12] propose d'utiliser pour l'évaluation des performances d'un système de fusion d'informations de bas niveau :

- la *précision* ;
- l'*obsolescence* : la rapidité avec laquelle le système propose une décision pour un niveau de précision imposé ;
- la *confiance*.

Dans le cas particulier d'un système de fusion d'informations utilisé pour des application en défense, l'évaluation de la qualité se fait en utilisant quatre classes de mesures [Waltz 90] : les mesures des paramètres des entités physiques, la mesure des performance (MOP - « measure of performance »), la mesure de l'efficacité (MOE - « measure of effectiveness ») et la mesure de l'efficacité de la force (MOFE - « measure of force effectiveness »).

1. **Les mesures des paramètres des entités physiques** : décrivent les caractéristiques, le comportement des éléments du système :
   - le rapport signal sur bruit (SNR), la bande de fréquences, la fréquence, le nombre d'opérations par seconde, le taux d'erreurs binaires, la résolution, le taux d'échantillonnage, le coût, etc.
2. **Les mesures de performance** : sont en général dépendantes des mesures des entités physiques et comme le nom l'indique, elles décrivent les performances des attributs du système :
   - la probabilité de détection, le taux de fausse alarme, la précision de l'estimation de la localisation, la probabilité d'identification, la distance d'identification, le temps entre la détection et la transmission, le délai de communication, la couverture spatiale et fréquentielle des capteurs, la précision de classification des cibles, etc.
3. **Les mesures de l'efficacité** : caractérisent les performances du système vis-à-vis de sa capacité d'assister le bon déroulement d'une mission (pour la définition des attributs voir [Llinas 08]) :
   - le taux de nomination des cibles, l'obsolescence de l'information, la précision de l'information, le temps de réaction (« warning time »), l'immunité aux contremesures, la survie des communications, le nombre de cibles qui sort de la détection, etc.
4. **Les mesures de l'efficacité de la force** : caractérisent les performances de la mise en application des décisions proposées (d'un niveau d'abstraction encore plus haut et qui traduit la capacité d'accomplir la mission par l'ensemble de forces) :
   - le résultat du combat, le coût du système, le taux de survie, le taux de diminution de la force ennemie, « weapons on target » (même si la détection des cibles ennemies augmente, un manque de munition ne va pas augmenter la MOFE), etc.

Un des critères de l'évaluation des performances d'un système de fusion haut niveau est l'**effectivité** du système. L'effectivité d'un système peut être définie comme la capacité du système à produire un effet [Costa 12]. Parmi les attributs de l'effectivité proposés, [Costa 12] cite :

- **le rendement** : la propriété du système à réaliser les tâches d'une manière économique ;
- **l'efficacité** : identique aux mesures de l'efficacité (MOE) présentées ci-dessus ;
- **la conformité** : la propriété du système à réaliser les tâches pour lesquelles il a été déployé.

Pour mesurer l'effectivité d'un système, il faut quantifier ces attributs. Cette même étude, [Blasch 10] propose :

- **le gain d'information :** dû à la combinaison de plusieurs sources d'information (par rapport à l'utilisation individuelle des informations fournies par les sources) ;
- **la qualité de l'information :** exprimée par les mesures des performances (précision, diminution de l'incertitude, confiance, fiabilité, etc.) ;
- **la robustesse :** traduit la consistance des résultats lors des étapes de test et de production.

Toujours dans la même étude, une méthode de combinaison de ces trois mesures a été proposée afin d'arriver à une mesure générale de l'effectivité :

$$\text{Effectivite} = \text{GainInfo} \times \text{QualInfo} \times \text{Robustesse} \tag{3.5}$$

Réseaux	IFS	ATR/ID	Tracking
Retard	Obsolescence	Temps d'acquisition/traitement	Taux de mise à jours
Probabilité d'erreur	Confiance	Prob. de détection/ FA	Prob. de détection
Variations du retard	Précision	Précision de la position	Covariance
Débit	Débit	Nombre d'images	Nombre de cibles
Coût	Coût	Nb. de plateformes	Nb. de plateformes

TABLE 3.6: Exemples des mesures pour différents domaines, selon [Blasch 10]

En comparant les mesures de performances (MOP) à celles de l'efficacité (MOE), il peut s'observer que ces dernières sont évaluées par rapport à l'utilisateur. Ainsi, l'humain fait partie intégrante de l'évaluation et donc, il est nécessaire de l'intégrer dans le processus d'évaluation.

En considérant un IFS comme un système fournissant un service pour l'utilisateur, il est possible de mettre en parallèle la qualité du service (la QoS - paragraphe 3.4) et les mesures de performances d'un IFS. Dans le tableau 3.6, est présentée une comparaison entre les différentes mesures utilisées dans l'évaluation de performances de réseaux de communications, d'IFS, de systèmes de reconnaissance/identification automatique de cibles (ATR/ID) et de systèmes de tracking.

3.5.4 Conclusion

La qualité d'un système de fusion d'information est décrite dans la communauté par rapport aux incertitudes de l'information. Une information peut être affectée par plusieurs sources d'incertitudes ou être la résultante d'une fusion d'informations chacune étant affectée par ses propres incertitudes. Par conséquent, il est nécessaire d'utiliser les moyens adaptés pour la représentation de chaque aspect de l'incertitude. Ce choix va influencer les performances d'un système de fusion d'information [Blasch 10] et [Dragos 13]. L'essentiel de recherches sur ce sujet est orienté vers le développement de théories mathématiques spécifiques pour chaque type d'incertitude. Récemment une nouvelle théorie mathématique a essayé de couvrir tous les aspects de l'incertitude dans un seul cadre : la théorie de l'information généralisée [Klir 06a]. Cette théorie mathématique est introduite dans le paragraphe A.4 de l'annexe A, ensemble avec ses principaux principes de raisonnement. Une synthèse très intéressante reliant les théories mathématiques et les types d'incertitudes est présentées dans la figure 3.12.

3.6 UTILISATION DES MODÈLES DE QUALITÉ DE L'INFORMATION DANS LA PRATIQUE

Dans le développements de modèles de qualité de l'information présentées dans ce chapitre, les chercheurs ont essayé de rester rigoureux en essayant de proposer des listes de dimensions de qualité couvrant tous les aspects. Cependant, dans la pratique, les organisations se sont préoccupées à trouver les problèmes de qualité spécifiques à leur contexte d'application. Et pour cela, elles utilisent des démarches ad-hoc sans être rigoureuses d'un point de vue théorique [Lee 02] et [Borek 14]. En conséquent, les organisations ont une vision sur la qualité de l'information orientée sur leur besoin de moment, perspective différente par rapport aux recherches académiques.

Parmi les modèles de la qualité d'information présentés dans ce chapitre, celui de Wang et Strong a connu le plus grand succès pratique. Dans le tableau 3.7 sont présentés quelques exemples d'organisations et leur modèle de qualité de l'information. Les dimensions de qualité ont été classifiées selon les quatre catégories du modèle de Wang et Strong afin d'avoir une meilleure correspondance.

Les recommandations concernant la qualité de l'information du ministère américain de la défense (DoD [17]) font appel au quatre étapes du processus TDQM, illustrées dans la figure 3.3. Leur

17. The Departement of Defense

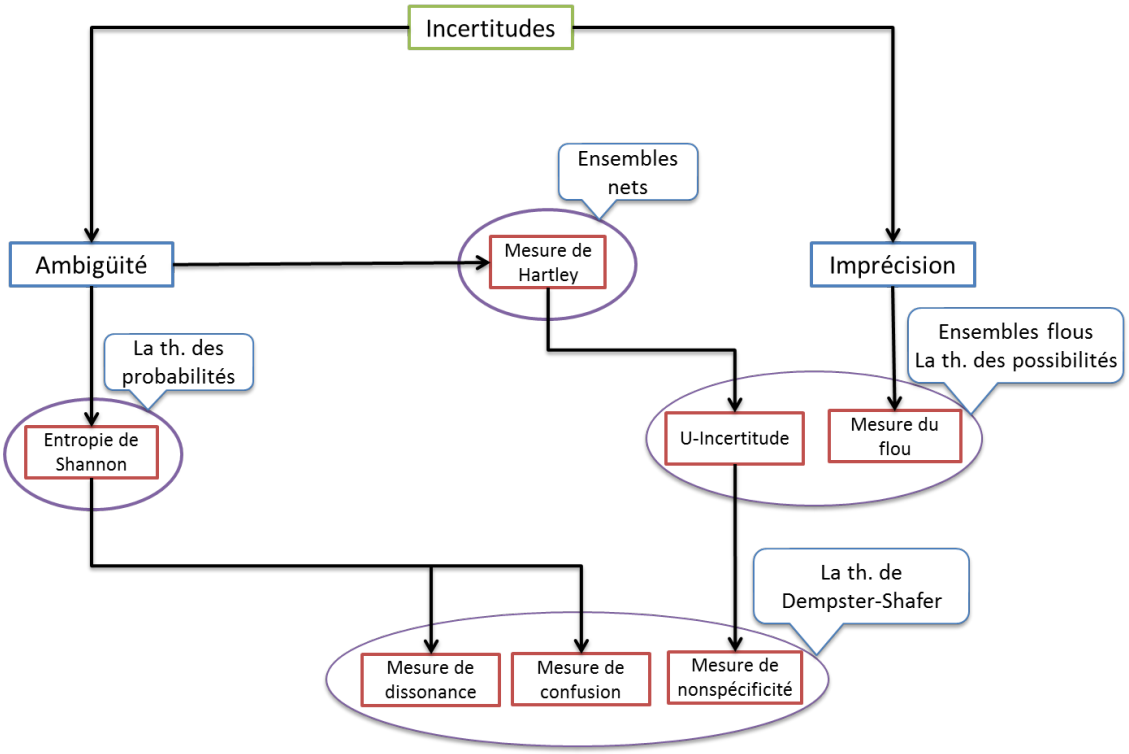


FIGURE 3.12: Les différentes mesures d'incertitude, selon [Klir 88]

Organisation	Intrinsèque	Contextuelle	Représentation	Accessibilité
US DoD [Cykana 96]	précision validité	complétude fraîcheur	unicité consistance	
HSBC [Gardyn 97]	précision	complétude fraîcheur	consistance	accessibilité
MITRE [Meyen 97]	Id. [Wang 96]	Id. [Wang 96]	Id. [Wang 96]	Id. [Wang 96]
Info. Resources [Kovac 97]	précision confiance	fraîcheur		
Unitech Syst. [Mandke 97]	précision confiance	complétude fraîcheur	consistance	sécurité privé
AT & T [Redman 92]	précision confiance	complétude fraîcheur (temps cycle) exhaustivité	clarté de la déf. redondance consistance interprétabilité portabilité	disponibilité flexibilité robustesse
OTAN STANAG 2511	réputation confiance			

TABLE 3.7: Qualité de l'information dans sept organisations

programme utilise les dimensions de qualité suivantes : la précision, la complétude, la consistance, l'obsolescence, l'unicité et la validité [Cykana 96]. À part les deux dernières dimensions, toutes les autres se retrouvent dans le modèle de Wang et Strong. L'unicité (le fait d'être la seule représentation de son type) et la validité (être suffisamment rigoureuses pour induire l'acceptation) pourraient être couvertes par la dimension *consistance*. Mais, dans leur contexte il existe des directives particulières pour ces deux dimensions. Par exemple, la validité est décrite dans la directive DOD 8320.1-M.

Dans le cas de HSBC, le modèle de qualité a été proposé pour le département de Management de Capitaux [18]. Ils ont surtout concerné les problèmes d'intégration des données dans des entrepôts et la qualité des informations extraites. La seule dimension ajoutée par rapport au modèle de Wang et Strong est l'exactitude. Mais une analyse montre que cette dimension couvre les aspects de la précision et de la conformité vis-à-vis des règles de business spécifiées par les utilisateurs.

L'entreprise MITRE a choisi de suivre *ad litreram* le modèle de Wang et Strong. Dans une de leurs études [Meyen 97], ils affirment que 35% des utilisateurs sont concernés par des problèmes d'accessibilité, 27% par la qualité intrinsèque, 24% par la qualité contextuelle et 14% par celle de la représentation. En plus, 43% des problèmes d'accessibilité étaient liés à la difficulté d'effectuer des opérations.

L'entreprise Information Resources Inc. est un fournisseur d'informations pour d'autres organisations. Ainsi, ils sont directement concernés par la qualité de l'information car leur produit commercialisé est l'information. Afin de mieux gérer la qualité des informations fournies aux clients ils ont développé un modèle de qualité appelé TRAQ : « timeliness + reliability + accuracy = quality » [19]. Ce modèle a été construit pour offrir à la fois des mesures objectives de la qualité des données et de l'information fournie par le système et un outil permettant d'améliorer et de gérer les performances du système.

Unitech Systems Inc. commercialise des outils logiciels pour l'évaluation de la qualité de l'information. Au lieu d'utiliser le mot qualité ils préfèrent employer la notion d'intégrité. Dans leur étude [Mandke 97], ils affirment que tout système d'information doit respecter les trois dimensions d'intégrité suivantes : la précision, la confiance et la consistance. Les autres dimensions sont, dans leur vision, dérivées de ces trois dimensions.

Lors de ses travaux effectués chez AT&T, Redman a proposé une des premières méthodologies d'évaluation de la qualité des données et de l'information. Dans son ouvrage [Redman 92], à la page 66, il fait un résumé de l'ensemble de dimensions de qualité identifiées dans ses travaux. Comme il a travaillé dans l'intégration de bases de données volumineuses et hétérogènes, la grande majorité de dimensions présentées concernent la qualité de représentation. Dans le tableau 3.3 sont présentées les plus importantes dimensions, mais pour une liste quasi-exhaustive on recommande [Redman 92].

Le standard STANAG 2511 de l'OTAN concerne l'évaluation de toute information obtenue dans l'objectif d'être utilisée par un service d'intelligence. Pour cela, deux critères sont évalués : la *réputation* de la source d'information et la *confiance* du contenu de l'information. Le système de quantification de ces critères utilise des valeurs alphanumériques de A à F pour la réputation et de 1 à 6 pour la confiance.

3.7 Synthèse sur les modèles de qualité de l'information

Le plus souvent la qualité de l'information est définie comme étant le degré d'utilité vis-à-vis de son usage par l'utilisateur. Des technologies de contrôle de la qualité ont été développées et utilisées depuis les années 70 dans l'industrie. Dans ce cadre industriel, ont été définies des méthodologies comme par exemple : « Total Quality Control » [Wilkinson 98] (le programme de

18. Asset Management
19. Trad. en fr. : « obsolescence + confiance + précision = qualité »

MIT sous la coordination de R. Wang), « Six Sigma » [Coskun 10] ou encore « Statistical Process Control » [Redman 92]. Malheureusement, il n'est pas encore prouvé que ces méthodologies et techniques peuvent directement s'appliquer à la modélisation et à l'évaluation de la qualité de l'information. Les inconvénients de ces méthodologies pour l'application à l'évaluation de la qualité de l'information sont liés aux propriétés particulières de l'information [Stvilia 08] : le manque de propriétés physiques, la dépendance du contexte, la non-linéarité de l'information (c'est-à-dire l'information de l'ensemble ne peut pas s'exprimer comme étant la somme de ses composantes), l'instabilité et l'apparition pas nécessairement au hasard des erreurs dans l'information.

Une observation générale caractérisant tous les modèles présentés dans ce chapitre est que la notion de qualité a tendance à être évaluée en utilisant quasiment toujours les mêmes dimensions : la précision, la confiance, l'obsolescence, l'accessibilité, la pertinence, etc. Les différences consistent dans la façon dont ces dimensions sont évaluées. De plus, dans ces modèles, la qualité de l'information est définie comme une extension de la qualité des données. La seule différence entre les deux évaluations est l'ajout dans le cas de la qualité de l'information de quelques dimensions supplémentaires.

Les modèles développés dans le cadre de MIS et de WIS considèrent le système d'information comme une « *boîte noire* », dans le sens qu'ils analysent la qualité en entrée du système (la qualité des données) et en sortie du système (la qualité de l'information). Ainsi, ils ont une vision restreinte du système d'information, sans s'intéresser aux effets des divers traitements. Cette vision est utile pour l'évaluation finale d'un produit informatique, mais pour un système d'information (complexe) il est nécessaire d'avoir une vision locale qui permette de justifier les choix des modules de traitement et d'indiquer la provenance de l'information.

Si les notions de données et d'information sont considérées comme différentes (c'est-à-dire considérant la définition du paragraphe 1.3.2), il peut s'observer dans un premier temps qu'une faible qualité de données va implicitement entrainer une faible qualité de l'information (le principe GIGO [20]). Dans un deuxième temps, une faible qualité de l'information n'entraine pas forcement une faible qualité de données. La faible qualité de l'information peut être le résultat d'un traitement non-adapté sur des données de bonne qualité. Une autre situation intéressante est celle où les données sont d'une qualité moyenne et l'utilisation d'une chaîne de traitement adaptée rend une information de très bonne qualité à l'utilisateur. Donc, comme conclusion finale, il faut dire que la qualité des données et la qualité de la chaîne de traitement (l'adéquation et les performances des traitements) rendent la qualité de l'information fournie à l'utilisateur.

Par rapport aux systèmes MIS et WIS, les systèmes de fusion d'information considèrent les différents aspects de l'information à travers le système. Ainsi, le modèle de Lefebvre (présenté dans le paragraphe 3.5.2.1) associe différents types d'incertitudes aux données et aux informations issues des divers traitements. C'est une idée très intéressante parce que les caractéristiques sémantiques de données/information évoluent au fur et à mesure des traitements. En conséquence, cette évolution devrait également induire un changement de définition de la qualité de cette information. Malheureusement, ce modèle contient un nombre réduit de dimensions de qualité, étant orienté vers l'évaluation des dimensions intrinsèques de la qualité. Cependant, le modèle de Rogova et Bossé (présenté dans le paragraphe 3.5.2.2) offre un cadre plus élargi, inspiré du modèle de Wang et Strong et de la taxonomie de Smets. Pourtant, Rogova et Bossé ne considèrent que l'analyse du module de fusion, d'où la principale limitation de ce modèle. Le modèle le plus complet d'évaluation d'un système IFS est celui présenté dans le paragraphe 3.5.3. Malheureusement, les mêmes critiques que dans le cas des systèmes MIS et IFS peuvent être soulevées.

Dans les cas où le système est vu comme une boîte noire, les dimensions de la qualité de l'information ont été évaluées en utilisant des questionnaires fournies aux utilisateurs. C'est la seule technique d'évaluation possible à cause de la complexité du système et de l'ignorance des différents traitements subis par les données/informations. La principale justification au recours à cette technique d'évaluation de la qualité a été le caractère subjectif de beaucoup de dimensions :

20. De l'anglais « Garbage In Garbage Out »

dépendantes du besoin de l'utilisateur et de son niveau de satisfaction. Mais dans la pratique, cette méthodologie d'évaluation de la qualité n'est que partiellement satisfaisante. Une des raisons pour laquelle ce genre de méthodologies n'est pas satisfaisant est l'impossibilité de saisir les influences, négatives ou positives, induites par une ou plusieurs dimensions de qualité (suite aux modifications de valeurs de qualité) vers les autres dimensions. Ainsi, par exemple, une augmentation de la précision de données, grâce aux enregistrements de plus longue durée peut avoir un impact négatif sur la « fraîcheur » (dans le cas d'un système de classification cette dimension est équivalente au rappel - dans le couple précision/rappel), une dimension de la qualité critique dans certains domaines d'applications. Un autre exemple de dépendances entre dimensions est celle entre la complétude et la consistance [Ballou 03]. De plus, si le système évolue suite à des mises à jours des modules de traitements ou à l'ajout d'autres modules de traitement, ce type de méthodologie va être incapable de s'ajuster à ces changements. Ainsi, elle oblige à un nouveau processus d'évaluation de la qualité de l'information par des formulaires, démarche qui demande beaucoup de temps.

« You should call it entropy, for two reasons. In the first place your uncertainty function has been used in statistical mechanics under that name, so it already has a name. In the second place, and more important, nobody knows what entropy really is, so in a debate you will always have the advantage. »

La réponse de J. von Neumann à C. Shannon (selon Tribus et McIrvine, 1971)

DEUXIÈME PARTIE : MÉTHODOLOGIE D'ÉVALUATION DE LA QUALITÉ DE L'INFORMATION

Dans cette partie, nous proposons une nouvelle méthodologie d'évaluation de la qualité de l'information. Le cœur de mon doctorat constitue en l'étude de la qualité au sein du SI, c'est-à-dire en chacun de ses modules. Grâce à la décomposition du SI, nous proposons, dans le premier chapitre de cette partie, d'étudier la qualité selon deux niveaux : local, en entrée et en sortie de chaque module, et globale, en sortie du SI. Dans le deuxième chapitre de cette partie nous introduisons un outil innovant permettant de modéliser l'influence du module de traitement sur la qualité. Dans le troisième chapitre de cette partie, nous montrons comment la qualité globale peut être automatiquement estimée en utilisant la qualité locale.

TELECOM Bretagne

4

Qualité locale versus Qualité globale

Dans le chapitre précédent 3, un état de l'art a été réalisé sur la définition et l'évaluation de la qualité de l'information. On a pu observer que (à notre connaissance) toutes les méthodologies considèrent le système d'information comme une boîte noire. Ainsi, celles-ci se sont plutôt intéressées à l'évaluation des performances moyennes du système d'information et non pas de la qualité de **chaque** information fournie par le système.

En même temps, dans le processus de prise de décisions, l'utilisateur a besoin connaître la qualité de chaque information individuelle qui lui est proposée par le système, et si possible, que celui-ci lui propose la possibilité d'obtenir une explication sur la provenance de cette qualité. Une solution à cette question est l'exploitation de la structure interne du système d'information afin de pouvoir modéliser les différents niveaux de qualité de l'information à l'intérieur du système d'information.

Pour cela, il est nécessaire de décomposer le système en ces modules élémentaires. Bien sûr, il faut pouvoir avoir accès aux modules de traitement afin de pouvoir les analyser. Cette hypothèse est vérifiée dans la pratique car, dans la plupart du temps, les modules de traitement sont déjà construits et donc ils existent physiquement. Ainsi, il est possible d'obtenir les connaissances nécessaires sur le comportement et le fonctionnement de chacun de ces modules. Suite à la décomposition du système d'information, la notion de qualité peut être définie et analysée à deux niveaux :

1. **niveau local** : caractérisant la qualité en entrée et en sortie de chaque module élémentaire de traitement ;
2. **niveau global** : caractérisant la qualité en sortie du système d'information.

La figure 4.1 présente ces deux visions de la notion de qualité, locale et globale, obtenues suite à une décomposition d'un système d'information.

Le paragraphe suivant 4.1 présente la décomposition du système d'information dans ces modules élémentaires. Ensuite, le paragraphe 4.2, présente la stratégie d'analyse de chacun de ces modules de traitement afin de pouvoir définir et évaluer la qualité locale.

4.1 Présentation du SI décomposé

Selon [Caseau 07], l'architecture d'un système d'information peut s'analyser en distinguant deux visions clefs du système d'information :

- **l'architecture fonctionnelle**, c'est-à-dire l'organisation des fonctions et des processus métiers qui sont supportées par le système ;
- **l'architecture technique**, c'est-à-dire l'organisation des modules de traitement de l'information. Ces modules ont pour rôle d'offrir les fonctionnalités métiers aux utilisateurs.

Dans le cas des systèmes d'information complexes, implémentés dans le cadre d'une organisation de taille importante, l'architecture fonctionnelle est la seule visible par l'utilisateur final. Ainsi, elle « correspond à une organisation hiérarchique des échanges de flux d'objets métiers entre les fonctionnalités métier » [Caseau 07]. Comme l'objectif de ce chapitre est de modéliser l'évolution de la qualité à travers le système, il ne sera considéré que l'architecture technique.

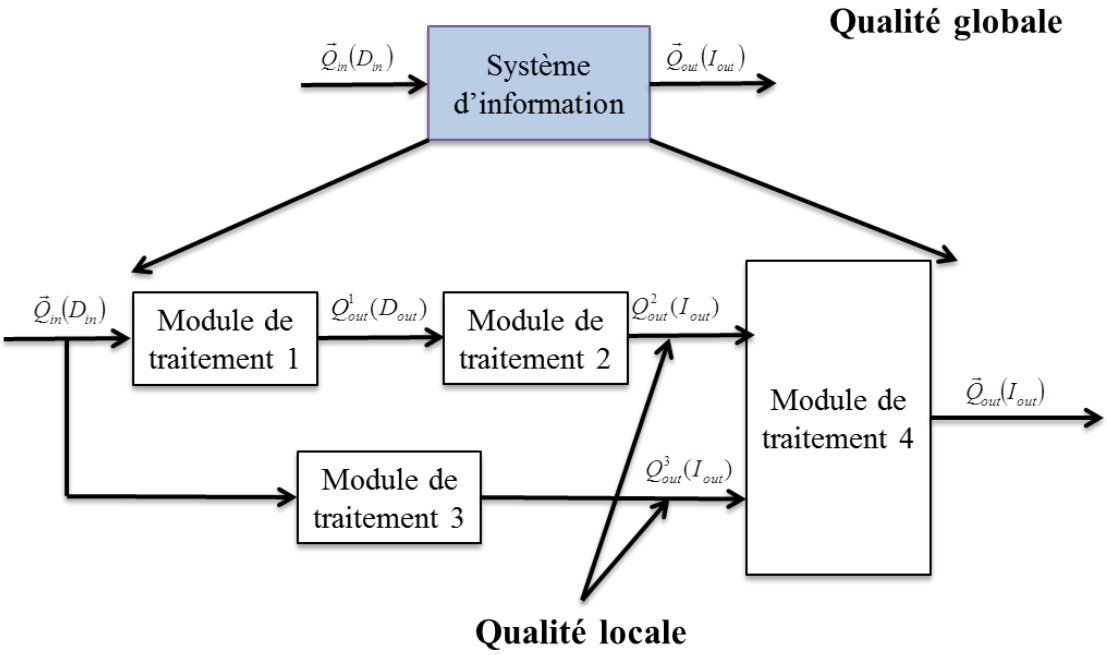


FIGURE 4.1: La qualité locale et globale dans un système d'information

Nous allons présenter maintenant l'approche hiérarchique de la décomposition de l'architecture technique. Plus précisément, il s'agit d'un processus de décomposition descendante [1] du système d'information. Comme déjà indiqué dans la section 1.1.3, nous sommes intéressés par des architectures modulaires. Ainsi, par la suite, nous allons étudier la décomposition des systèmes d'information modélisés par des diagrammes d'architecture composés de modules de traitement élémentaires, représentés par des boîtes, des flèches ou des liens, qui représentent les interactions entre ces modules. Une illustration de la décomposition d'un système d'information modulaire est présentée dans la figure 4.1. Dans ce cas, le système d'information est composé de quatre modules élémentaires de traitement.

Le module élémentaire (atomique) de traitement de l'information peut être considéré comme une boîte noire dans le sens qu'il [Borysowich 11] :

- a des entrées et des sorties prédéfinies et fixes ;
- réalise des traitements qui sont inconnus et non-pertinents pour les autres modules ;
- a un comportement prévisible, c'est-à-dire la fonctionnalité de l'algorithme de traitement est correctement définie et implémentée.

Ainsi, un module élémentaire de traitement de l'information peut être utilisé par les autres modules en interactions avec lui, sans faire appel à sa structure interne. Un des avantages majeurs d'avoir un système décomposable en modules élémentaires est la possibilité de remplacer un module par un autre, équivalent, sans influencer le fonctionnement des autres. C'est la principale raison pour laquelle il est préférable d'avoir des modules ayant des couplages faibles, c'est-à-dire être quasi-indépendants. Dans ces situations, il faut seulement connaître les données/informations échangées par les modules et leurs paramètres de fonctionnement. Cet aspect sera traité plus en détail dans le paragraphe suivant 4.2.

Les flèches connectant deux modules de traitement d'un diagramme d'architecture sont supposées traduire l'existence d'une ou plusieurs liaisons entre ces deux modules. Comme l'objectif de

1. En anglais : « top down »

cette thèse est l'étude de la qualité, la nature et le nombre d'échanges de données ou d'informations entre deux modules de traitement, c'est-à-dire le flux de données/informations, ne seront pas explicitement représentés, sauf dans le cas où la nature des données ou des informations sera mise en relation avec la définition de la qualité.

Par la suite, nous faisons l'hypothèse que l'architecture du système d'information peut être obtenue, par exemple, de la part de l'analyste du système. Ainsi, la décomposition du système d'information en modules élémentaires est supposée réalisée. A la fin de ce paragraphe, le problème de mise à jour d'un système d'information est traité, soit par une mise à jour d'un ou de plusieurs modules, soit par le remplacement d'un ou de plusieurs modules.

Comme chaque module élémentaire d'un système d'information est inter-connecté avec d'autres modules, lors d'une mise à jour il faut respecter quelques règles :

- le nouveau module devrait accomplir les mêmes fonctions que l'ancien. Donc, une mise à jour d'un module oblige de fournir toutes les fonctionnalités que l'ancien module était censé de livrer. Cela n'empêche pas d'améliorer les performances de ces fonctionnalités ou d'en ajouter de nouvelles.
- les entrées du nouveau module devrait être compatibles avec les données/informations qui lui sont fournies par les autres modules. Ainsi, le nouveau module devrait être capable d'utiliser les mêmes flux des données/informations que l'ancien. Une source importante d'erreurs est l'ignorance du format de représentation de données/information. Par exemple, dans le cas de traitement d'images, il existe des formats avec compression (jpeg, gif, etc.) ou sans compression (raw) et donc, le nouveau module devrait être compatible avec le format d'images qui lui est fourni.
- les sorties du nouveau module devrait être adaptées aux entrées des modules les utilisant. Comme dans le cas précédent, le nouveau module est obligé de fournir, dans le bon format, les données/informations demandées par les autres modules avec lesquels il interagit.

Ces principes à respecter dans le processus de mise à jour sont illustrés dans la figure 4.2. En bas de la figure 4.2, sont présentés trois modules successifs de traitements, i-1, i et i+1, faisant partie du même système d'information [2]. Leurs entrées et sorties, représentées par des figures géométriques, sont adaptées aux autres modules. Soit la situation dans laquelle le Module i devrait être remplacé par un autre et que l'analyste du système dispose de quatre modules implémentant la même fonction que le Module i. En respectant les règles d'adaptation du nouveau module aux modules avec lesquels il va interagir, seul le Module i_3 pourrait être utilisé. L'utilisation des autres modules va entraîner des problèmes de communication entre ces modules et le système d'information peut être en danger de ne plus fonctionner correctement.

4.2 Évaluation de la qualité locale de l'information

Dans ce paragraphe, le processus d'analyse d'un module élémentaire de traitement est présenté d'un point de vue de la qualité. Plus précisément, la stratégie de définition et d'évaluation de la qualité des données et/ou de l'information en entrée ou en sortie d'un module élémentaire est étudiée. Nous avons appelé cette qualité (cf. [Todoran 13]), *qualité locale*.

Dans les chapitres précédents, la notion de qualité a été définie et présentée pour plusieurs domaines d'application. Pourtant, aucune définition formelle n'a pas été donnée. Ainsi, dans le paragraphe suivant 4.2.1, nous proposons une définition formelle fondée sur les observations faites lors des chapitres 2 et 3. Puis, dans le paragraphe 4.2.2, le processus d'analyse de la qualité en sortie d'un module de traitement quelconque est défini et illustré. Dans le paragraphe 4.2.3, nous proposons une modélisation informatique de la qualité. Celle-ci a pour objectif de rendre le processus d'analyse de la qualité facilement applicable dans la pratique.

2. Afin de mieux visualiser les caractéristiques d'entrée et de sortie des modules, le système d'information est vu sous la forme d'un pipeline, les modules s'imbriquant un dans l'autre

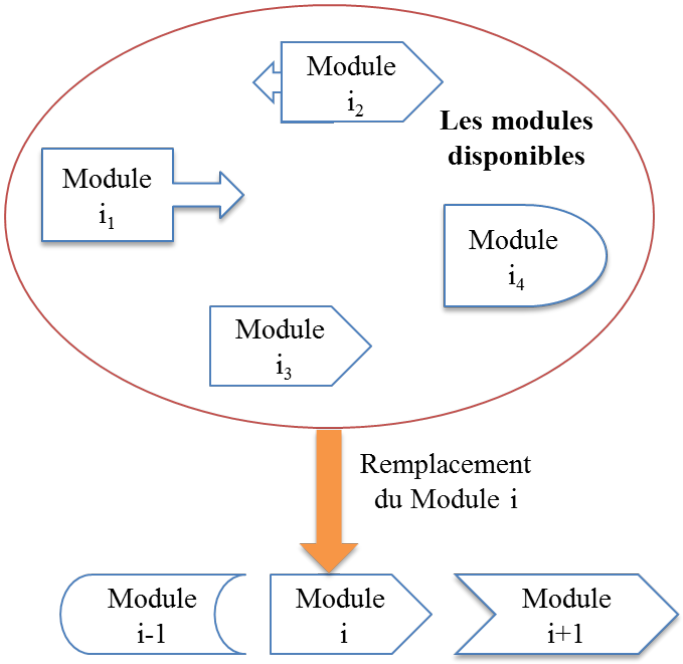


FIGURE 4.2: La mise à jour d'un module de traitement de l'information

4.2.1 Formalisation du concept de qualité des données et de l'information

Lors de l'introduction sur la définition de la qualité, paragraphe 1.4, il a été précisé que la qualité traduit les propriétés voulues de l'entité étudiée, dans ce cas les données et les informations. Dans le paragraphe 1.3, lors de la définition des données et de l'information, deux caractéristiques fondamentales de celles-ci ont été identifiées : la *quantité* (couvrant le type et la valeur) et la *sémantique*.

Les listes des dimensions de qualité proposés dans les différents types de contextes d'application et présentées dans le chapitre 3 peuvent être considérées comme quasi-exhaustives, dans le sens qu'elles couvrent tous les aspects des données et de l'information. Ainsi, ces listes de dimensions constituent une bonne base pour la construction d'une méthodologie d'évaluation de la qualité. Comme un processus d'évaluation fait appel à des critères, nous avons choisi d'appeler les dimensions de qualité : critères de qualité.

Cependant, en fonction des caractéristiques des données ou de l'information, seule une partie de ces critères est applicable. En conséquence, il est nécessaire de les adapter, en ajoutant des méta-informations, afin de les rendre plus facilement utilisables. Plus précisément, il faut utiliser ces critères dans un cadre permettant la sélection de « bons » critères. De plus, la simple utilisation de critères de qualité n'est pas suffisante. Il est nécessaire d'associer à chaque critère de qualité les outils adaptés pour leur quantification, c'est-à-dire les *mesures de qualité*.

En poursuivant les idées énoncées ci-dessus, nous présentons une formalisation de la notion de qualité des données ou de l'information dans la figure 4.3. Le module de traitement générique de cette figure est supposé recevoir une information en entrée et fournir en sortie également une information.

Observation : quand un module de traitement, faisant partie d'un système d'information est analysé, il est suffisant d'évaluer la qualité de données/informations à sa sortie parce que la qualité à son entrée a déjà été évaluée lors de l'étude du module précédent.

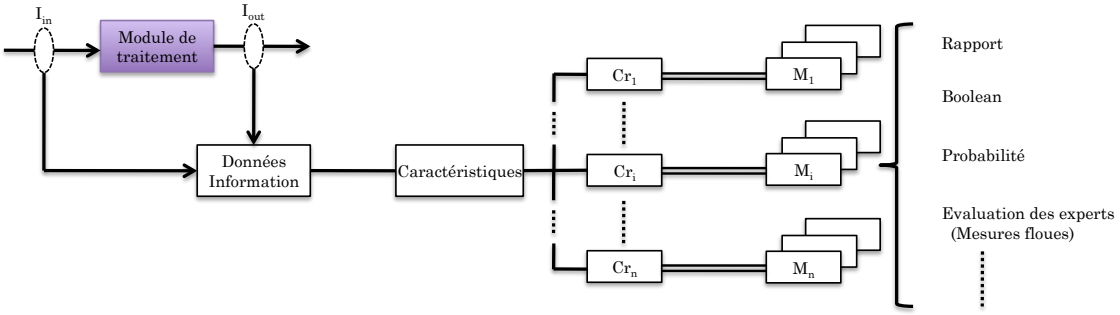


FIGURE 4.3: La formalisation de la qualité de l'information

Supposons que la qualité de l'information en sortie du module de traitement est à analyser. Cette information possède des **caractéristiques** dépendante du fonctionnement du module de traitement et du contexte d'application. En fonction de ces caractéristiques seulement quelques **critères de qualité** (notés Cr_i) seront utilisés pour décrire la qualité de l'information. Parmi les caractéristiques de données ou d'information se retrouvent :

- le type : série temporelle, image, liste d'identités, etc. ;
- la valeur : numérique (valeurs booléennes, naturelles, réelles, dans un intervalle $[a, b]$, etc), chaîne de caractères, linguistique, etc. ;
- la sémantique : ce que cette information représente, la signification du message transmis.

Soit l'exemple d'un module de traitement d'images implémentant un débruitage d'images et un autre, implémentant une segmentation d'images. Les deux modules de traitement vont fournir en sortie une image, ayant donc le même type, mais le contenu sémantique de ces deux images sera différent et donc, impose l'utilisation de critères de qualité différents. L'association de critères de qualité aux caractéristiques des données et/ou de l'information est illustrée dans le cas des deux exemples concrets aux chapitres 7 et 8.

Afin d'utiliser les critères de qualité, dans la pratique il est nécessaire d'associer à chacun d'entre eux, des **mesures de qualité** (notés M_j). Ces mesures de qualité ont comme objectif la quantification des critères de qualité en les associant des valeurs numériques ou symboliques. La démarche classique dans le développement de mesures de qualité est de les définir dans l'intervalle unitaire $[0, 1]$. Des exemples de mesures de qualité prenant de valeurs dans l'intervalle unitaire sont les mesures de probabilité, de possibilité, d'évidence, etc.

Les valeurs numériques ne sont pas nécessairement adaptées pour tous les critères de qualité. Ainsi, dans la plupart des cas, une représentation linguistiques des valeurs de qualité se prêtent mieux à quantifier les critères fortement dépendants de contexte. L'explication est que ce genre de critères est habituellement évalué par des experts, qui utilisent dans le processus d'évaluation des termes exprimés en langage naturel. Bien sûr, même si l'évaluation de ces critères est faite en utilisant le langage naturel, afin de pouvoir l'incorporer dans les futurs traitements, il est nécessaire de la « numériser ». Un outil adapté à ce genre de problème est la théorie des ensembles flous. Dans cette théorie toute évaluation linguistique est numériquement quantifiable à l'aide d'une mesure floue [Ehikioya 99]. La cohabitation de plusieurs mesures de qualité définies dans des théories mathématiques différentes peut générer des problèmes d'interprétation. À cause de cela, le modèle mathématique le mieux adapté pour définir et construire des mesures de qualité est la théorie de l'information généralisée. Dans cette théorie, il coexistent des mesures de probabilité, d'évidence, de possibilité, floues, etc. Une description plus détaillée de cette théorie se retrouve dans l'annexe A.

Soit l'exemple d'un module de traitement réalisant l'estimation de la position d'un aéronef et délivrant cette information directement à l'utilisateur. En sortie du module la valeur de l'information sera composée de trois nombres réels correspondant aux trois dimensions spatiales (x, y, z).

La qualité de cette information sera évaluée par les critères suivants avec leurs mesures de qualité spécifiques :

- *La précision* : mesurée par une déviation standard pour chaque dimension spatiale ;
- *L'obsolescence* : le temps écoulé entre la prise de mesures et le moment de la présentation de l'information à l'utilisateur.

Dans le cas où l'utilisateur a besoin d'autres informations complémentaires de la position de l'aéronef, par exemple son identité ou sa vitesse, des critères de qualité additionnels sont nécessaires. Dans ce cas, il s'agit d'une information finale composée de multiples informations élémentaires. Un critère de qualité adapté pour ce genre de situations est la *complétude*, indiquant la présence de toutes les informations élémentaires.

4.2.2 Processus d'analyse de l'information en sortie d'un module

Dans la figure 4.4 est présenté, sous la forme d'un diagramme de flux, le processus d'analyse de l'information [3] en sortie d'un module de traitement. Pour cela, nous allons utilisé la définition formelle de la notion de qualité comme fil directeur pour le développement du processus. Ci-dessous, chaque étape de ce processus est décrite :

1) La première étape du processus est l'identification des caractéristiques (opération *idCaractéristiques*) de l'élément d'information étudié, *El_Information*.

2) La deuxième étape est d'associer à chaque caractéristique de l'information un ou plusieurs critères de qualité. Ainsi, tant qu'il reste de caractéristiques à quantifier (représenté par la boucle *caract_à_quantifier*), deux actions sont possibles :

2a) La première concerne le cas où la caractéristique courante peut être quantifiée (branche *quantifiable*) par un ou plusieurs critères de qualité et dans ce cas, l'opération *trouveCritère(s)* est lancée.

2b) La deuxième représente la situation quand la caractéristique de la qualité ne peut pas être quantifiée (branche *pas quantifiable*), c'est-à-dire il n'y existe pas un correspondant dans la liste des critères disponibles. Dans ce cas, la seule solution est de demander à un expert ou à l'utilisateur de réaliser l'évaluation.

3) Une fois les critères de qualité sélectionnés, dans la troisième étape il est nécessaire de leur associer des mesures de qualité, comme indiqué dans la figure 4.3. Deux situations peuvent être identifiées :

3a) Dans la première (branche *mesurable*), le critère de qualité est associé à une ou plusieurs mesures de qualité et dans ce cas l'opération *trouveMesure* est lancée. Ensuite, chacune des mesures est reliée à une certaine métrique indiquant l'unité et l'échelle de mesure (données par l'opération *associeMétrique*).

3b) Dans la deuxième situation (branche *non-mesurable*) le critère de qualité n'a pas une mesure de qualité associée. La procédure, dans ce cas, est de faire appel à une évaluation extérieure (opération *évalExt*) en appelant un expert. En même temps, comme il s'agit d'une caractéristique de l'information qui est quantifiable, il est nécessaire d'ajouter une mesure de qualité (opération *ajoutMesure*). Celle-ci pourrait être donnée par l'expert, sous la forme d'une formule mathématique ou sous la forme d'une évaluation générale : la caractéristique X de l'information en sortie du module courant est $valeur_X$.

La sous-étape **3b)** décrit tous les critères de qualité pour lesquels il n'existe pas de mesures de qualité exprimées sous la forme de formules mathématiques. Il s'agit surtout des critères de qualité subjectifs. Ainsi, il est intéressant de donner un exemple afin d'illustrer le processus de quantification d'un tel critère.

3. Par la suite de cette section, comme la discussion sur l'évaluation de la qualité locale est identique pour le cas des données et de l'information, par simplicité de présentation que le terme information sera utilisé pour représenter la notion de données et de l'information.

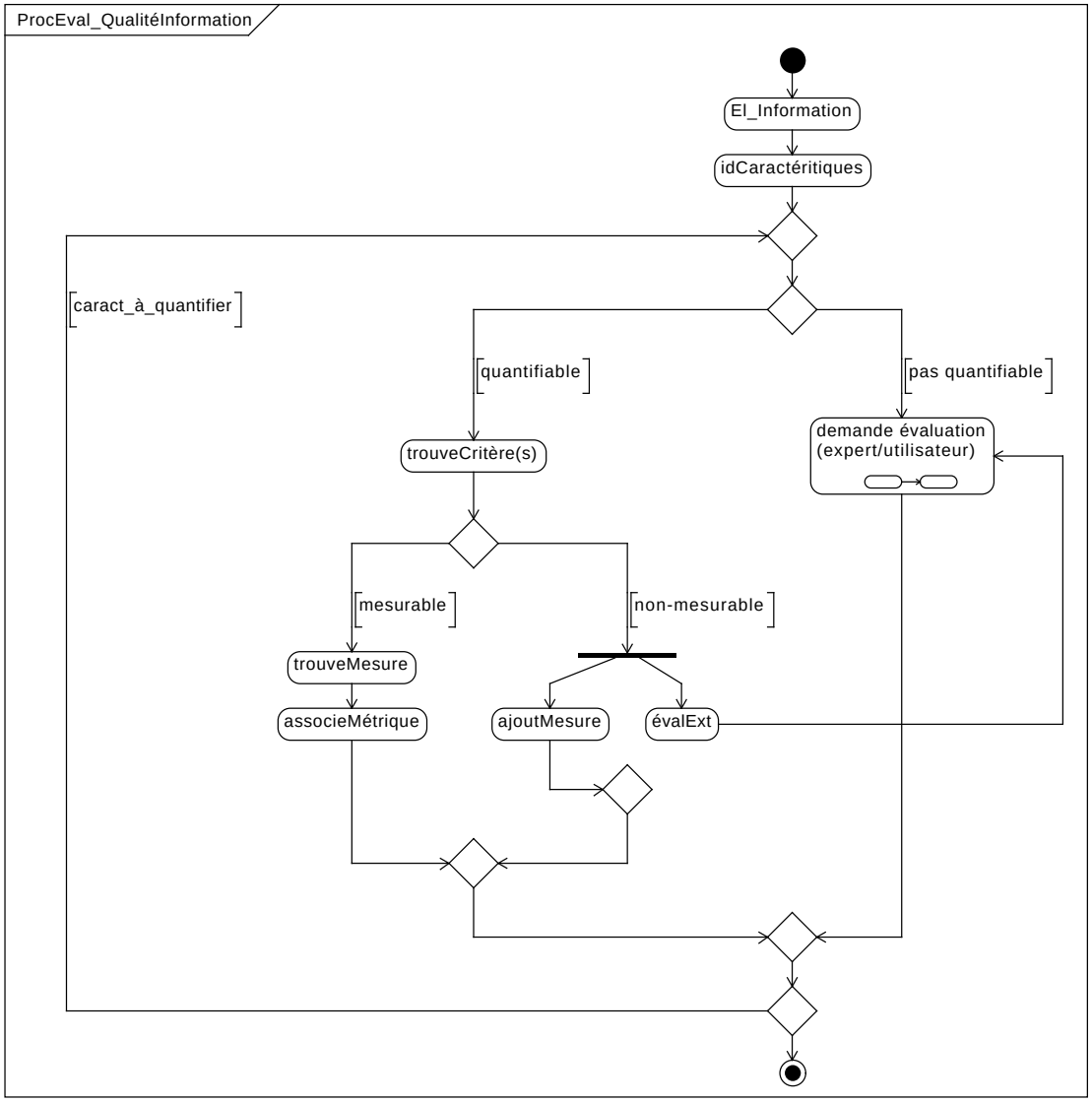


FIGURE 4.4: Le processus d'évaluation de la qualité de l'information

Exemple 1 **:** Soit une source d'information quelconque et le critère de qualité *l'objectivité*. Trouver une mesure pour ce critère est très difficile et même impossible sans avoir de connaissances *a priori* - par exemple si cette source d'information est étudiée sur une longue durée de temps, il est possible de développer une mesure statistique. Cependant, l'objectivité de cette source pourrait être évaluée par un expert du domaine qui l'a utilisée auparavant. Dans ce cas, son évaluation, par exemple sous une forme *peu objective*, peut être considérée comme généralement valable pour ce critère.

Comme les données sont différentes de l'information au niveau sémantique, la qualité de ces deux entités est également différente. Cette différence se traduit par un changement de définition des critères de qualité et donc, implicitement, par le besoin d'utilisation de mesures de qualité différentes. En conséquence, il est nécessaire d'utiliser pour chaque sens d'un même critère de qualité, une mesure de qualité différente. Soit l'exemple du critère de qualité la *complétude*. Lorsque ce critère est utilisé pour évaluer la qualité des données il exprime le degré de présence de toutes les valeurs. En même temps, lorsqu'il est employé dans l'évaluation de la qualité de l'information, la complétude exprime la présence de tous les éléments informatifs (et pas les autres) qui sont nécessaire à l'utilisateur dans le processus de prise de décisions. Ainsi, dans ce dernier cas, la complétude est définie par rapport aux besoins de l'utilisateur. Cet exemple concernant la complétude de l'information est illustré à la figure 1.9.

Par la suite, afin de montrer l'influence de la sémantique sur la qualité, nous proposons une classification des critères de qualité présentés dans le chapitre 3, en critères adaptés pour la qualité des données (tableau 4.1) et en critères adaptés pour la qualité de l'information (tableau 4.2). De plus, pour chacun de ces critères seront proposés des exemples de mesures possibles de qualité.

Même si non exhaustive, la liste du tableau 4.1 contient les principaux critères de qualité qui peuvent caractériser les données. Comme le domaine de management des systèmes d'information met l'accent principalement sur les données extraites des bases de données, la grande majorité des critères du tableau 4.1 est issue du modèle de Wang et Strong, voir le paragraphe 3.3.1. Dans la deuxième colonne de ce tableau, pour chaque critère de qualité sont présentées quelques exemples de mesures de qualité. Celles-ci sont principalement issues du chapitre 2, traitant de la qualité des données. Néanmoins, il existe des critères de qualité qui n'ont pas de mesures numériques, exprimées par des formules mathématiques. Dans ce cas, il est nécessaire de faire appel aux évaluations faites par des experts ou, si un expert n'est pas disponible, de demander aux utilisateurs de donner leur avis.

Dans la troisième colonne du tableau 4.1 est illustrée l'applicabilité de ces critères de qualité dans le cas de deux types différents de données : des données issues de capteurs et des données non-structurées. Le premier type de données concerne les données collectées par divers types de capteurs : électrique, infrarouge, électro-optiques, mécanique, acoustique, radar, sonar, vidéo, etc. Ce type de données a comme caractéristique principale un format et une structure bien définis, représentés par des valeurs numériques issues du processus de prise de mesures. Le deuxième type de données consiste de données semi ou non-structurées. Cette catégorie de données sont principalement issues de rapports d'experts, du Web (par exemple en format XML qui est semi-structuré), etc. Par rapport à l'autre type de données, celles-ci ont comme caractéristique principale la coexistence de valeurs numériques et symboliques avec des symboles, images, graphiques et d'autres objets (comme par exemple des équations mathématiques). De plus, tous ces objets se retrouvent dans un format ad-hoc. Ainsi, le critère de qualité, *précision*, qui est applicable pour les valeurs numériques, est mieux adapté aux données issues de capteurs. Cependant, le critère de qualité, *objectivité* est adapté et très important pour les données issues de sources humaines qui sont sujets des biais à cause de leur subjectivité. Ce critère est applicable pour les données non-structurées.

Comme dans le cas de l'étude de la qualité des données, le tableau 4.2 contient les principaux critères de qualité qui peuvent être définis afin de caractériser l'information. Si dans le cas des critères de qualité de données, le modèle de Wang et Strong est le mieux adapté, dans ce cas, nous avons choisi le modèle de Rogova et Bossé adapté aux systèmes de fusion d'informa-

Critère de qualité	Mesures de qualité	Type de données : ♢ Issues de capteurs ♡ Non-structurées
Précision	Déviation standard	♢
Confidence	Standards, évaluation utilisateur/expert	♢ ♡
Objectivité	Évaluation utilisateur/expert	♡
Pertinence	Évaluation utilisateur/expert	♢ ♡
Obsolescence	Fréquence de rafraîchissement	♢ ♡
Complétude	Proportion de valeurs manquantes, complétude de la population	♢ ♡
Quantité	No. d'attributs, no. d'entités, volume (bits)	♢ ♡
Interprétation	Évaluation utilisateur/expert	♡
Consistance	Par rapport au type du format, redondances	♢ ♡
Accessibilité	Temps d'accès, taux de défaillances, temps de remise en état	♢ ♡
Sécurité	Niveau de sécurité	♢ ♡

TABLE 4.1: Critères de qualité pour les données en association avec exemples de mesures de qualité

tion (voir le paragraphe 3.5.2.2). Dans la seconde colonne de ce tableau, pour chaque critère de qualité sont proposés des exemples de mesures possibles de qualité. Par rapport aux critères de qualité pour les données, dans le cas de l'évaluation de l'information, les critères de qualité ont un caractère fortement dépendant du contexte d'application et de l'utilisateur. Ainsi, les mesures de qualité doivent être définies en concordance avec les attentes et les besoins de l'utilisateur dans le contexte considéré. Afin de montrer l'applicabilité de ces critères dans des situations concrètes, deux types/formats d'informations sont considérées. Dans le premier, les informations sont présentées sous la forme d'un rapport (par exemple décrivant la situation ukrainienne) et dans le deuxième, le système d'information fournit une liste d'informations élémentaires, c'est-à-dire une énumération d'informations élémentaires (par exemple l'ensemble des objets identifiés dans un certain périmètre). La présentation des informations sous la forme d'un rapport permet de fournir à l'utilisateur, en plus des informations élémentaires, d'autres informations comme les interactions entre celles-ci ou encore une prévision sur l'évolution possible de la situation actuelle. En conséquence, pour ce type d'information il est nécessaire d'évaluer la facilité de compréhension et son objectivité, parce que dans la plupart des cas, des sources humaines (des experts) sont utilisées dans la constitution de rapports.

Suite à cette analyse concernant les critères et les mesures de qualité, pour les données et pour l'information, nous pouvons conclure que la quantification d'un critère de qualité peut se faire en employant une ou plusieurs mesures de qualité. Ainsi, par exemple, pour l'évaluation de la qualité d'une entité multidimensionnelle, le critère *précision* devrait être mesuré pour chaque dimension. De plus, en fonction de ce qui est disponible, il est possible d'utiliser des mesures incommensurables pour le même critère :

- *une mesure probabiliste*, s'il existe des connaissances sur les probabilités *a priori* et/ou si des informations statistiques sont disponibles ;
- *une mesure floue*, si l'évaluation de critères de qualité fait appel à des experts.

4.2.3 Modélisation informatique de la qualité

À partir de la formalisation de la définition de la qualité, figure 4.3, une solution informatique peut être proposée afin de facilité son implémentation dans la pratique. Ainsi, la qualité peut se

Critère de qualité	Mesures de qualité	Type d'info. : ♣ Rapport ♠ Liste d'info.
Véridicité	Degré de validité	♣ ♠
Confiance	Standards, évaluation utilisateur/expert	♣ ♠
Objectivité	Évaluation utilisateur/expert	♣
Réputation	Éval. utilisateur/expert, fondée sur les préférences personnelles et sur l'expérience professionnelle	♣ ♠
Pertinence	Degré d'applicabilité et d'utilité pour la tâche courante	♣ ♠
Obsolescence	Degré avec lequel la fraîcheur de l'information est appropriée à l'utilisation	♣ ♠
Complétude	Degré avec lequel tous les éléments d'information sont présents	♣ ♠
Compréhension	Évaluation utilisateur/expert	♣
Intégrité	Degré avec lequel l'information est consistante	♣ ♠

TABLE 4.2: Critères de qualité pour les données en association avec exemples de mesures de qualité

modéliser sous la forme d'un diagramme de classe UML [4] [Todoran 14a].

Une telle modélisation est illustrée dans la figure 4.5. Comme dans la description formelle de la qualité de l'information, le diagramme de classe UML part de l'élément des données ou de l'information à analyser, modélisé par la classe *El_Info*. Chaque élément d'information est décrit par une ou plusieurs caractéristiques, composant une liste de caractéristiques, *listeCar*. La modélisation des caractéristiques de l'information est réalisée en utilisant une nouvelle classe *Caractéristique*. Parmi les attributs de cette classe, se trouve le type de valeur de l'information (par exemple un réel, une chaîne de caractères, etc.) et la description de l'information, c'est-à-dire la signification de l'information. Par exemple, l'information en sortie d'un module peut représenter l'identité d'une entité (la description de l'information). Dans ce cas, la valeur de cette information est le nom de l'entité (par exemple « Airbus A380 ») et son type est une chaîne de caractères. Comme déjà mentionné dans la section 4.2.2 et dans la figure 4.4, il est nécessaire d'indiquer si cette caractéristique est quantifiable ou pas. Cela est représentée par la méthode de classe *estMesurable*() retournant une valeur booléenne. Cette valeur indique, pour chacune de ces caractéristiques, si un processus de mesure peut être mis en place, c'est-à-dire elle indique les caractéristiques qui peuvent être évaluées.

Avec l'élément d'information caractérisé, la prochaine étape est de lui associer sa qualité, représentée par la classe *QualitéInfo*. Cette classe est composée d'une liste non-vide de critères de qualité, *listeCritères*. La modélisation d'un critère de qualité est réalisée par une classe, *CritèreQual*. Cette classe contient deux attributs, le *nom* et la *catégorie*. L'attribut *nom* indique le nom du critère utilisé et l'attribut *catégorie* indique s'il s'agit d'un critère de qualité pour les données ou pour l'information. Toujours à ce niveau, si besoin, un nouveau critère de qualité peut être ajouté en appelant la méthode de classe *ajoutNouveauCritère*(). Chaque critère de qualité est composé d'un ou de plusieurs mesures de qualité, représentées par la classe *MesureQual*. Chaque instance de cette classe, donc chaque mesure de qualité, a comme attribut la *valeur* prise par cette mesure, le *type* de mesure (par exemple probabiliste, floue, possibiliste, etc.) et une *description* présentant la façon de calcul de cette mesure.

De plus, une métrique est associée pour chaque mesure de qualité. La classe *Métrique* a comme attributs l'*unité* et l'*échelle* de la mesure de qualité.

Dans le diagramme de classes UML, il est également représenté le contexte d'application du module de traitement, sous la forme d'une classe qui a pour rôle de décrire le fonctionnement du

4. Unified Modeling Language (UML) `http://www.uml.org/` est un outil de modélisation graphique orienté objet.

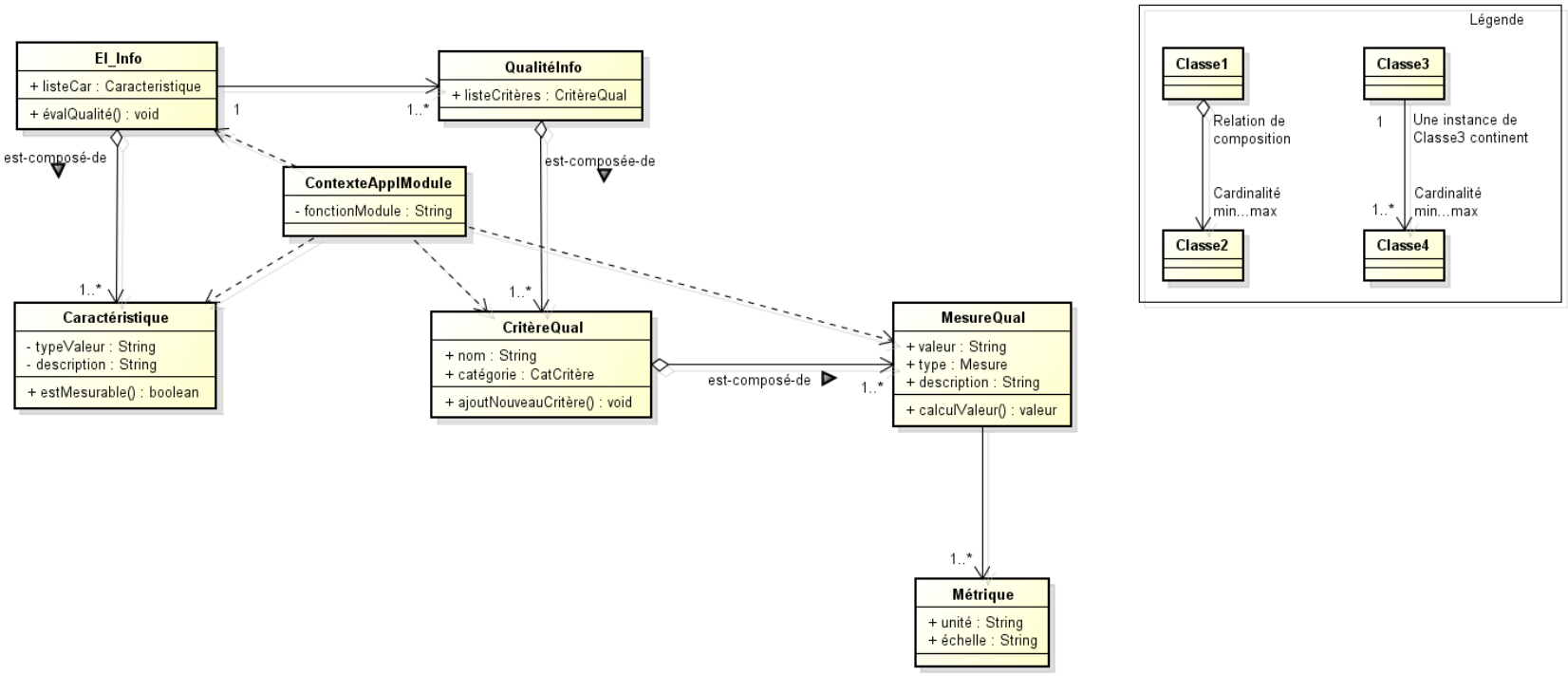


FIGURE 4.5: La modélisation de la qualité de l'information sous la forme d'une diagramme UML

module de traitement et donc, il influence :

- la détermination des caractéristiques de l'information. Celle-ci est dépendante de la classe *Caractéristique* du contexte car la description et la signification de l'information est directement dépendante du contexte d'application du module ;
- la sélection de critères de qualité (dépendance de la classe *CritèreQual* du contexte) : en fonction du besoin de l'application courante seulement un sous-ensemble de critères sont utiles ;
- le choix des mesures de qualité (dépendance de la classe *MesureQual* du contexte) : en fonction du contexte d'application, certaines mesures peuvent être préférables aux autres. C'est le cas de l'exemple 2, dans lequel plusieurs mesures d'entropie sont disponibles.

Avant de conclure ce paragraphe, un exemple est considéré afin d'illustrer l'utilité de cette modélisation informatique.

Exemple 2 : Soit le cas d'un module de traitement implémentant l'identification d'un objet parmi un ensemble de N possibles. Habituellement, l'algorithme de discrimination entre les différents identités associe à chaque identité un score de probabilité [5] (le cas d'un algorithme probabiliste) :

$$Pr_i \rightarrow \mathrm{Id}_i \quad (\forall)\ \mathrm{Id}_i\ 0 < i \leq N \tag{4.1}$$

Supposons qu'à sa sortie il fournit l'identité ayant le plus élevée niveau de probabilité :

$$\mathrm{Id}_{finale} = \{\mathrm{Id}_j | Pr_j \geq Pr_i,\ (\forall) Pr_i\ 0 < i \leq N\} \tag{4.2}$$

Cette description du fonctionnement du module correspond au contexte d'application du module. L'identité de l'objet représente l'information en sortie du module. Supposons que le critère de qualité pour cette information est le degré de *confiance*. Le contexte d'application du module, chaque identité associée à une probabilité, indique d'utiliser une mesure de qualité exploitant ces informations. Une solution est la mesure d'entropie (le type de mesure) : plus l'entropie est faible, plus la discrimination entre les identités est forte et donc la confiance est plus élevée. Dans la littérature il existe plusieurs formules de calcul pour l'entropie, développées sous différentes théories mathématiques, voir l'annexe A.8. Ainsi, une description de cette mesure de qualité peut être : entropie développée dans la théorie des probabilités et représentée par l'entropie de Shannon :

$$S = -\sum_{i=1}^{N} Pr_i \log_2(Pr_i) \tag{4.3}$$

Cette métrique, l'entropie de Shannon, prend des valeurs dans l'intervalle $[0, \log_2(N)]$ et donc une possible transformation pour obtenir le degré de confiance, exprimé dans l'intervalle unitaire $[0, 1]$ est :

$$\mathrm{Confiance} = 1 - \frac{S}{\log_2(N)} \tag{4.4}$$

4.3 Conclusion

Dans ce chapitre, le système d'information a été décomposé afin d'avoir accès à ses modules élémentaires de traitement de l'information. Suite à cette décomposition, nous avons défini deux visions différentes de la notion de qualité : **la qualité locale**, caractérisant la qualité des données ou de l'information en entrée et en sortie de chaque module, et **la qualité globale**, caractérisant la qualité du système entier.

5. ou de possibilité (le cas d'un algorithme possibiliste) ou encore de croyance (le cas d'un algorithme implémenté dans le cadre de la théorie de Dempster-Shafer)

Comme les définitions habituelles de la qualité des données et de l'information ne fournissent pas d'indications sur l'utilisation dans un cas concret, nous avons proposé un nouveau processus de définition et d'évaluation de la qualité en sortie d'un module de traitement quelconque, c'est-à-dire la qualité locale. Pour cela, nous avons proposé, dans un premier temps, une définition formelle de la qualité des données et de l'information. Ainsi, nous avons réalisé la connexion entre les caractéristiques des données et de l'information et les critères de qualité adaptés pour les quantifier.

Dans un deuxième temps, grâce à cette définition formelle de la qualité, nous avons proposé un processus d'évaluation de la qualité des données et de l'information en sortie d'un module, sous la forme d'un diagramme de flux. Ce processus, intuitif et facile à comprendre, permet d'illustrer le raisonnement pour le choix de bons critères de qualité, par rapport aux caractéristiques des données et de l'information, et leur quantification, à l'aide des mesures de qualité.

De nos jours, le paradigme le plus utilisé dans la conception de systèmes d'information est celui orienté-objet. Ainsi, nous avons choisi d'utiliser la définition formelle de la qualité et le processus d'évaluation de la qualité afin de modéliser sous la forme d'un diagramme de classes UML, un processus de définition et de modélisation de la qualité des données et de l'information. Cet outil permettra de facilement implémenter dans la pratique le modèle de qualité adapté, sous une forme informatique.

Bien sûr, l'évaluation de la qualité locale n'est utile que pour l'analyse du module de traitement. Notre objectif est de pouvoir évaluer la qualité du système d'information entier. Ainsi, dans les deux prochains chapitres nous allons faire appel à la qualité locale pour faciliter le passage à la qualité globale.

5 Modélisation de l'influence d'un module de traitement sur la qualité de l'information

Dans ce chapitre, nous analysons la possibilité de modéliser l'influence du module de traitement sur la qualité de l'information en sortie. Plus précisément, la modélisation d'un module de traitement est étudiée par une « fonction » permettant de relier les qualités de l'information à l'entrée et à la sortie du module. L'intérêt étant de modéliser l'influence du module sur la qualité, seuls les échanges en terme de qualité seront représentés et analysés. Par conséquent, le flux de données et d'informations à travers le système d'information ne sera pas étudié. Néanmoins, lorsque les valeurs de l'information devront être prises en comptes, elles seront explicitement représentées.

Avant de commencer l'étude sur la transformation de la qualité à travers un module de traitement, il est précisé que pour des raisons de simplicité de l'exposé, seul le terme information sera utilisé. Ainsi, toute discussion portant sur la qualité de l'information est généralisable à celle portant sur la qualité des données. De plus, comme précisé dans le paragraphe 4.2, il est supposé que tous les modules de traitement sont accessibles à l'étude, c'est-à-dire que les connaissances sur les entrées, les sorties et le fonctionnement du module sont supposées connues.

Dans le paragraphe 5.1 nous introduisons un outil innovant sous la forme d'une fonction mathématique, permettant de modéliser l'influence d'un module de traitement sur la qualité. Ensuite, dans les paragraphes 5.2 et 5.3, nous présentons deux façons de calculer cette fonction : *analytique* et *non-analytique*.

5.1 Fonction de transfert de qualité

Sous ces hypothèses, la figure 5.1 présente un module de traitement élémentaire d'un système d'information. À son entrée, le module reçoit l'information I_{in} et à sa sortie il fournit l'information I_{out}. En parallèle avec le flux d'informations, le flux de qualité est également représenté. Le paragraphe 4.2.2 présente le processus d'évaluation de la qualité locale. Ce processus permet d'évaluer la qualité de l'information en entrée du module, Q_{in} et la qualité en sortie du module, Q_{out}. Chaque fois que la qualité de l'information en entrée du module change, la qualité en sortie du module change aussi. Malheureusement, le processus d'évaluation de la qualité locale demande beaucoup de temps et donc, il n'est pas réaliste de l'utiliser chaque fois qu'un changement de qualité intervient. Ainsi, il est préférable de développer un outil permettant de mettre à jour automatiquement la qualité en sortie du modèle une fois qu'un changement de qualité en entrée a été détecté.

Dans le domaine du traitement du signal, il est possible de déterminer le signal $s_o(t)$ en sortie d'un filtre. Le signal $s_o(t)$ est déterminé par le résultat de la convolution entre le signal en entrée du filtre, $s_i(t)$ et la fonction de transfert du filtre, $h(t)$: $s_o(t) = s_i(t) * h(t)$. Ainsi, il n'est pas nécessaire d'utiliser une sonde pour déterminer le signal à chaque endroit du système, il suffit de connaître le signal en entrée du système et ensuite, en utilisant les fonctions de transfert de filtres, le signal peut être directement calculé à chaque instant.

Pour pouvoir réaliser une évaluation automatique de la qualité en sortie d'un module, dans [Todoran 13] nous avons introduit une fonction $\mathcal{Q}_f$ qui a pour rôle de caractériser l'influence sur la qualité du module de traitement. Cette fonction est appelée : **fonction de transfert de qualité**, en analogie avec la fonction de transfert du domaine du traitement du signal, figure 5.1.

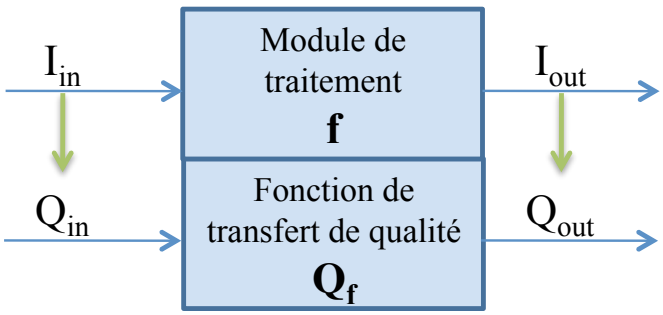


FIGURE 5.1: La fonction de transfert de qualité d'un module de traitement

L'estimation de la *confiance* d'une information en sortie d'un module par rapport à son entrée est un problème classique qui a été étudié en profondeur dans le cadre de systèmes de contrôle [Srivastava 83]. Dans ce cas d'application, le critère de qualité nommé *confiance* décrit complètement la qualité en entrée, C_{in} et la qualité en sortie C_{out}. Afin de pouvoir estimer la confiance à sa sortie, [Srivastava 85] a modélisé le module de contrôle à l'aide des paramètres suivants :

- P_w qui décrit la probabilité du module de contrôle en état de fonctionnement ;
- P_c qui décrit la probabilité du module de contrôle de prendre une décision correcte sachant que l'information à son entrée est correcte ;
- P_e qui décrit la probabilité du module de contrôle de prendre une décision correcte sachant que l'information en entrée est incorrecte, c'est-à-dire le degré de détection et de correction d'erreur.

Ainsi, le module de contrôle est caractérisé par un modèle probabiliste. De plus, si les confiances en entrée et en sortie du module sont définies dans un cadre probabiliste, l'expression analytique de la confiance en sortie du module de contrôle est donnée par [Srivastava 85] :

$$C_{out} = C_{in}\left[1 - P_w + P_w(P_c - P_e)\right] + P_w P_e \tag{5.1}$$

L'équation 5.1 est un exemple de fonction de transfert de la qualité reliant la qualité en entrée du module, C_{in} et celle en sortie du module, C_{out}. Malheureusement, le modèle du domaine de système de contrôle prend en compte une seule dimension de la qualité de l'information, la *confiance*. Cependant, dans les chapitres précédents, nous avons montré que la qualité est une notion multidimensionnelle.

Avant de rentrer plus en détail sur la définition et le développement des fonctions de transfert de qualité, nous présentons une discussion sur l'influence des caractéristiques de l'information en entrée du module, I_{in} sur la qualité de l'information en sortie, Q_{out}. Sous l'hypothèse que le système opère dans un environnement complexe et dynamique, il est évident que les caractéristiques de l'information en entrée de chaque module de traitement évoluent dans le temps. En conséquence de cette évolution, la qualité de l'information évolue également puisqu'elle est directement définie en lien avec les caractéristiques de l'information, cf. paragraphe 1.4. En même temps, un changement de l'information en entrée du module de traitement n'implique pas obligatoirement un changement de l'information en sortie du module. Tout dépend du module de traitement de l'information. Cependant, même si la valeur de l'information reste identique, sa qualité peut changer. Soit, le cas d'un module réalisant une classification et la situation dans laquelle le niveau de bruit diminue (pour visualisation la figure 5.2 pourrait être utile). Les conséquences de cette diminution de bruit sont que la qualité en entrée du module s'améliore, les entités qui ont été correctement classifiées vont continuer à l'être (donc la même valeur pour l'information en sortie du module), mais la qualité de l'information en sortie est meilleure. C'est-à-dire *la confiance* (le critère de qualité) dans le fonctionnement du module augmente.

Suite à cette discussion, la fonction de transfert de qualité $\mathcal{Q}_f$ doit être définie par rapport à

l'information et la qualité de information en entrée du module :

$$Q_{out} = \mathcal{Q}_f(I_{in}, Q_{in}) \tag{5.2}$$

La notion de qualité est par sa nature multidimensionnelle et donc, en conséquence, la fonction de transfert de qualité $\mathcal{Q}_f$ le sera également. Afin d'illustrer le comportement de cette fonction, il a été choisi l'exemple d'un module de traitement implémentant une classification d'une signature radar (cet exemple sera analysé en profondeur dans le chapitre 7). Ce module reçoit en entrée des données radar et fournit en sortie une information sur l'identité de la cible. Dans ce cas, nous supposons que le processus d'évaluation de la qualité locale caractérise la qualité des données en entrée par les critères de qualité : $\{Cr_{\text{Précision}}, Cr_{\text{Obsolescence}}, Cr_{\text{Quantité}}\}$ et la qualité de l'information en sortie du module par les critères de qualité $\{Cr_{\text{Confiance}}, Cr_{\text{Obsolescence}}\}$. Une partie de ces critères, $Cr_{\text{Obsolescence}}$, exprime les mêmes caractéristiques pour les données et l'information en entrée et en sortie du module, mais avec un sens différent, tandis que les autres caractéristiques qui ont été utilisés à l'entrée ne sont plus en sortie, $Cr_{\text{Quantité}}$. De plus, en fonction des caractéristiques de l'information à qualifier, ces critères peuvent être évalués par différentes mesures de qualité, μ^j_{Cr} qui représente la j-ème mesure de qualité employée pour le critère de qualité Cr. L'équation (5.3) illustre cet exemple et le rôle de la fonction de transfert de qualité de mappage d'une instance de qualité à une autre.

$$Q_{in} = \begin{pmatrix} Cr_{\text{Précision}} \\ Cr_{\text{Obsolescence}} \\ Cr_{\text{Quantité}} \end{pmatrix} \xrightarrow{\mathcal{Q}_f} Q_{out} = \begin{pmatrix} Cr_{\text{Confiance}} \\ Cr_{\text{Obsolescence}} \end{pmatrix} \tag{5.3}$$

En conséquence, la fonction de transfert de qualité est définie sur l'espace de critères de qualité en entrée du module et renvoie des valeurs dans l'espace de critères de qualité en sortie du module :

$$\mathcal{Q}_f : \text{QualCritères}_{\text{Entrée}} \rightarrow \text{QualCritères}_{\text{Sortie}} \tag{5.4}$$

Ainsi, afin d'obtenir la fonction de transfert de qualité pour un module de traitement, une analyse en deux étapes est nécessaire :

1. établir les relations entre les critères de qualité en entrée du module et celle en sortie ;
2. déduire la fonction de transfert de la qualité.

De plus, pour que la fonction de transfert de qualité soit déterminée il faut avoir une connaissance parfaite du module de traitement. Cette connaissance doit s'exprimer, dans un premier temps, par les critères de qualité utilisés en entrée et en sortie du module de traitement. L'utilisation des vecteurs binaires prenant des valeurs unitaires pour chaque critère utilisé et des valeurs nulles pour chaque critère inutilisé est un exemple de représentation de la connaissance sur les critères de qualité. En reprenant l'exemple représenté par l'équation (5.3) et sous l'hypothèse que les critères de qualité en entrée sont pris du tableau 4.1 et que ceux en sortie proviennent du tableau 4.2, la représentation vectorielle de l'équation (5.5) est donnée par :

$$\left\{ \begin{array}{ll} Q_{in} & = [1\,0\,0\,0\,1\,0\,1\,0\,0\,0\,0] \\ Q_{out} & = [0\,1\,0\,0\,0\,1\,0\,0\,0] \end{array} \right\} \tag{5.5}$$

Dans un deuxième temps, il faut avoir une connaissance sur le fonctionnement du module de traitement : la fonction réalisée, les paramètres internes, etc. (comme dans le cas de l'exemple du module de contrôle, l'équation 5.1).

Maintenant, la stratégie de développement de la fonction de transfert de la qualité d'un module de traitement quelconque est présentée. Parfois, il se peut que ce module ait un comportement simple et qu'une connaissance parfaite pour sa caractérisation soit disponible. Dans ce cas, il est possible d'arriver à une expression analytique de la fonction de transfert de qualité. Dans d'autres cas, le module de traitement a un comportement complexe ou partiellement connu, difficilement

exprimable par des formules mathématiques simples. Dans ce type de situations, l'évaluation de la qualité de l'information nécessite l'utilisation d'un nombre important de critères de qualité avec de possibles dépendances entre eux. Dans ce genre de situation, il est préférable de viser une méthode non-analytique d'estimation de la fonction de transfert de qualité.

Avant de commencer la présentation de ces deux techniques de développement de la fonction de transfert de qualité, l'hypothèse de départ est rappelée : les modules de traitement sont accessibles à l'étude, c'est-à-dire les entrées, les sorties et le fonctionnement générale du module sont supposés connus.

5.2 ÉVALUATION ANALYTIQUE DE LA FONCTION DE TRANSFERT DE QUALITÉ

Dans ce cas, la connaissance sur les critères de qualité utilisés en entrée et en sortie du module de traitement est supposée connue, car ceux-ci ont été analysés et déterminés lors de l'étape d'évaluation de la qualité locale du système d'information. Pourtant, il n'y a pas une relation entre les valeurs prise par les critères de qualité en entrée et les valeurs de critères de qualité en sortie du module de traitement. Il est rappelé que les valeurs de critères sont données par des mesures de qualité.

Afin d'illustrer la détermination analytique de la fonction de transfert de qualité, prenons un exemple simple d'un module de traitement implémentant une classification bayésienne en deux classes. Ce problème de classification peut également être vue comme un problème de détection indiquant la présence ou l'absence d'un signal.

Notons chaque observation à classifier par X et supposons qu'elle soit donnée par la relation suivante :

$$X = S + N \tag{5.6}$$

Dans cette dernière équation, S représente le signal utile (à détecter) qui, pour la simplicité de l'exemple, peut prendre les valeurs 0 ou A et N est un bruit blanc gaussien additif de moyenne nulle et variance unitaire. Ainsi, ce problème peut se formuler sous la forme d'un test binaire d'hypothèses :

$$\begin{array}{ll} \text{Hypothèse } H_1 : \text{présence signal} & : \ X = A + N \\ \text{Hypothèse } H_0 : \text{absence signal} & : \ X = N \end{array} \tag{5.7}$$

Suite à l'équation 5.6, l'observation à classifier, X est une variable aléatoire gaussienne. Comme toute variable aléatoire, elle est complètement caractérisée par sa densité de probabilité $f_{X|H_i}(x)$, conditionnée par l'hypothèse H_i, $i = 0, 1$. La théorie Bayésienne utilise comme critère de test d'hypothèse le *rapport de plausibilités* des probabilités conditionnelles [Van Trees 68] :

$$\Lambda(X) = \frac{f_{X|H_1(x)}}{f_{X|H_1(x)}} \underset{H_1}{\overset{H_0}{\lessgtr}} \eta \tag{5.8}$$

Ainsi, si le rapport de plausibilité, $\Lambda(X)$ est inférieure au seuil η l'hypothèse H_0 est validée (signal absent) et dans le cas contraire, $\Lambda(X) > \eta$ l'hypothèse H_1 est validée (signal détecté). Toujours dans le cadre de la théorie bayésienne, la valeur du seuil fait appel aux probabilités *a priori*, d'absence du signal, P_0 et de sa présence, P_1 et elle est déterminée suite à la minimisation du coût de prise de décisions sous les deux hypothèses [Van Trees 68] :

$$\eta = \frac{P_0(C_{10} - C_{00})}{P_1(C_{01} - C_{11})} \tag{5.9}$$

Dans cette dernière équation, C_{ij} représente le coût de prendre une décision en faveur de l'hypothèse i dans la situation quand j est la vraie, $(i, j) = \{0, 1\} \times \{0, 1\}$.

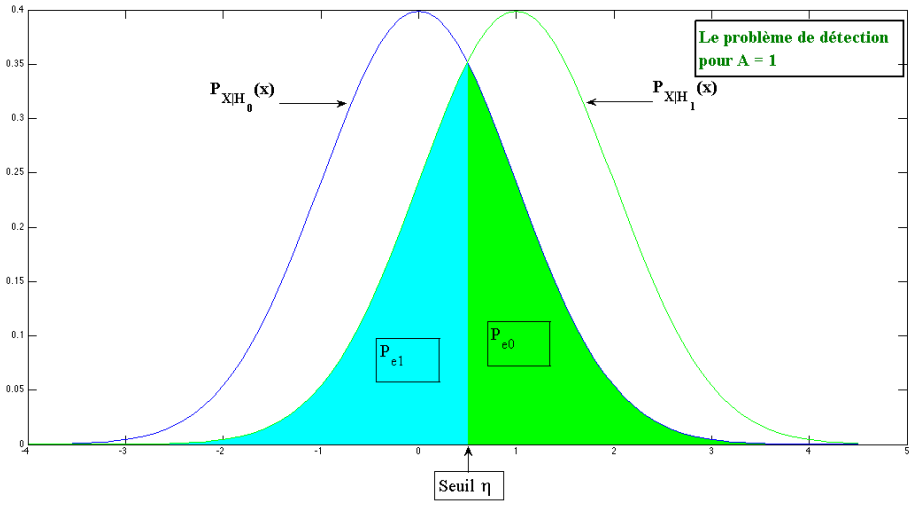


FIGURE 5.2: Exemple d'évaluation analytique de la fonction de transfert de qualité

En développant le rapport de plausibilité l'équation (5.8) se réécrit comme :

$$\Lambda(X) = \exp\left\{\frac{1}{2\sigma^2}A(2X - A)\right\} \underset{H_1}{\overset{H_0}{\lessgtr}} \eta \tag{5.10}$$

Ce problème de classification est illustré dans la figure 5.2. La classification de l'entrée est faite en comparant l'observation au seuil η : si la valeur est inférieure à ce seuil, l'observation appartient à la classe 0 (aucun signal détecté), sinon l'observation appartient à la classe 1 (signal détecté).

Dans le cas de cet exemple, la qualité en entrée du module de traitement est donnée par le critère de qualité *la précision*, mesurée par le niveau du rapport signal sur bruit, SNR [1]. La qualité en sortie du module de traitement est donnée *la confiance* de la détection, mesurée par la probabilité de bonne détection (ou par son inverse la probabilité d'erreur). La probabilité de bonne détection a, dans ce cas simple, la formule analytique suivante [Bisdikian 07] :

$$\begin{aligned} P_d &= Pr(\Lambda(X) \geq \eta | H_1) \\ &= \int_{\eta}^{+\infty} f_{\Lambda(X)|H_1}(u)du \\ &= 1 - \Phi\left(\frac{\sigma^2 \ln(\eta) - \frac{1}{2}A^2}{\sigma A}\right) \end{aligned} \tag{5.11}$$

dans cette équation Φ représente la distribution cumulative d'une variable aléatoire gaussienne standard, $\mathcal{N}(0, 1)$:

$$\Phi(u) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{u} \exp\left(-\frac{u^2}{2}\right) du \tag{5.12}$$

Le rapport $\psi = \frac{A^2}{\sigma^2}$ est une mesure du rapport signal sur bruit et donc l'équation (5.11) peut encore s'écrire sous la forme suivante :

$$P_d = 1 - \Phi\left(\frac{\ln(\eta)}{\sqrt{\psi}} - \frac{\sqrt{\psi}}{2}\right) \tag{5.13}$$

En conclusion, dans le cas de ce module une formule analytique de calcul de la qualité en sortie du module de traitement peut être déterminée en fonction de sa qualité en entrée (le rapport signal sur bruit) et en ayant une connaissance du comportement et du fonctionnement de ce module :

1. De l'anglais « Signal-to-Noise Ratio »

– il s'agit d'un détecteur bayésien ;
– la valeur du seuil η, qui est un paramètre interne du module.

Cet exemple présente une fonction de transfert de qualité unidimensionnelle, reliant la précision d'observations à la confiance de la détection. dans le cas multidimensionnel, l'expression générale de la relation entre la qualité en entrée et la qualité en sortie du module de traitement sera de la forme :

$$\overrightarrow{Q}_{out} = \mathcal{Q}_f\left(\overrightarrow{Q}_{in}, I_{in}\right) \tag{5.14}$$

Dans l'équation 5.14, les valeurs des critères de qualité en sortie du module sont exprimées en fonction des valeurs des critères de qualité en entrée du module et de la valeur prise par l'information en entrée du module. Cette dernière dépendance, de la quantité de l'information, doit nécessairement être prise en compte parce que le fonctionnement du module de traitement est dépendant d'elle et donc la qualité le sera également.

5

5.3 ÉVALUATION NON-ANALYTIQUE DE LA FONCTION DE TRANSFERT DE QUALITÉ

Il y a des cas où il est impossible d'exprimer la fonction de transfert de qualité sous une forme analytique. C'est la situation où la qualité en entrée et en sortie du module de traitement s'exprime par un nombre important de critères de qualité, avec des dépendances non-évidentes entre eux. Une autre situation est quand nous avons une connaissance partielle du comportement et/ou du fonctionnement du module de traitement. Cette dernière situation se rencontre lorsque le module de traitement est complexe d'un point de vue des traitements numériques (exemple : un réseau de neurones) ou quand il n'est pas possible d'avoir accès à l'intérieur du module de traitement pour voir les paramètres de fonctionnement, la façon dont il a été développé, etc. Dans ce dernier cas, le module de traitement est vu plus ou moins comme une boîte noire offrant accès seulement aux entrées et aux sorties de ce module. Dans tous les cas, nous faisons la supposition qu'au moins la connaissance des caractéristiques des entrées et des sorties du module et le comportement générale de fonctionnement du module peuvent être obtenus. En utilisant ces connaissances, incomplètes, le comportement vis-à-vis de l'influence du module sur la qualité peut alors être analysé en faisant varier le couple d'entrée l'information et la qualité de l'information, $\overrightarrow{Q}_{in}, I_{in}$ et en observant (mesurant) la qualité correspondante en sa sortie. Cette stratégie d'analyse du module de traitement permet d'obtenir des couples de qualité de type $\overrightarrow{Q}_{in}, \overrightarrow{Q}_{out}$ pour chaque module du système d'informations. Par la suite, ces couples vont permettre de déterminer des relations entre la qualité en entrée du module et la qualité en sortie du module en utilisant des méthodes statistiques comme par exemple la régression linéaire ou non-linéaire.

Maintenant, un exemple d'estimation de la fonction de transfert de qualité $\mathcal{Q}_f$ sera présenté, en faisant appel à cette méthode non-analytique, c'est-à-dire en mesurant les paires $\overrightarrow{Q}_{in}, \overrightarrow{Q}_{out}$. Pour la simplicité de l'exposition, un cas unidimensionnel d'évaluation de la qualité sera considéré, en prenant le même exemple que celui du cas d'évaluation analytique de la $\mathcal{Q}_f$, paragraphe 5.2. Dans ce cas, il est supposé que la connaissance sur le module de traitement est :

– le fonctionnement générique du module : il s'agit d'un détecteur bayésien ;
– le signal en entrée du module de traitement est unidimensionnel et susceptible d'être affecté par un bruit blanc gaussien additif de moyenne nulle ;
– la sortie du module est la classe estimée du signal (présence ou absence).

Sans avoir la connaissance de la valeur du seuil η, paramètre interne du module de traitement, la détermination analytique de la fonction de transfert de qualité $\mathcal{Q}_f$ n'est pas possible (équation (5.13)). Comme dans le cas traité dans le paragraphe 5.2, la valeur de la qualité en entrée est représentée par le niveau du rapport signal sur bruit, ψ et la qualité en sortie par la probabilité de bonne détection, P_d.

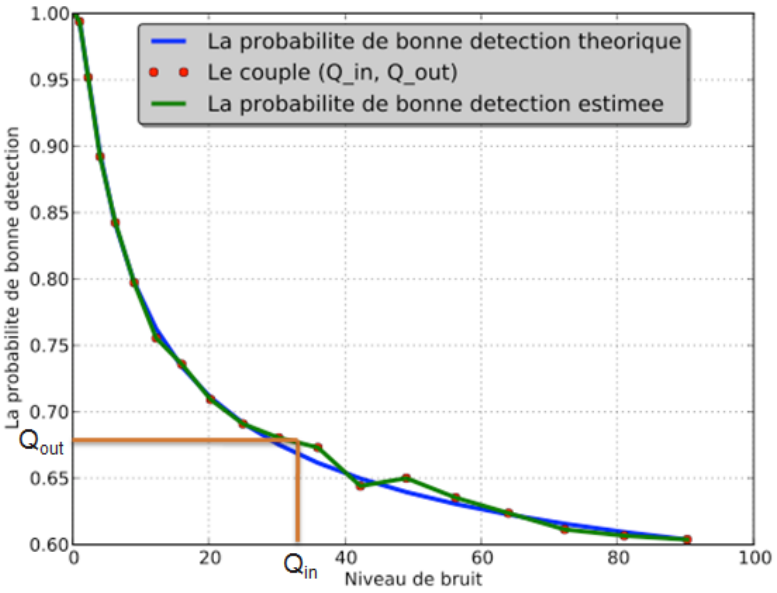


FIGURE 5.3: L'évaluation non-analytique de la fonction de transfert de qualité

L'évaluation de la qualité en sortie a été réalisée par une simulation Monte-Carlo consistant dans la génération de 10000 échantillons pris à partir d'une loi Bernoulli ayant la probabilité de succès $p = 0.5$. La valeur du paramètre du signal utile, A a été fixée à 5 et le bruit $N \sim \mathcal{N}(0, \sigma^2)$ avec σ^2 variant de 0.25 à 100. Dans la figure 5.3 se trouvent les résultats de cette simulation. Les points rouges représentent les estimations de la probabilité de bonne détection en fonction du niveau de bruit. La courbe en bleu représente la probabilité de bonne détection théorique. De petits biais peuvent s'observer entre les valeurs estimées par la simulation et les valeurs théoriques de $\mathcal{Q}_f$. Avec la fonction de transfert de qualité ainsi déterminée, pour chaque nouvelle qualité en entrée Q_{in}, la qualité en sortie Q_{out} peut être directement évaluée en utilisant une interpolation entre les deux points les plus proches de la fonction $\mathcal{Q}_f$, comme illustré dans la figure 5.3.

Cet exemple présente l'estimation de la fonction de transfert de qualité pour le cas unidimensionnel. Pour le cas multidimensionnel, il est proposé d'évaluer indépendamment chaque critère de la qualité en sortie du module, Cr_{out}, par rapport à tous les critères de qualité en entrée, Cr_{in}. L'algorithme d'estimation pour ce cas général est présenté sous la forme de pseudo-code, algorithme 1.

Cette évaluation individuelle de critères de qualité en sortie est justifiée en raison des possibles dépendance entre les critères de qualité. Ainsi, la fonction de transfert de qualité pour un module de traitement sera représentée par M fonctions, avec M le nombre de critères de qualité en sortie du module. Chacune de ces M fonctions sera représentée par une surface N-dimensionnelle, avec N le nombre de dimensions de qualité en entrée.

Un exemple concret de détermination d'une fonction de transfert de qualité multidimensionnelle sera présenté dans le chapitre 7.

5.4 CONCLUSION

Dans ce chapitre, nous avons introduit un nouveau concept : la fonction de transfert de qualité. L'objectif de cette fonction est de modéliser l'influence du module de traitement sur la qualité.

Entrées : Critères de qualité en entrée Cr_{in}
Critères de qualité en sortie Cr_{out}
Sorties : La fonction de transfert de qualité $\mathcal{Q}_f$ pour chaque Cr_{out}

N = nombre(Cr_{in}) ;
M = nombre(Cr_{out}) ;
pour *i = 1 à M* **faire**
  pour *j = 1 à N* **faire**
    varie les valeurs de Cr^j_{in} ;
    enregistre Cr^i_{out} ;
  fin
  $\mathcal{Q}_f(Cr^i_{out})$ = interpolation de Cr^i_{out}
fin

ALGORITHME 1: L'algorithme d'évaluation de la fonction de transfert de qualité d'un module de traitement

Ainsi, elle réalise la liaison entre la qualité en entrée du module et la qualité en sortie. L'importance de cette fonction est de mettre à jour automatiquement la qualité en sortie d'un module une fois qu'un changement d'information ou de qualité d'information est identifié à son entrée.

Comme la fonction de transfert de qualité a été définie dans un cadre générique, pouvant modéliser tout module de traitement, sa construction peut poser des problèmes. Pour cela nous avons identifié deux situations. La première est lorsque le comportement et le fonctionnement du module de traitement sont complètement connus et que le nombre de critères de qualité est réduit, par exemple inférieur à 5. Dans cette situation, il est envisageable de déterminer une expression analytique pour la fonction de transfert de qualité.

La deuxième situation est lorsque les connaissances sur le comportement et le fonctionnement du module de traitement sont incomplètes ou lorsque le nombre de critères de qualité est trop important. Dans ce cas, il est très difficile à déterminer une formule analytique pour la fonction de transfert de qualité. Ainsi, nous avons proposé un processus d'estimation de celle-ci sous l'hypothèse d'avoir au moins accès aux entrées et aux sorties du module de traitement. Cette estimation suppose de faire varier l'information et la qualité de l'information de l'entrée et d'enregistrer les valeurs des critères de qualité en sortie. Grâce à cette technique, il est possible d'enregistrer des couples $((Q_{in}, I_{in}), Q_{out})$ qui, suite à une interpolation, peuvent être considérés comme une bonne estimation de la fonction de transfert de qualité.

Ainsi, il est possible de relier la qualité en sortie du module à celle de son entrée. Dans le chapitre suivant nous allons corroborer cet outil avec le processus d'évaluation de la qualité locale afin d'évaluer d'une manière automatique la qualité de l'information en sortie du système d'information.

6 Évaluation de la qualité globale de l'information

Dans ce chapitre, nous présentons la stratégie d'utilisation de l'évaluation de la qualité locale et de la fonction de transfert de qualité, dans l'évaluation de la qualité globale, c'est-à-dire la qualité du système d'information. Jusqu'à présent, la procédure d'évaluation de la qualité locale, chapitre 4, ainsi que la manière de quantifier l'influence d'un module de traitement sur la qualité, chapitre 5, ont été définies.

Les études précédentes ont été faites en considérant un seul module de traitement. Maintenant, l'objectif est de changer d'échelle et de passer à l'analyse du système d'information entier. La première étape est l'analyse de deux modules de traitements successifs. Puis, l'étude sera étendue au système d'information entier et finalement la qualité globale sera évaluée. Le passage de la qualité locale à la qualité globale est présenté dans le paragraphe 6.1. Puis, dans le paragraphe 6.2, il est proposé d'agréger les valeurs de qualité afin de l'exprimer dans une seule dimension.

6.1 Passage de la qualité locale à la qualité globale

Dans la figure 6.1 sont présentés deux modules de traitement successifs qui doivent être concaténés. Afin que ces deux modules de traitement puissent être connectés ensemble, l'information en sortie du premier module de traitement doit être adaptée à l'entrée du deuxième module de traitement, cf. paragraphe 4.1. Ce besoin d'adaptation est aussi applicable au domaine de la qualité.

Ainsi, une première condition pour la qualité est formulée par le principe suivant :

> **Principe 1 :** Dans le cas de deux modules de traitement de l'information successifs, la qualité en sortie du premier, Q^i_{out} détermine la qualité en entrée du deuxième module, Q^{i+1}_{in} par rapport aux critères de qualité, c'est-à-dire $Q^{i+1}_{in} \subseteq Q^i_{out}$.

L'explication de ce principe est immédiate. En effet, les caractéristiques (valeurs et sémantiques) de l'information en sortie du premier module sont identiques aux caractéristiques de l'information en entrée du module suivant. En conséquence, les mêmes critères de qualité seront employés. Cependant, il est possible que certains critères ne soient pas disponibles pour le deuxième module.

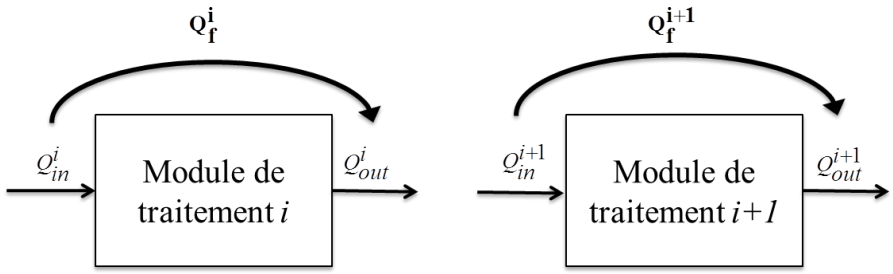


Figure 6.1: La concaténation de deux modules de traitement

Ainsi, dans l'énoncé de ce principe l'opérateur inclusion-égalité a été utilisé, au lieu d'une simple égalité.

Maintenant, nous allons faire la transition vers l'évaluation de l'information en sortie du système d'information. Le point de départ est l'observation qu'une variation de la qualité en entrée d'un module de traitement est directement responsable d'une variation de la qualité en sortie de ce même module, à l'aide de la fonction de transfert de qualité. En se basant sur ce raisonnement, dans [Todoran 14c] nous avons énoncé *un principe de propagation des variations de la qualité* :

> **Principe 2 :** Les variations locales de la qualité des données et de l'information se propagent à travers le système jusqu'à sa sortie.

Ce principe est illustré dans la figure 6.2. Cette architecture d'un système d'information modulaire contient deux flux principaux d'information. Supposons qu'un changement de qualité apparaisse après le premier module de la branche supérieure, Module i. Par application du deuxième principe, ce changement de qualité va se propager, en cascade, en aval de ce module, après chaque module de traitement, pour finalement arriver à la sortie du système d'information. Cette propagation du changement de la qualité est représentée par la ligne rouge pointillée dans la figure 6.2.

Dans la présentation des objectifs attendus de cette méthodologie d'évaluation de la qualité de l'information, il a été également fait l'hypothèse qu'elle devrait être capable de prendre en compte les évolutions éventuelles du système d'information : la mise à jour des modules de traitement, l'ajout d'un nouveau module de traitement, etc. Toujours à la figure 6.2, la situation du remplacement du Module i+1, par un autre est illustrée, en respectant les conditions présentées au début du paragraphe 4.2. Dans ce cas, la seule procédure à suivre est l'estimation de la fonction de transfert de qualité du nouvel module, $\mathcal{Q}_{f'}^{i+1}$. Après son estimation, la qualité en sortie de ce module sera automatiquement mise à jour, ainsi que les autres qualités de l'information en aval de ce module, marquées par une flèche rouge. Ainsi, il n'est pas nécessaire d'évaluer à nouveau le système dans sa globalité, il suffit d'estimer une seule fonction de transfert de qualité et de laisser le reste du processus inchangé.

Il est intéressant d'analyser aussi la procédure à suivre afin de détecter les éventuels changements de qualité dans le système. Pour cela, une hypothèse réaliste est de considérer que les données en entrée du système (et donc, implicitement, également leur qualité) évoluent beaucoup plus vite que les paramètres internes du système (c'est-à-dire les mises à jour des modules, les ajouts/suppressions de modules, etc.). Ainsi, dans [Todoran 13], nous avons proposé d'utiliser une sonde mesurant la qualité à des intervalles de temps réguliers, déterminés par les spécificités de l'application. Si la base de données ou le capteur sont considérés comme un module, la qualité à la sortie de ce module sera décrite par une liste de critères de qualité $Cr_i, i \in 1, \ldots, n$, avec les valeurs données par des mesures de qualité $m_i^j, i \in 1, \ldots, nj \in 1, \ldots, N_i$, comme présenté sur la figure 6.3.

En conclusion, la qualité locale conjointement avec le concept de fonction de transfert de qualité permettent d'évaluer automatiquement [1] la qualité de l'information en sortie d'un système d'information.

6.2 Évaluation du score de qualité totale de l'information

Pour chaque niveau de traitement du système d'informations, plusieurs critères de qualité sont utilisés afin d'évaluer la qualité. Dans certaines situations, il est préférable d'exprimer le niveau de cette qualité dans une seule dimension, c'est-à-dire d'évaluer la qualité par une seule valeur. C'est le cas quand l'utilisateur final veut avoir une idée de l'évolution de la qualité à travers les différents traitements subis par les données et par les informations dans le système d'informations. Dans cette situation, l'utilisation d'une seule valeur pour caractériser la qualité permet d'avoir une

1. Une fois la qualité locale et les fonctions de transfert de qualité déterminées pour chaque module de traitement

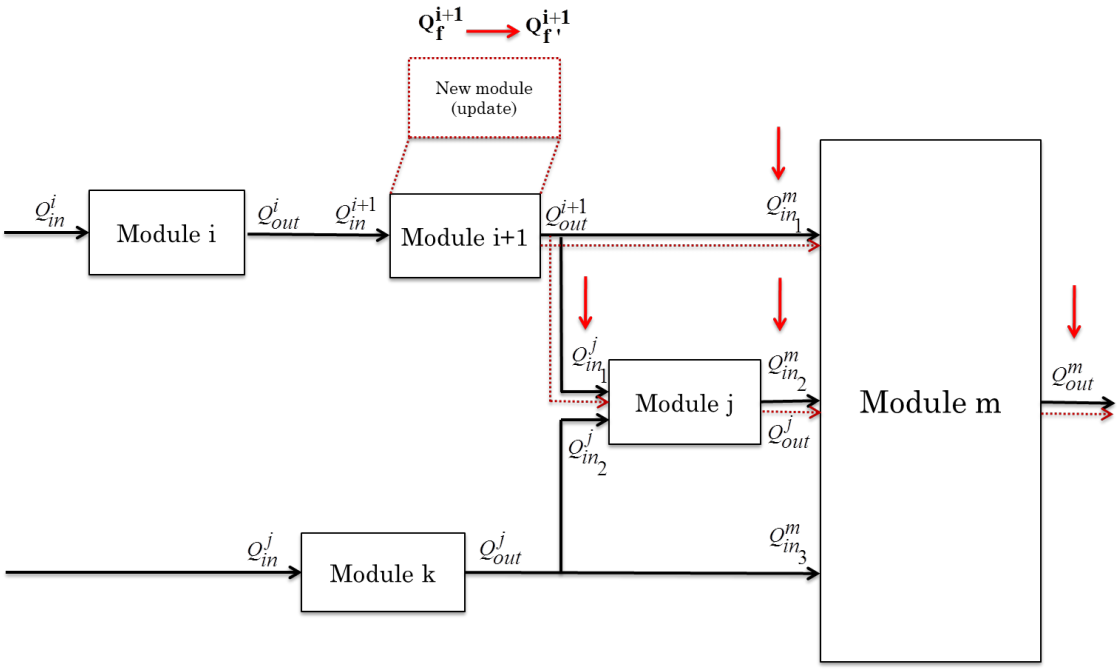


FIGURE 6.2: Évaluation dynamique de la qualité d'information

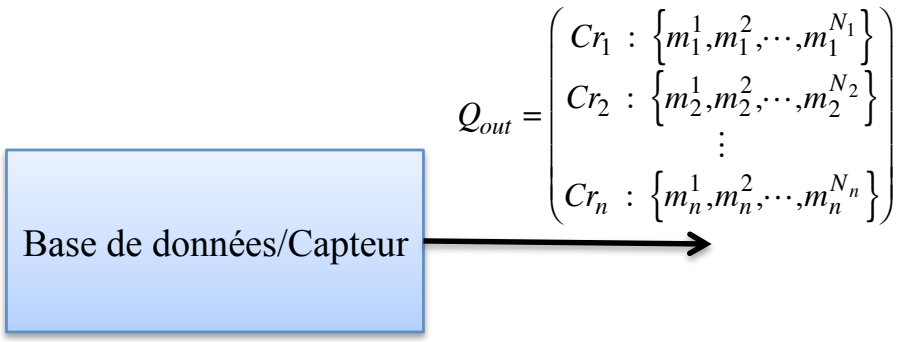


FIGURE 6.3: La qualité d'une base de données ou d'un capteur mesurée par une sonde

vision d'ensemble synthétique, qui ne nécessite pas l'analyse par l'utilisateur de tous les critères de qualité, opération difficile, qui demande beaucoup de temps de la part de l'utilisateur. Cette vision unidimensionnelle de la qualité est généralement préférée par les décideurs. En fait, les décideurs ont besoin d'une vision de la qualité plus détaillée seulement dans les situations où l'information proposée est en contradiction avec leurs propres croyances, attentes, etc. Dans ce cas, la méthodologie présentée dans ce mémoire de thèse, permet d'étendre la vision de la qualité et de lui proposer la traçabilité de la qualité à travers le système [Todoran 14a] :

« *La traçabilité permet au décideur de visualiser quelles sources d'information soutiennent l'information qui lui est proposée et comment sa qualité a évolué à travers le système d'information.* »

Le problème est que les différents critères de qualité peuvent être évalués en utilisant des mesures de qualité noncommensurables, c'est-à-dire des mesures qui n'ont pas le même domaine de définition, la même importance ou la même signification. Ainsi, dans le paragraphe 4.2, nous avons montré qu'un critère de qualité pourrait être mesuré par une mesure floue ou par une probabilité. De plus, il existe des dépendances entre les critères de qualité qui devront être prises en compte.

Par conséquent, le processus d'agrégation de ces critères de qualité n'est pas si trivial. Néanmoins, l'existence de multiple critères caractérisant la même entité suggère de s'orienter vers le domaine d'analyse multi-critères [Todoran 14c]. Ceci est un outil mathématique très utilisé dans le domaine d'aide à la décision. Il a pour objectif la comparaison d'alternative, décrites par un groupe de critères. Dans notre cas d'analyse, la principale différence, par rapport à l'analyse multi-critères traditionnelle, est que l'analyse sera faite non seulement à la sortie du système, mais aussi après chaque module de traitement.

Le problème général de construction d'un opérateur d'agrégation de toutes les dimensions de la qualité peut être décrit par l'équation suivante :

$$Q_{\text{totale}} = \mathcal{H}_n(\text{Cr}_1(q_1^1, q_1^2, \ldots), \ldots, \text{Cr}_n(q_n^1, q_n^2, \ldots)) \tag{6.1}$$

Dans cette équation, Cr_i représente le $i$ème critère de qualité et q_i^j sa $j$ème mesure de qualité (car un critère peut être évalué par plusieurs mesures de qualité, paragraphe 4.2). En analysant plus en détail l'équation 6.1, on peut observer que l'agrégation doit se faire à deux niveaux :

- **agrégation intra-critère :** correspond à la combinaison de mesures de qualité du même critère de qualité ;
- **agrégation inter-critère :** correspond à la combinaison de critères de qualité, critères qui sont exprimés par une seule valeur issue de l'agrégation intra-critère.

Comme indiqué dans toute étude sur les opérateurs d'agrégations (voir par exemple [Calvo 02]), la restriction des valeurs en entrée et en sortie à un intervalle fixé (une échelle) $I = [a, b] \subseteq (-\infty, +\infty)$ n'affecte pas la généralité et les propriétés de l'opérateur. En fait, il s'agit d'un problème de passage d'une échelle à une autre. Ainsi, sans perdre en généralité, le cas de l'intervalle unitaire $I = [0, 1]$ sera considéré.

Avant de proposer des solutions à ce problème de construction d'opérateurs d'agrégation adaptés pour l'évaluation du score de qualité totale, dans le paragraphe suivant, une courte présentation des différents types d'opérateurs rencontrés dans la littérature est réalisée.

6.2.1 Vers l'agrégation de mesures de qualité

Dans beaucoup de situations, il est nécessaire de combiner (agréger) plusieurs mesures qui, dans le cas le plus général, ne sont pas forcement commensurables. Pour cela, dans ce paragraphe les principales classes d'opérateurs d'agrégation seront présentées, en mettant l'accent sur leurs propriétés.

Cette présentation d'opérateurs d'agrégation sera faite dans le cadre du domaine de l'aide à la décision multi-critères (MCDA [2]).

2. De l'anglais « Multi-criteria decision aid »

Nous notons :
- X l'ensemble d'alternatives (de critères de qualité dans notre cas) ;
- $\mathcal{H}$ l'opérateur d'agrégation ;
- $x \succsim y$: x est préféré à y
- u la fonction d'utilité, ayant comme propriété $x \succsim y \Rightarrow u(x) > u(y)$;
- $x_{(i)}$ indique que les indices ont été permutés pour respecter : $0 \leq f(x_{(1)}) \leq \ldots \leq f(x_{(n)}) \leq 1$ (avec f une fonction continue strictement monotone) ;
- $A_{(i)} \triangleq \{x_{(i)}, \ldots, x_{(n)}\}$.

Selon [Grabisch 95], la seule propriété nécessaire pour un opérateur d'agrégation est celle de **monotonicité** :

$$x \succsim y \Rightarrow \mathcal{H}(u_1(x_1), \ldots, u_n(x_n)) > \mathcal{H}(u_1(y_1), \ldots, u_n(y_n)) \tag{6.2}$$

6.2.1.1 Opérateurs d'agrégation communs

Dans ce paragraphe, les principales catégories classiques d'opérateurs d'agrégation seront présentées :

1. **Opérateurs basés sur la moyenne** :
   Ce type d'opérateurs d'agrégation se situent entre les opérateurs *min* et *max*. Ci-dessous sont présentés quelques exemples d'opérateurs d'agrégation basés sur la moyenne :

   • *la moyenne arithmétique* :
   $$\frac{1}{n}\sum_{i=1}^{n} a_i \tag{6.3}$$

   • *la moyenne géométrique* :
   $$\prod_{i=1}^{n} a_i^{\frac{1}{n}} \tag{6.4}$$

   • *la famille de Dyckhoff-Pedrycz* :
   $$\left(\frac{1}{n}\sum_{i=1}^{n} a_i^{\alpha}\right)^{\frac{1}{\alpha}} \tag{6.5}$$

   • *la moyenne quasi-arithmétique* (qui couvre tous les autres opérateurs) :
   $$f^{-1}\left[\frac{1}{n}\sum_{i=1}^{n} f(a_i)\right] \tag{6.6}$$

   avec f une fonction continue strictement monotone.

2. **La médiane** :
   L'opérateur médiane donne comme résultat la valeur du milieu d'une séquence ordonnée :
   $$med(a_{(1)}, \ldots, a_{(2q-1)}) \triangleq a_{(q)} \tag{6.7}$$
   La médiane est le seul opérateur de la famille des moyennes qui possède la propriété d'associativité [Grabisch 95].

3. **Opérateurs compensatoires**
   Si l'opérateur d'agrégation est regardé d'un point de vue de l'utilisateur humain, il peut s'observer qu'il est compensatoire [Zimmermann 80]. L'effet compensatoire signifie qu'un critère de très bonne qualité peut compenser un autre d'une qualité moins bonne. La famille d'opérateurs moyenne quasi-arithmétique, l'équation (6.6), et les opérateurs algébriques basés sur la disjonction « OU » (par exemple le max) sont des exemples d'opérateurs complètement compensatoires. De l'autre côté, les opérateurs basés sur la conjonction « ET » sont non-compensatoires. [Zimmermann 80] a montré que ni les opérateurs complètement compensatoires, ni les opérateurs non-compensatoires sont adaptés pour modéliser l'agrégation des informations issues des experts. De plus, dans cette même étude il a été montré que la moyenne arithmétique donne des résultats biaisés parce qu'elle ne prend pas en compte les interactions possibles entre les différents critères. Pour prendre en compte l'effet compensatoire, [Zimmermann 80] propose une nouvelle famille d'opérateurs, ayant un comportement se situant entre les opérateurs non-compensatoires et ceux complètement compensatoires. Cette famille d'opérateurs a été appelée $\gamma-$opérateurs :

$$\mathcal{H}(a_1,\ldots,a_n) = \left(\prod_{i=1}^{n} a_i\right)^{1-\gamma} \left(\bigoplus_{i=1}^{n} a_i\right)^{\gamma} \tag{6.8}$$

   où l'opérateur $\oplus$ désigne la somme probabiliste :

$$\bigoplus_{i=1}^{n} a_i \triangleq 1 - \prod_{i=1}^{n}(1-a_i) \tag{6.9}$$

   Dans [Grabisch 95], d'autres exemples d'opérateurs d'agrégation compensatoires sont présentés. Un trait commun de tous ces opérateurs d'agrégation est qu'ils sont intuitifs, mais comme ils sont construits d'une manière ad-hoc, leurs propriétés exactes ne sont pas connues de manière précise (sauf la compensation) [Grabisch 95].

4. **Opérateurs avec pondération**
   Dans beaucoup de cas, les mesures à agréger n'ont pas la même importance et donc une pondération est nécessaire. Par la suite avec w_i sera noté le poids associé à la $i-$ème mesure. De plus, il sera considéré que les poids sont normalisés, c'est-à-dire $\sum_{i=1}^{n} w_i = 1$. Dans la théorie des possibilités, voir annexe A.2, il existe des exemples d'agrégations avec pondération en utilisant les opérateurs *min* et *max* :

$$W_{w_1,\ldots,w_n}(a_1,\ldots,a_n) = \min_i\{\max\{(1-w_i), a_i\}\} \tag{6.10}$$

$$W_{w_1,\ldots,w_n}(a_1,\ldots,a_n) = \max_i\{\min\{(w_i, a_i)\}\} \tag{6.11}$$

   La famille d'opérateurs d'agrégation quasi-arithmétiques peut être transformée afin de prendre en compte la pondération [Grabisch 95] :

$$M^f_{w_1,\ldots,w_n}(a_1,\ldots,a_n) = f^{-1}\left[\sum_{i=1}^{n} w_i f(a_i)\right] \tag{6.12}$$

   Une autre famille importante d'opérateurs d'agrégation avec pondération est celle proposée dans [Yager 91], qui caractérise des opérateurs de moyenne pondérée ordonnée (OWA [3]). Dans le cas le plus simple, OWA est une moyenne pondérée :

$$OWA_{w_1,\ldots,w_n}(a_1,\ldots,a_n) = \sum_{i=1}^{n} w_i a_{(i)} \tag{6.13}$$

3. De l'anglais "Ordered Weighted Averaging"

Comme dans le cas de la moyenne quasi-arithmétique pondérée, équation 6.12, l'opérateur OWA peut être généralisé :

$$OWA_{w_1,\ldots,w_n}(a_1,\ldots,a_n) = f^{-1}\left[\sum_{i=1}^{n} w_i f(a_{(i)})\right] \tag{6.14}$$

En analysant les méthodes d'agrégation proposées dans ce paragraphe, il peut s'observer qu'elles ont des limitations. Ainsi, par exemple, il existe des méthodes simples à interpréter, comme la moyenne pondérée ou l'OWA, mais avec l'inconvénient d'être restrictives (par exemple elles ne peuvent pas prendre en compte les interactions entre les mesures à agréger). D'autres méthodes sont plus flexibles mais avec l'inconvénient d'être difficile à interpréter, comme les opérateurs avec compensation. Une autre limitation est la difficulté à prendre en compte les éventuelles dépendances entre les mesures à agréger. Une solution possible à ces limitations est l'intégrale floue qui sera introduite dans la section suivante.

6.2.1.2 L'intégrale floue

L'intégrale floue sera considérée comme un opérateur sur $[0,1]^n$ et donc elle sera appliquée pour le cas de fonctions qui prennent les valeurs dans l'intervalle unitaire.

Quelques exemples d'intégrales floues seront maintenant présentées. Pour la simplicité seulement le cas discret sera considéré, avec un ensemble d'alternatives fini $X = \{x_1,\ldots,x_n\}$, qui peut être vu comme un ensemble de critères, d'attributs, d'experts, de capteurs, etc.

Définition 18 : Soit un espace mesurable flou $(X, \mathcal{P}(X), \mu)$. L'intégrale de Sugeno d'une fonction $f : X \to [0,1]$, par rapport à la mesure floue μ, est définie par :

$$\mathcal{S}_\mu(f(x_1),\ldots,f(x_n)) \triangleq \max_i\{\min(f(x_{(i)}), \mu(A_{(i)})\} \tag{6.15}$$

Définition 19 : Soit un espace mesurable flou $(X, \mathcal{P}(X), \mu)$. La quasi-intégrale de Sugeno d'une fonction $f : X \to [0,1]$, par rapport à la mesure floue μ, est définie par :

$$\mathcal{S}^{\mathsf{T}}_\mu(f(x_1),\ldots,f(x_n)) \triangleq \max_i\{f(x_{(i)}) \,\mathsf{T}\, \mu(A_{(i)})\} \tag{6.16}$$

avec T représentant une t-norme, c'est-à-dire $\mathsf{T}(x,y) \leq \min(x,y)$.

Définition 20 : Soit un espace mesurable flou $(X, \mathcal{P}(X), \mu)$. L'intégrale de Choquet d'une fonction $f : X \to [0,1]$, par rapport à la mesure floue μ est définie par :

$$\mathcal{C}_\mu(f(x_1),\ldots,f(x_n)) \triangleq \sum_{i=1}^{n} \left((f(x_{(i)} - f(x_{(i-1)})\right) \mu(A_{(i)}) \tag{6.17}$$

avec $f(x_{(0)}) = 0$.

Observations sur l'intégrale de Choquet [Grabisch 00] :

– si la mesure floue μ est additive, l'intégrale de Choquet reviens à une somme pondérée :

$$\mathcal{C}_\mu(f) = \sum_{i=1}^{n} \mu(x_i) f_i \tag{6.18}$$

– si la mesure floue μ est symétrique, l'intégrale de Choquet reviens à un opérateur OWA :

$$\mathcal{C}_\mu(f) = \sum_{i=1}^{n} (\mu_{n-i+1} - \mu_{n-i}) f_{(i)} \tag{6.19}$$

avec $\mu_i \triangleq \mu(A)$ et avec $|A| = i$.

– une intégrale de Choquet commutative reviens à un opérateur OWA.

Dans les définitions précédentes la mesure floue μ doit être vue comme une représentation des poids (mesure d'importance, de confiance, etc.) soit pour des critères individuels, soit pour un ensemble de critères [Grabisch 95]. Grâce à cela, les intégrales floues ont la possibilité de prendre en compte les éventuelles interactions entre les critères.

Une possibilité pour la quantification de l'interaction entre les différents critères est la valeur (l'indice) de Shapley [Grabisch 97], utilisée pour la première fois dans la théorie des jeux coopératifs. Pour tout élément $x_i \in X$ la valeur de Shapley de l'élément x_i est donnée par :

$$\phi_i \triangleq \sum_{K \subset X \setminus \{x_i\}} \frac{(n - |K| - 1)!|K|!}{n!} [\mu(K \cup \{x_i\}) - \mu(K)] \tag{6.20}$$

Ainsi, la valeur de Shapley est un vecteur ϕ_i, $i = \overline{1,n}$. En considérant que la contribution de l'élément $\{x_i\}$ dans la coalition K est donnée par $\mu(K \cup \{x_i\}) - \mu(K)$, la valeur de Shapley ϕ_i exprime la contribution moyenne de l'élément $\{x_i\}$ dans toutes les coalitions, la moyenne étant pondérée par un coefficient dépendant du cardinal de la coalition [Grabisch 00]. Dans la pratique la valeur de Shapley peut être utilisée comme une mesure de l'importance de chaque source d'information [Gader 04].

Quelques propriétés de la valeur de Shapley :

– $\sum_{i=1}^{n} \phi_i = \mu(X) = 1$

– si la mesure μ est additive : $\phi_i = \mu(x_i)$ $(\forall) x_i \in X$

– si la mesure μ est dépendante seulement du cardinal de l'ensemble : $\phi_i = \mu(x_i) = \frac{1}{n}$ $(\forall) x_i \in X$

En gardant le même raisonnement que pour la définition de la valeur de Shapley, l'interaction entre deux éléments x_i et x_j peut s'exprimer par :

$$I_{i,j} \triangleq \sum_{K \subset X \setminus \{x_i, x_j\}} \frac{(n - |K| - 2)!|K|!}{(n-1)!} [\mu(K \cup \{x_i, x_j\}) - \mu(K \cup \{x_i\}) - \mu(K \cup \{x_j\}) + \mu(K)] \tag{6.21}$$

En fonction de la coopération entre ces deux éléments, la quantité $\mu(\{x_i, x_j\}) - \mu(x_i) - \mu(x_j)$ peut prendre des valeurs [Grabisch 00] :

– Positives : exprimant une coopération productive, car les éléments x_i et x_j pris ensemble donne une information plus riche que les deux éléments pris individuellement. Dans ce cas, les deux éléments sont complémentaires.

– Négatives : la coopération n'est pas productive, car les éléments x_i et x_j pris ensemble donne une information moins riche que les deux éléments pris individuellement. Dans le cas extrême la prise en compte d'un seul élément est suffisante, l'autre étant redondant. Dans cette situation, les éléments sont substitutifs.

– Nulle : il n'y a pas de coopération, les deux éléments étant indépendants.

Dans [Grabisch 97] a été proposée une généralisation de l'interaction pour toutes les coalitions :

$$I(A) \triangleq \sum_{K \subset X \setminus A} \left(\frac{(n - |K| - |A|)!|K|!}{(n - |A| + 1)!} \times \sum_{B \subset A} (-1)^{|A|-|B|} \mu(K \cup B) \right), \quad \forall A \subset N \tag{6.22}$$

Dans la construction de l'intégrale de Choquet, il est nécessaire de définir $2^n - 2$ coefficients. Dans la pratique ce nombre peut être grand et donc la construction de l'intégrale devient complexe. Pour résoudre ce problème, une solution est de considérer un nombre limité à k interactions possibles entre les critères et donc de réduire le nombre de coefficients à $\sum_{i=1}^{k} C_i^n$.

> **Définition 21 :** Une mesure floue est appelée mesure k-additive si $I(A) = 0$ pour toutes les coalitions contenant plus de k éléments et s'il existe au moins une coalition A, contenant exactement k éléments et pour laquelle $I(A) \neq 0$ [Grabisch 00].

Dans [Warren 99], il a été proposée une méthodologie pour le calcul de l'intégrale de Choquet en quatre étapes en supposant que la formule de calcul de l'intégrale est donnée par l'équation (6.19) :

1. Pour un ensemble donné de valeurs d'importance de coefficients (des poids) ω, déterminer une constante non-additive $\lambda : [-1, +\infty]$ solution de l'équation de Sugeno :

$$\lambda + 1 = \prod_{i=1}^{n} (1 + \omega_i \lambda) \tag{6.23}$$

   Les valeurs des poids ω_i peuvent être obtenues par des techniques d'apprentissage automatique ou par des connaissances *a priori* sur la confiance dans les sources d'information [Schuck 10].
2. Ordonner (décroissant) les valeurs $f(x_i)$.
3. Pour les coefficients ordonnés, calculer les poids de sous-ensembles $\mu(A_i)$ en commençant avec la plus grande valeur :

$$\mu(A_i) = \mu(A_{i-1}) + \omega_i + \lambda \omega_i \mu(A_{i-1}) \tag{6.24}$$

   $\mu(A_i)$ représentent des poids associés aux sous-ensembles monotones, $\mu(A_0) = 0$
4. Utiliser l'équation (6.19) pour la construction de l'intégrale.

Dans l'équation (6.24) une valeur positive du paramètre λ exprime un effet super-additif, c'est-à-dire en augmentant le poids du sous-ensemble. Cet effet peut être utile dans l'utilisation d'une synergie entre les valeurs individuelles de la fonction $f(x_i)$ [Warren 99]. Ainsi, les valeurs proches entre elles vont augmenter la valeur finale de l'agrégation et les valeurs divergentes vont la diminuer. Il faut aussi remarquer que les valeurs proches entre elles traduisent la consistance et donc le fait d'avoir une valeur finale de l'agrégation directement dépendante du degré de consistance est tout à fait naturel. Dans ce cas de super-additivité, la valeur de l'intégrale de Choquet reste toujours inférieure à la valeur obtenue par l'opérateur de moyenne pondérée. Une valeur négative du paramètre λ exprime un effet de redondance et dans ce cas, les valeurs de la fonction de l'ordre inférieur vont avoir une moindre influence sur le résultat de l'agrégation. Dans ce cas, de sous-additivité, la valeur de l'intégrale de Choquet reste toujours supérieure à la valeur obtenue par l'opérateur de moyenne pondérée.

6.2.1.3 Conclusion sur l'agrégation de dimensions de la qualité

Tous les opérateurs d'agrégation possèdent la propriété de **monotonicité**, exprimée par l'équation 6.2. En plus de cette propriété, dans le tableau 6.1, les principales propriétés de trois catégories d'opérateurs d'agrégation sont présentées. De plus, pour chacune de ces catégories leurs principaux inconvénients sont également exposés.

En revenant au problème formulé au début du paragraphe 6.2, il faut proposer des opérateurs adaptés à l'agrégation intra et inter critères. La première agrégation qui doit être réalisée est celle intra-critère, pour chaque critère de qualité. Dans ce cas, les mesures de qualité à agréger

Catégories	Propriétés	Inconvénients
Moyenne pondérée	• idempotente • linéaire • additive • symétrique (pour $\omega_i = \frac{1}{n}$) • permet l'association de poids d'importances aux critères • facile d'interpréter	• ne peut pas prendre en compte les interactions (dépendances) inter-critères
Opérateurs compensatoires	• prend en compte les dépendances inter-critères • intuitifs dans un certain degré	• propriétés exactes variables en fonction de la méthode de construction de l'opérateur
L'intégrale floue	• non-additive • non-linéaire • fondée sur une mesure floue • prend en compte les dépendances inter-critères	• difficile à interpréter

TABLE 6.1: Les propriétés et les inconvénients de trois catégories d'opérateurs d'agrégation

ont été définies pour quantifier la même caractéristique de la qualité. Ainsi, elle peuvent être transformées par des opérations spécifiques, comme par exemple la représentation sur la même échelle, afin d'arriver à des entités *commensurables*. Par conséquent, il est proposé d'utiliser un opérateur « *moyenne pondérée* », car ses propriétés linéaires et additives sont désirables pour ce cas. Soit le cas du critère de qualité, la *précision*, pour une information de géo-localisation. Supposons que deux mesures sont utilisées pour quantifier ce critère : la déviation standard longitudinale, ϵ_{Long} et latitudinaire, ϵ_{Lat}. En utilisant l'opérateur de Dyckhoff-Pedrycz avec $\alpha = 2$ (équation (6.5)), la valeur de ce critère est donnée par :

$$Cr_{\text{précision}} = \left(\sqrt{\frac{\epsilon^2_{\text{Long}} + \epsilon^2_{\text{Lat}}}{2}} \right) \tag{6.25}$$

La deuxième agrégation à réaliser est celle inter-critères. Comme dans ce cas les critères de qualité à agréger expriment des caractéristiques de qualité différentes, il est nécessaire d'utiliser un opérateur capable de prendre en compte les dépendances entre ceux-ci. Ainsi, dans [Todoran 14c] et [Todoran 14a], nous avons proposé d'utiliser l'intégrale de Choquet. Un exemple de son application et de ses capacités de prendre en compte les dépendances inter-critères d'une manière non-linéaire est présenté dans le paragraphe suivant.

6.2.2 Exemple d'agrégation de critères de qualité

Afin d'illustrer l'application des opérateurs d'agrégation dans un cas concret d'application, un système de reconnaissance automatique de cibles radar est considéré dans ce paragraphe. Ce type d'application, qui sera traité plus en détail dans le chapitre 7, est caractérisé comme ayant un degré élevé de risque à cause de possibles contremesures qui pourraient être envisagées. Sans rentrer dans les détails, il est supposé que l'information finale est qualifiée par les critères de qualité : { *Confiance*, *Complétude*, *Obsolescence*}.

À cause du caractère critique du système, supposons que l'utilisateur ait besoin d'être informé sur la qualité des informations en utilisant une seule valeur, qu'elle soit numérique, symbolique ou en code de couleurs. De plus, supposons que pour l'utilisateur :

1. La confiance et la complétude sont plus importante que l'obsolescence.

Information	Confiance	Complétude	Obsolescence	Qualité totale
Info1	0.90	0.80	0.50	0.720
Info1	0.50	0.60	0.90	0.680
Info1	0.65	0.70	0.70	0.695

TABLE 6.2: Exemple d'évaluation du score de qualité totale en utilisant l'intégrale de Choquet

2. La confiance et la complétude sont presque de même importance et il existe une forte corrélation entre les deux, signifiant qu'une information d'un bon niveau de confiance est aussi complète et vice-versa.
3. Une information d'un bon niveau de confiance ou complète, qui est en plus actuelle et qui décrit une information d'une très bonne qualité doit être mise en évidence.

Ces trois spécifications sont très difficiles à utiliser pour la construction d'un opérateur d'agrégation sous la forme d'une moyenne pondérée. Ainsi, dans ce genre de situation il est préférable d'utiliser un opérateur non-linéaire, comme l'intégrale de Choquet. Cette liste de spécifications peut servir à construire la mesure floue μ, utilisée par l'intégrale de Choquet. Ci-dessous, nous donnons à titre d'exemple une possibilité de construction de la mesure floue :

– La première spécification peut être transformée en :

$$\mu(\{\text{Confiance}\}) = 0.4\ ;\ \mu(\{\text{Complètude}\}) = 0.4\ ;\ \mu(\{\text{Actualité}\}) = 0.3 \tag{6.26}$$

– En utilisant la deuxième spécification :

$$\begin{aligned}&\mu(\{\text{Confiance}, \text{Complètude}\}) = \mu(\{\text{Complètude}, \text{Actualité}\}) = 0.6\ < \\ &\mu(\{\text{Confiance}\}) + \mu(\{\text{Complètude}\}) = 0.8\end{aligned} \tag{6.27}$$

– À partir de la troisième spécification il peut être déduit que :

$$\mu(\{\text{Confiance}, \text{Actualité}\}) = \mu(\{\text{Complètude}, \text{Actualité}\}) = 0.9 \tag{6.28}$$

plus grand que la somme individuelle $0.4 + 0.3 = 0.7$.

Les résultats d'application de l'intégrale de Choquet pour trois informations ayant des qualités différentes sont présentés dans le tableau 6.2. Dans le premier cas, l'information est d'une très bonne confiance et complète, mais elle manque d'actualité. Dans le deuxième cas, l'information est presque en temps-réel, mais avec un niveau médiocre de confiance et de complétude. Finalement, dans le troisième cas, l'information est d'un niveau moyen de confiance, de complétude et d'obsolescence.

Ainsi, en conclusion de cet exemple, l'intégrale de Choquet indique que l'utilisateur préfère la première information, même si les autres deux ont des scores de qualité proches.

L'expression des préférences de l'utilisateur sous la forme présentée dans cet exemple, accompagnée par l'utilisation de l'intégrale de Choquet peut aider l'analyste du système dans le développement du module de fusion. Ainsi :

– en fonction des préférences de l'utilisateur, seulement quelques sources d'information pourrait être considérées : celles offrant une information d'un niveau élevé de confiance ou d'actualité (pas obsolète).
– les modules de traitement de l'information peuvent être adaptés à ces préférences, par exemple en passant d'un algorithme à un autre en fonction de valeurs de critères de qualité.

6.3 CONCLUSION

Ce chapitre présente la dernière étape de notre méthodologie d'évaluation de la qualité de l'information d'un système d'information, c'est-à-dire la qualité globale. Afin d'obtenir la qualité de l'information proposée à l'utilisateur, nous avons utilisé l'évaluation de la qualité locale, présentée

dans le chapitre 4 et la notion de fonction de transfert de qualité, définie et analysée dans le chapitre 5.

L'évaluation de la qualité globale est fondée sur le principe de propagation de la qualité, car il permet de mettre à jour la qualité en sortie du système d'une manière automatique. Ainsi, tout changement de qualité à l'intérieur du système d'information est directement translaté à la sortie. De plus, si l'utilisateur a besoin d'avoir une explication de la provenance de la qualité, il est possible de lui présenter la traçabilité de la qualité à travers le système.

Également, l'agrégation des critères de qualité a été étudiée. Comme un critère peut être quantifié par plusieurs mesures de qualité, nous avons proposé de réaliser une agrégation en deux étapes. La première étape consiste dans l'agrégation des mesures de qualité pour chaque critère de qualité, afin d'avoir une seule valeur de qualité par critère. La deuxième étape consiste dans l'agrégation des valeurs de critères afin d'obtenir un score global de qualité exprimé par une seule valeur. Une étude de plusieurs familles d'opérateurs d'agrégation a été faite, en mettant l'accent sur leurs propriétés, afin de trouver les opérateurs d'agrégation les mieux adaptés pour ces deux étapes. Ainsi, il a été observé que dans le cas de l'agrégation intra-critère, les mesures de qualité peuvent être considérées comme commensurables. Par conséquent, il est possible d'utiliser une moyenne arithmétique pondérée. Pourtant, dans le cas de l'agrégation inter-critères, les valeurs sont incommensurables, exprimant des aspects différents de la qualité. Par conséquent, nous avons proposé comme solution l'intégrale de Choquet. Par rapport à un opérateur commun, elle est non-linéaire et permet de prendre en compte les interactions entre les divers critères. Un exemple a permis d'illustrer l'utilité de cet opérateur et a montré comment ce score global de qualité peut aider l'analyste du système dans la conception d'un système d'information.

« Creating information from data is complicated by the fact that, like beauty, what is truly "information" is largely in the eyes of the beholder »

Mika Endsley

TROISIÈME PARTIE : VALIDATION DE LA MÉTHODOLOGIE

Dans cette partie est réalisée la validation de la méthodologie en considérant deux systèmes d'information de deux domaines d'application différents. La première application est un système de reconnaissance automatique de cibles radar et la deuxième est un système d'aide au codage médical. Pour chaque système d'information, les trois étapes de notre méthodologie seront illustrées :

1. la décomposition du système dans ces modules élémentaires et l'évaluation de la qualité locale ;
2. construction de la fonction de transfert de qualité pour chaque module ;
3. évaluation de la qualité globale, du système entier.

TELECOM Bretagne

7 Étude d'un système de reconnaissance automatique de cibles radar

Dans ce chapitre, nous traitons un premier cas applicatif afin d'illustrer et de valider la mise en pratique de notre méthodologie. Dans le premier paragraphe 7.1, nous présentons une courte introduction de la problématique des systèmes de reconnaissance de cibles radar. L'accent de cette présentation sera mis sur les besoins informatifs et qualitatifs des utilisateurs des systèmes d'information dans le domaine de la défense. Ensuite, dans le paragraphe 7.2, un système multi-capteurs de reconnaissance automatique de cibles radar est choisi afin de servir pour la validation de la méthodologie d'évaluation de la qualité de l'information que j'ai proposé. Les trois étapes de notre méthodologie seront appliquées et l'analyse complète du système sera réalisée.

7.1 Introduction

Le système d'information que nous allons étudier est un système de reconnaissance automatique de cibles radar. Ce type d'application a un niveau de risque très élevé à cause des possibles contremesures qui peuvent être mise en œuvre. De plus, l'environnent militaire évolue très vite car les enjeux stratégiques et géopolitiques obligent à garder un avantage technologique face aux adversaires.

Un des problèmes majeurs au domaine de la défense est le manque de solution rapide et robuste pour l'identification d'objets dans l'espace de combat [Tait 05]. Le développement de systèmes de type « *Identification Friend or Foe* (IFF) » (identification ami ou ennemi) pour les cibles aériennes a permis une amélioration de la probabilité de reconnaissance des cibles aériennes amies. Cependant, en ce qui concerne les cibles ne fournissant pas une réponse positive, le système IFF peut seulement dire qu'elles soient suspectes. Ainsi, il peut y avoir des situations dans lesquelles les cibles amies qui n'ont pas pu coopérer, sont considérées comme des menaces. De plus, l'opinion publique est très virulente vis-à-vis de ce genre d'incidents tragiques [1] et donc, elle constitue un autre élément qui oblige à améliorer les performances des systèmes d'identification de cibles [2]. Ce problème, lié à l'identification des cibles hostiles, est aussi d'actualité dans le domaine civile concernant la lutte anti-terrorisme.

À présent, les systèmes radar, sonar, lidar, satellite, etc. sont des technologies arrivées à maturité. Ainsi, elles permettent d'enregistrer beaucoup de données qui peuvent être utilisées pour assurer une augmentation du niveau de sécurité en employant des mesures prophylactiques.

En conséquence, avec le déploiement de plus en plus de capteurs, le spectre des applications utilisant des données issues de ces capteurs continue à se diversifier et à augmenter. Ainsi, de plus en plus d'études se sont intéressées à la qualité des données capteur et de l'information issues de leurs traitements. Les études menées jusqu'à présent se sont focalisées surtout sur les traitements bas niveaux, c'est-à-dire pré-traitement de données, extraction d'information, etc. L'amélioration de la qualité des données par des calibrations appropriées des capteurs, la corrélation des données capteur issues des capteurs de proximité, la corrélation d'informations obtenues suite à de différents

1. voir l'effet CNN qui reporte chaque incident rapidement
2. les statistiques d'anciens combats disent qu'environ 10% de victimes sont dues au feu de leur propre armée

mécanismes d'extraction d'information sont parmi les stratégies les plus communes d'amélioration de la qualité d'un système d'information utilisant de données capteurs [Bisdikian 07].

Tandis que la quantité des données et de l'information continue à augmenter, dans ce contexte d'application, le temps de traitement et de présentation aux décideurs joue un rôle primordial. Ainsi, il est très important de qualifier les informations en fonction de leur degré de pertinence pour la situation actuelle, d'où le besoin d'accompagner les informations par des méta-informations présentant à l'utilisateur la qualité de ces dernières.

Dans la conception des systèmes complexes de reconnaissance, qui est le cas pour un système de détection, de localisation et d'identification de cibles radar, l'utilisation d'un seul module d'extraction d'informations, de classification, etc. devient problématique car un tel module aura des performances réduites ou dans le cas contraire, une très grande complexité (due à la nécessité de bien fonctionner dans toutes les situations, donc d'être le plus général possible). À cause de cela, les fonctions de maintenance, de compréhension des traitements réalisés, d'accès aux ressources, etc. vont être difficiles à réaliser. Une solution à ce problème est l'utilisation d'un système basé sur de multiples modules d'extraction d'informations, de multiples classifieurs et de multiples modules de fusion à chaque niveau de traitement.

Comme suggère [Appriou 01], dans la pratique, il est intéressant d'utiliser des capteurs ayant des résolutions spatiales orthogonales afin de pouvoir augmenter la précision de détection, de localisation et d'identification des cibles à l'aide d'une conjonction d'observations locales.

En conclusion, tout système d'information dans le domaine de la défense doit fournir des réponses (décisions) sous fortes contraintes de temps. Ainsi, ces systèmes d'information doivent utiliser toutes les informations disponibles et doivent fournir aux utilisateurs des informations d'une très bonne qualité. Un autre point important dans ce domaine est la nécessité d'offrir aux utilisateurs la possibilité de spécifier la qualité minimum nécessaire que le système d'information doit respecter. En fonction de ce niveau de qualité, le système d'information va pouvoir privilégier différentes sources d'information, différents moyens de communication de l'information (niveau de sécurité, rapidité, etc.) ou différents traitements de l'information. Mais pour tout cela, il faut premièrement être capable d'évaluer la qualité de l'information.

7.1.1 Le besoin pour la défense et pour le domaine civil

Le besoin actuel pour un système de détection, de localisation et d'identification est de posséder un système capable d'identifier les menaces possibles, avec une confiance élevée, à des distances plus grandes que le champ visuel. Comme support à cette affirmation, nous avons la déclaration du Contre-amiral Ph. Balisle, directeur de Navy Surface Warfare Division, États-Units [3]. Lors de ce discours, il a parlé de l'importance du système radar et des fonctionnalités qu'il doit avoir : détection automatique, identification et suivi des cibles, reconnaissance de cibles non coopératives (RCNC) [4]. Il a également insisté sur le besoin de pouvoir détecter les aéronefs hostiles et les missiles à des distances plus grandes que celles du champ d'action du système d'armement.

Une vision réaliste d'un théâtre d'événements militaires doit considérer une combinaison de cibles amies, ennemies et neutre, voir la figure 1.2. Ces cibles peuvent être aériennes, terrestres, marines pouvant appartenir aux forces militaires ou étant des objets civils. Ainsi, un système de reconnaissance de cibles devrait fonctionner dans toutes les conditions pour pouvoir réaliser une évaluation de la situation avec un degré de confiance acceptable.

La possibilité de pouvoir reconnaître tous les types de cibles à des grandes distances va permettre d'obtenir une **supériorité d'information** qui, à son tour, va permettre d'adopter des tactiques et des stratégies adaptées à la situation.

La reconnaissance de cibles est aussi utile pour le domaine civil. Pour mentionner quelques applications il y a : la lutte anti-terroriste, la lutte anti-drogue, le contrôle de l'immigration illégale,

3. Sa déclaration en face de Seapower Sub-committee of the Senate Armed Surface Committee on Surface Weapon Systems se trouve à l'adresse `http://www.navy.mil/navydata/testimony/seapower/pmbalisle020309.txt`

4. En anglais « NCTR : Non-Cooperative Target Recognition »

etc. Ainsi, dans le cas de la lutte anti-terroriste, un système radar côtier collaborant avec un système de contrôle du trafic aérien aura besoin d'intégrer la fonctionnalité de RCNC afin d'identifier les bateaux et les avions hostiles (cette fonctionnalité est complémentaire aux techniques coopératives de reconnaissance de cibles - par exemple l'IFF).

D'autres applications civiles pouvant profiter de la reconnaissance de cibles sont la lutte contre l'exploitation illégale des ressources dans des zones interdites : poissons, ressources minérales, etc. La sécurité des aéroports peut également s'améliorer par l'intégration des fonctions de reconnaissance de : cibles en approche et leur distinction de cibles se situant au sol, véhicules de l'aéroport pour pouvoir les contrôler, etc.

7.1.2 Description d'un système de reconnaissance de cibles radar

Très souvent dans la littérature, il y a une confusion dans l'utilisation de termes comme classification, identification et reconnaissance de cibles. Par la suite nous allons utiliser les définitions de [Tait 05] :

- **L'identification de cibles** est définie comme une description de la cible en termes plus précis (par exemple dire qu'une cible est un F-18 avec un degré de confiance satisfaisant peut être vu comme une identification de cibles). Dans la situation quand une cible est associée à un ensemble d'identité possibles (par exemple Rafale, MIG-29, Tornado ou F-16), il s'agit d'une classification très précise mais avec une identification partielle et/ou ambiguë.
- **La classification de cibles** ou encore **catégorisation de cibles** se définit comme étant une attribution des cibles à une classe générale. Ainsi, par exemple, il peut y avoir les catégories : missiles, aéronefs avec ailes et hélicoptères. D'autres catégories peuvent être aéronef militaire ou civil ; ou aéronef avec ou sans pilote ; ou aéronef ami, ennemi ou neutre ; ou aéronef hostile ou pas ; missile supersonique ou sous-sonique ; hélicoptère, aéronef d'attaque, aéronef de transport militaire, avion civil ou avion de petit taille. En conclusion, une cible aérienne peut être incluse dans beaucoup de catégories et pour cela il faut que la demande de classification de cibles soit accompagnée du type de catégorisation et classification dont l'utilisateur a besoin.
- **La reconnaissance de cibles** couvre la classification de cibles et l'identification de cibles. Dans la littérature est utilisée très souvent l'expression « Reconnaissance de Cibles Non Coopératives (RCNC) » pour exprimer l'ensemble des techniques et technologies utilisées pour obtenir une signature de haute résolution de la cible servant pour la prise de décision.

Le processus qui permet l'identification et la reconnaissance de cibles radar est présenté à la figure 7.1. La première étape est l'utilisation d'un radar adapté aux cibles qui doivent être détectées et identifiées (par exemple un radar de haute résolution). De plus, afin d'obtenir des signatures de bonne qualité, la forme d'onde du signal émis joue un rôle très important, voir la référence [Tait 05] pour plus de détails. En même temps, le système radar doit être construit de manière à minimiser les distorsions des signaux émis et reçus. Une fois que le signal radar reçu et traité, les signatures de cibles radar sont obtenues (la cible est vue comme étant formée d'un ensemble des parties et chacune de ces parties sera imagée par le système radar - le signal complexe reçu est composé d'un module et d'une phase). Pour la tâche de reconnaissance de cibles radar, il faut comparer les signatures obtenues avec des signatures de référence. Ces signatures de référence doivent être également de très bonne résolution, pour tous les types de cibles radar et prises dans des conditions assez diverses (par exemple sous des différents angles). Le moyen d'obtenir ces signatures de référence peut être soit en utilisant des modèles mathématiques, soit en utilisant des mesures de haute résolution de cibles radar d'intérêt. Bien sûr, une combinaison de ces deux moyens peut aussi être employée.

En fonction de la méthode et des algorithmes de reconnaissance de cibles utilisées, les signatures de référence seront traitées de manière différente. Par exemple, il est possible d'extraire seulement quelques caractéristiques de ces signatures de référence, à l'aide de modélisations mathématiques. Les algorithmes de reconnaissance de cibles radar font une comparaison entre les signatures ob-

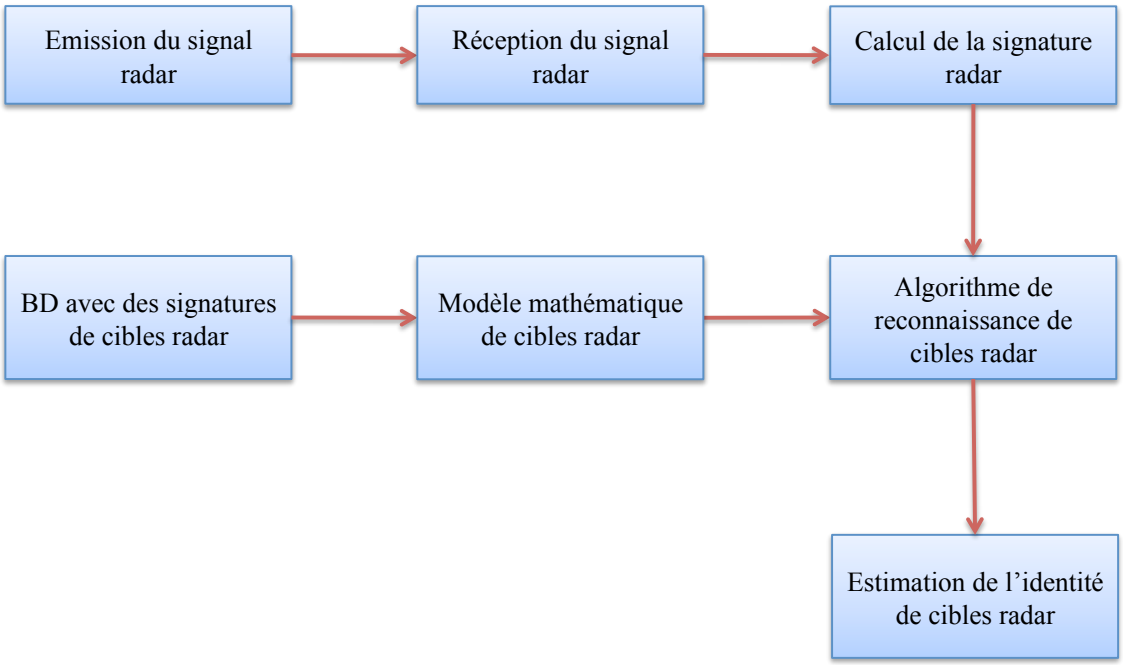


FIGURE 7.1: Processus d'identification de cibles radar

tenues à partir du signal reçu et les signatures de la banque de données (BD). La comparaison entre les deux signatures pourrait se faire par exemple, en utilisant une mesure de plausibilité qui va indiquer le degré de ressemblance entre la signature mesurée et celles de référence. La qualité de la reconnaissance de cibles radar est dépendante de la qualité des données radar, la similarité entre les propriétés physiques des cibles, la qualité des modèles mathématiques des cibles et les performances des algorithmes de reconnaissance [Tait 05].

Dans la littérature, la grande majorité des solutions pour le problème de détection, de localisation et d'identification de cibles radar sont exprimées dans un cadre mathématique probabiliste [Waltz 90], [Tait 05].

Dans la figure 7.2 se trouve l'architecture d'un système d'identification de cibles radar proposée par [Toumi 07]. Cette architecture a été développée en respectant la méthodologie CRISP-DM [5] pour l'extraction et la gestion des connaissances [6].

De cette architecture, il peut s'observer que le signal radar est transformé en deux types de données différents : une représentation temporelle unidimensionnelle et une représentation bidimensionnelle sous la forme d'une image. Ces deux types de données ne sont pas encore adaptés au processus d'identification et donc, une transformation est faite afin d'extraire un ensemble de paramètres (des descripteurs) permettant la discrimination de cibles radar. Le choix des descripteurs à extraire est une tâche difficile parce qu'ils doivent être, d'une part de taille minimale et d'autre part, ils doivent contenir une quantité suffisante d'information pour pouvoir réaliser l'identification indépendant des transformations géométriques de la cible (rotation, translation, changement d'échelle, etc.). Après la classification du signal unidimensionnel et de l'image, une fusion des informations représentant l'identité de la cible est utilisée afin de fournir à l'utilisateur final l'identité la plus probable.

Le système présenté figure 7.2 représente le système classique mono-capteur d'identification de cibles radar. Comme indiqué dans l'introduction de ce chapitre, dans la pratique plusieurs

5. Acronyme de « Cross Industry Standard Process for Data Mining »
6. En anglais *Knowledge Descovery in Databases* - KDD

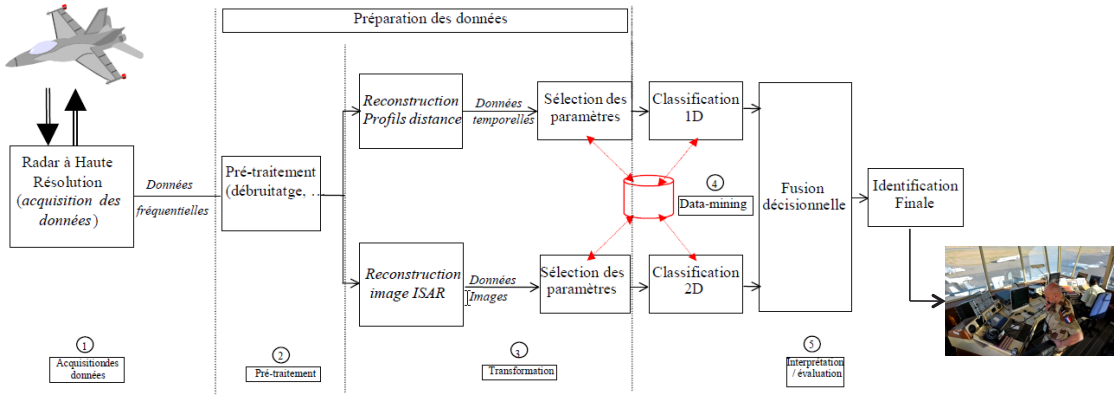


FIGURE 7.2: Architecture pour l'identification de cibles radar [Toumi 07]

Capteur	Signature détectable	Dépendances de la signature
MMW Radar	Surface équivalente radar, vélocité	Forme, composition du matériel, cavités, rugosité régularité de la surface, polarisation du récepteur, direction de déglacement par rapport au capteur
Infrarouge FLIR/IRST	Émission et réflectivité	Radiance de la part des moteurs, des sources naturelle (Soleil) ou radiation réfléchie
Mesures de soutien électronique	Émission électronique	Capteurs actifs et sources de transmission comme l'équipement de communication, de navigation et toute autre source de radiation électromagnétique
Acoustique	Énergie acoustique	Bruit des moteurs, bruit de l'objet lors de son mouvement en air ou sur la surface terrestre

TABLE 7.1: Caractéristiques des signatures de quatre types de capteurs, selon [Klein 14]

capteurs sont utilisés afin d'augmenter les performances de ce type de système. Dans le tableau 7.1, les caractéristiques principales des signatures de quatre types de capteurs sont présentées. En les analysant, on s'aperçoit que ces capteurs vont fournir des données représentant des caractéristiques physiques différentes (du même objet).

En conséquence, dans le paragraphe suivant, nous présentons une possible architecture simplifiée d'un système multi-capteur afin de servir pour la validation de notre méthodologie d'évaluation de la qualité.

7.2 VALIDATION DE LA MÉTHODOLOGIE PAR UN SYSTÈME MULTI-CAPTEURS DE RECONNAISSANCE AUTOMATIQUE DE CIBLES

La validation de la méthodologie d'évaluation de la qualité de l'information est réalisée en exécutant ses trois étapes :

1. Définition et évaluation de la qualité locale, c'est-à-dire la qualité en sortie de chaque module élémentaire de traitement ;
2. Estimation de la fonction de transfert de qualité pour chacun des modules élémentaires de traitement du système ;
3. Évaluation de la qualité globale, c'est-à-dire la qualité de l'information proposée à l'utilisateur final.

Capteur	Avantages	Inconvenant
Radar	Toute condition météo Fréq. basses pénètrent le feuillage Zone étendue d'observation Opérationnel jour et nuit Données : distance, vitesse et image	Résolution modérée Pas couvert Sensible au brouillage
IR-EO	Résolution fine de l'image spatiale et spectrale Opérationnel jour et nuit	Affecté par les cond. météo Pénétration faible du feuillage Maximisation du SNR plus difficile Besoin d'un mécanisme de balayage ou d'un grand réseau de détecteurs pour couvrir une zone d'observation large

TABLE 7.2: Performances des capteurs radar et infrarouge, selon [Klein 14]

Par la suite chacune de ces trois étapes sera présentée.

7.2.1 Évaluation de la qualité locale

La première étape de notre méthodologie, correspondant au chapitre 4, est la décomposition du système d'information en modules élémentaires afin de pouvoir réaliser l'évaluation de la qualité locale. L'architecture simplifiée d'un système multi-capteurs de reconnaissance automatique de cibles radar est présentée dans la figure 7.3. Ce système fait appel aux données issues de trois capteurs différents : un radar, un capteur infrarouge-électrooptique (IR-EO) et un système d'identification ami ou ennemi (IFF). Dans le tableau 7.2 une comparaison des performances du système radar et du système IR-EO est réalisée. L'objectif de cette comparaison est de montrer qu'en fonction des conditions de l'environnement sous observation, chaque capteur a ses avantages et inconvénients rendant les données fournies sensibles aux diverses imperfections, cf. chapitre 2.

Ensuite, ce système utilise trois extracteurs d'information, spécialisés dans l'identification des différentes caractéristiques des objets à reconnaître. En effet, le module implémentant la classification de cibles radar peut être encore décomposé pour finalement arriver à un vision du système présentée dans la figure 7.2. Dans cette étude, nous ferons l'hypothèse que nous n'avons pas accès à une vision plus détaillée et par conséquent, le module de classification de cibles radar sera traité comme un module élémentaire. D'un point de vue de leur fonctionnement, ces trois extracteurs d'information ont un comportement semblable à trois classifieurs.

Le premier module d'extraction d'information, le *Classificateur de la signature radar*, reçoit à son entrée une signature radar et réalise une identification de la cible radar détectée. Un exemple de signature radar est présentée dans la figure 1.7, correspondant à un F4 (l'identité de la cible radar). Le deuxième module, le *Classificateur de la signature IR-EO*, reçoit à son entrée une signature infrarouge-électrooptique (IR-EO) et réalise également une identification de la cible radar détectée. Le troisième module, l'*IFF*, reçoit à son entrée une réponse de type IFF et fournit l'allégeance de la cible, c'est-à-dire amie, ennemie ou neutre.

Un premier module de fusion, *Fusion d'identité*, collecte les informations fournies par les deux premiers modules afin de délivrer l'identité de la cible. Un deuxième module de fusion, *Fusion d'informations*, collecte l'identité de la cible, fournie par le premier module de fusion, et l'allégeance de la cible de la part du module IFF. Il a pour rôle de corroborer les deux informations afin de délivrer une information finale, complète, sur la cible détectée, c'est-à-dire son identité et son allégeance.

Avec le système d'information décomposé, il est possible de définir les critères de qualité adaptés à chaque module de traitement en fonction des caractéristiques des données/informations (c'est-à-dire la qualité locale). Dans le tableau 7.3 sont présentés, pour chaque module de traitement, les

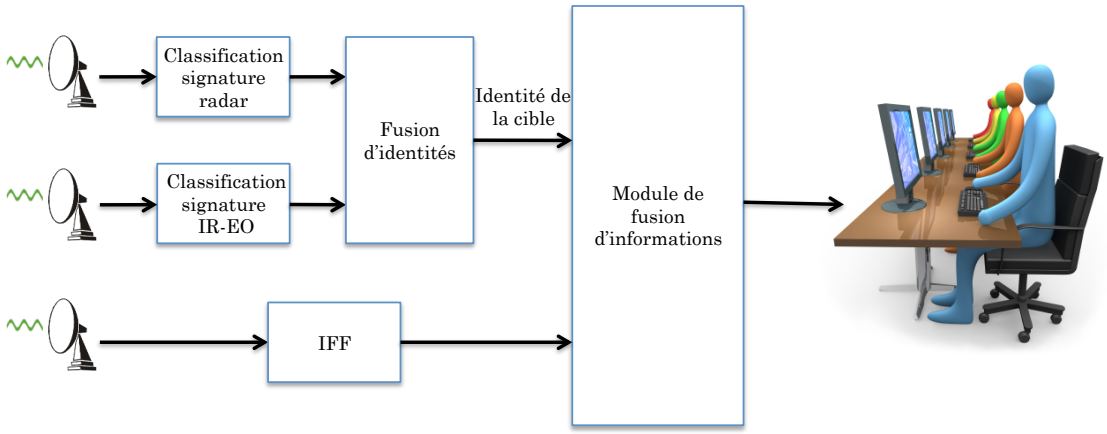


FIGURE 7.3: Architecture simplifiée d'un système multi-capteurs de reconnaissance automatique de cibles

critères de qualité en entrée et en sortie, accompagnés par des mesures de qualité [Todoran 14b].

7.2.2 Construction de la fonction de transfert de qualité pour chaque module de traitement

La deuxième étape de la méthodologie, correspondant au chapitre 5, a pour objectif de construire la fonction de transfert de qualité de chaque module de traitement. Ceci est réalisé par l'analyse individuelle de chaque module. Afin de facilité l'étude de l'influence du module sur la qualité, une interface développée en Python a été créée [Todoran 14b]. Dans la figure 7.4 est présentée l'analyse du module responsable de la classification de signature radar.

À l'aide de cette interface, en variant la qualité en entrée du module, Q_{in} de nouvelles valeurs de qualité peuvent être obtenues en sortie, Q_{out}. Ainsi, en enregistrant les valeurs de qualité en entrée et ses correspondants en sortie, des couples (Q_{in}, Q_{out}) peuvent être obtenus pour chaque critère de qualité. Chaque critère de qualité en sortie du module, $Cr_i \in Q_{out}$, sera fonction de tous les critères de qualité en entrée. Dans la figure 7.5, est illustrée la fonction de transfert de qualité pour le module de classification des signatures radar. Comme la qualité en sortie de ce module est composée de deux critères de qualité, la *confiance* et l'*obsolescence*, la fonction de transfert de qualité est représentée par deux graphiques, un pour chaque critère. Le graphique de gauche de la figure 7.5 présente la confiance en sortie du module, en fonction des critères d'entrée, la *précision* et la *quantité de données*. À première vue, il peut paraître étrange de ne pas considérer l'effet du critère *obsolescence* sur la *confiance*. Mais, il ne faut pas oublier qu'il existe de fortes dépendances entre les trois critères de qualité de l'entrée. Ainsi, afin d'avoir des données plus précises, il est nécessaire d'attendre que la cible approche du système radar. Donc, la précision des données est directement dépendante du temps attendu pour leur enregistrement. La même observation peut se faire pour l'autre critère, la quantité de données. Afin d'avoir une meilleure résolution (pixels plus petits) il est nécessaire d'imager une cible proche du système. En conclusion, la *précision* et la *quantité de données* sont en forte corrélation (négative) avec le critère de qualité temporel, l'*obsolescence*. Ainsi, ce dernier n'a pas été considéré pour le calcul de la *confiance*, car son influence est déjà prise en compte dans les deux autres critères de qualité de l'entrée.

Le graphique de droite de la figure 7.5 représente l'*obsolescence* de l'information sur l'identité de la cible, en fonction des critères de qualité d'entrée, l'*obsolescence* et la *quantité de données*. Plus précisément, l'obsolescence de l'information est déterminée par l'obsolescence des données utilisées pour son extraction et par le temps de traitement nécessaire, directement dépendant de

Module	Information en entrée	Qualité en entrée		Information en sortie	Qualité en sortie	
		Critère	Mesure		Critère	Mesure
Classification Signature Radar	Image radar	Quantité données	Taille de l'image	La classe d'identité	Confiance	Degré de confiance
		Précision	Niveau de bruit		Obsolescence	Timestamp
		Obsolescence	Timestamp			
Classification Signature IR-EO	Image IR-EO	Quantité données	Taille de l'image	La classe d'identité	Confiance	Degré de confiance
		Précision	Niveau de bruit		Obsolescence	Timestamp
		Obsolescence	Timestamp			
Identification IFF	Signal 1-D	Précision	Niveau de bruit	Classe d'allégeance	Confiance	Degré de confiance
		Obsolescence	Timestamp		Obsolescence	Timestamp
Fusion d'identité	Classe d'identité	Confiance	Degré de confiance	Classe d'identité	Confiance	Degré de confiance
		Obsolescence	Timestamp		Obsolescence	Timestamp
Fusion d'informations	Id. & allégeance	Confiance	Degré de confiance	Identité & allégeance	Confiance	Degré de confiance
		Obsolescence	Timestamp		Obsolescence	Fraîcheur
					Complétude	Degré de présence de toutes les informations

TABLE 7.3: Les critères de qualité accompagnés de leurs mesures de de qualité

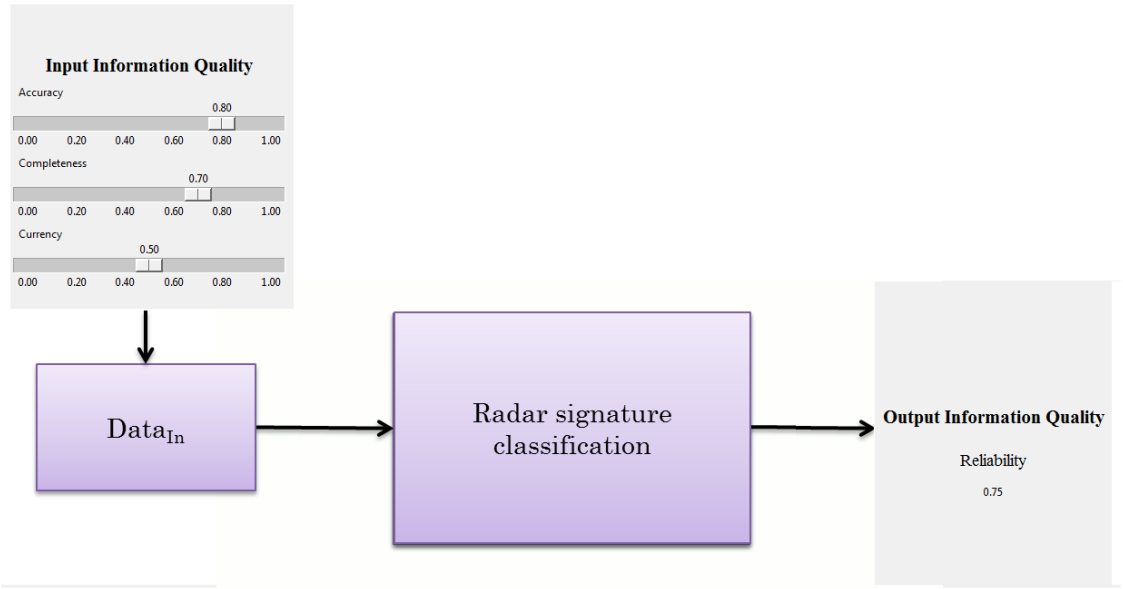


FIGURE 7.4: Interface d'analyse du module responsable avec la classification de signature radar afin de déterminer sa fonction de transfert de qualité

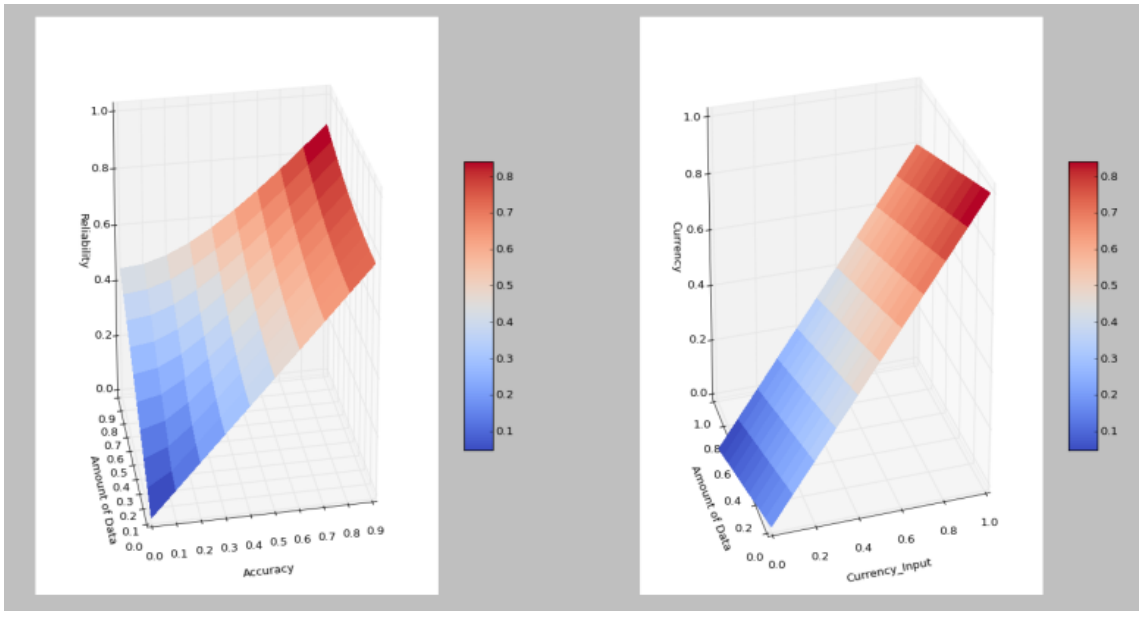


FIGURE 7.5: La fonction de transfert de qualité pour le module de classification de signatures radar (à gauche, *Confiance* et à la droite, l'*Obsolescence*)

la quantité de données. De ce graphique, il peut s'observer que l'obsolescence de données a la plus grande influence sur l'obsolescence de l'information fournie par le classificateur de signature radar.

Comme l'objectif de ce paragraphe est de montrer la faisabilité de la construction de fonctions de transfert de qualité pour les modules de ce système, les autres modules ne seront pas étudiés car leur analyse est identique à celle du module de classification de signature radar.

7.2.3 Évaluation de la qualité globale du système

Dans la troisième étape de la méthodologie, correspondant au chapitre 6, est présentée l'évaluation de la qualité globale du système entier. Dans la figure 7.6 est présentée une interface permettant de changer les valeurs de la qualité locale (un ou plusieurs critères de qualité). À l'entrée et à la sortie de chaque module de traitement se retrouvent les critères de qualité définis dans le paragraphe 7.2.1 avec leurs valeurs courantes. Le changement d'une de ces valeurs de qualité a un impact immédiat sur les autres valeurs de qualité situées en aval de la direction du flux d'informations. Ainsi, la qualité globale, caractérisant les performances du système entier, est automatiquement mise à jour suite au principe de propagation de la qualité énoncé dans le 6.1.

De plus, si l'un des modules est remplacé par un autre (illustré en rouge dans la figure 7.6), la méthodologie d'évaluation de la qualité reste toujours applicable. La seule demande, comme il a déjà été présenté dans le paragraphe 6.1, est d'analyser ce nouveau module afin de déterminer sa fonction de transfert de qualité. Ainsi, quand le remplacement est réalisé, la qualité en aval de ce module est automatiquement recalculée et mise à jour.

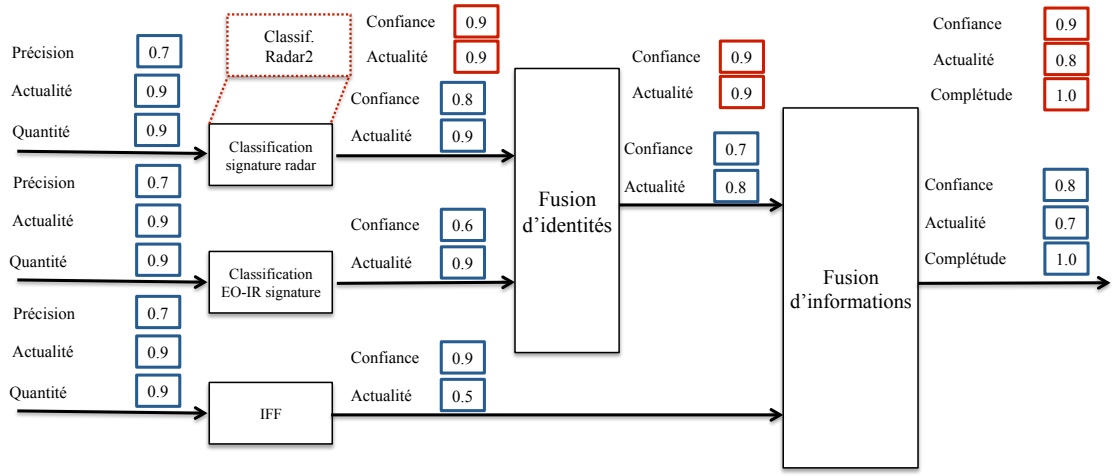


FIGURE 7.6: L'influence du changement de la qualité locale sur la qualité globale (illustration de la propagation de la qualité)

7.3 CONCLUSIONS

Les systèmes d'information dans le domaine de la défense doivent être robustes et réactifs dans toutes les situations. Par conséquent, il est indispensable d'accompagner chaque information proposée aux utilisateurs par des niveaux expliqués de qualité afin de permettre à ces utilisateurs de les incorporer facilement dans le processus cognitif de prise de décision. C'est pour cela que notre méthodologie est d'un réel intérêt pour ce type d'application.

Dans ce chapitre, la validation de notre méthodologie d'évaluation de la qualité de l'information a été réalisée dans le cadre d'un système multi-capteurs de reconnaissance automatique de cibles

radar. Les trois étapes composant la méthodologie ont été, une après l'autre, appliquées afin de tester le fonctionnement et la fiabilité de celle-ci.

Lors de la première étape de la méthodologie, nous avons défini pour chaque module élémentaire de traitement les critères de qualité à son entrée et à sa sortie. De plus, nous avons proposé des mesures de qualité spécifiques pour chacun de ces critères de qualité.

Dans la deuxième étape de la méthodologie, nous avons pris le cas du module de classification de cibles radar et nous l'avons analysé afin de modéliser son influence sur la qualité sous la forme d'une fonction de transfert de qualité. Comme en sortie de ce module la qualité de l'information utilise deux critères de qualité, la fonction de transfert de qualité est composée de deux fonctions, une pour chaque critère. Cela respecte l'algorithme de calcul de la fonction de transfert de qualité, présenté algorithme 1.

Dans la troisième étape, nous avons montré comment la qualité locale et la fonction de transfert de qualité permettent d'estimer automatiquement la qualité globale, c'est-à-dire la qualité des informations proposées à l'utilisateur final. De plus, nous avons montré que dans le cas d'une évolution du système, suite à une mise à jour d'un de ses modules, notre méthodologie est suffisamment flexible pour demander un minimum de modifications : un calcul de la nouvelle fonction de transfert de qualité propre à ce nouveau module.

8

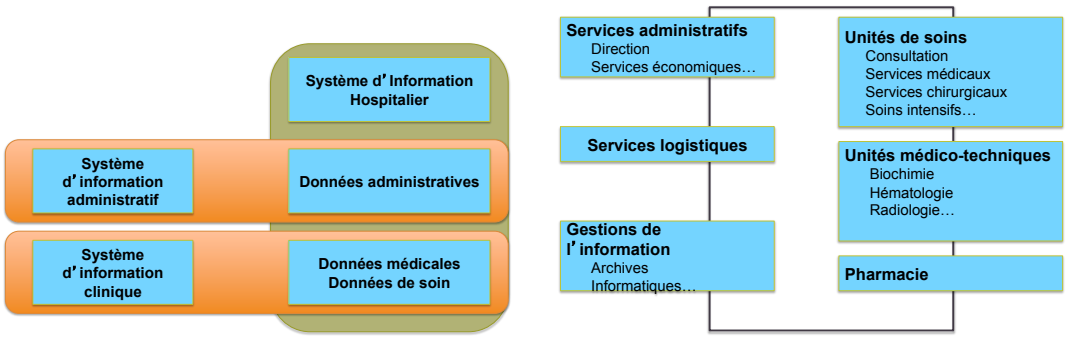
Étude d'un système d'information hospitalier

Un système d'information hospitalier est destiné à faciliter la gestion de l'ensemble des informations médicales et administratives d'un hôpital. La figure 8.1a présente le système hospitalier en interaction avec deux autres systèmes d'information : un système d'information administratif et un système d'information clinique. De ce fait, le système hospitalier a accès à un volume important de données. Il fait intervenir plusieurs acteurs (utilisateurs) : administrateurs de bases de données, opérateurs alimentant ces bases de données, utilisateurs finaux (des médecins), etc. Ainsi, le système d'information hospitalier peut également être vu par rapport aux services soutenus, figure 8.1b.

Une partie d'une architecture d'un système informatique hospitalier est présentée dans la figure 8.2. Cette architecture reçoit en entrée quatre grandes sources de données :

- Le PMSI, acronyme du « Programme de Médicalisation des Systèmes d'Information ». Le PMSI est un dispositif qui permet de mesurer l'activité et les ressources des établissements médicaux grâce à des données quantifiées et standardisées. Parmi les données qui peuvent se retrouver dans le PMSI, il y a les données concernant les séjours des patients dans l'hôpital : informations des patients (sexe, âge, etc.), le diagnostic principal, les diagnostics reliés, les diagnostics associés, etc. ;
- Les Résultats Biologiques : issus des analyses de laboratoire ;
- Les DMI [Papin 08], acronyme de « Dispositifs Médicaux Implantables » : prothèses, implants, défibrillateurs, etc. ;
- Les Médicaments Onéreux : médicaments (anticancéreux) innovants et onéreux.

À partir de ces bases de données, un entrepôt de données est construit afin de pouvoir faciliter la tâche d'extraction d'informations utiles pour le médecin. Un entrepôt de données est une collection de données, orientée sujet, intégrée, variant en temps et non-volatile et qui sert comme support au processus de prise de décisions [Chen 01]. Cet entrepôt de données est la pierre angulaire du système



(a) Différents systèmes d'information dans un environnement hospitalier

(b) Les services d'un système d'information hospitalier

FIGURE 8.1: Le système d'information hospitalier

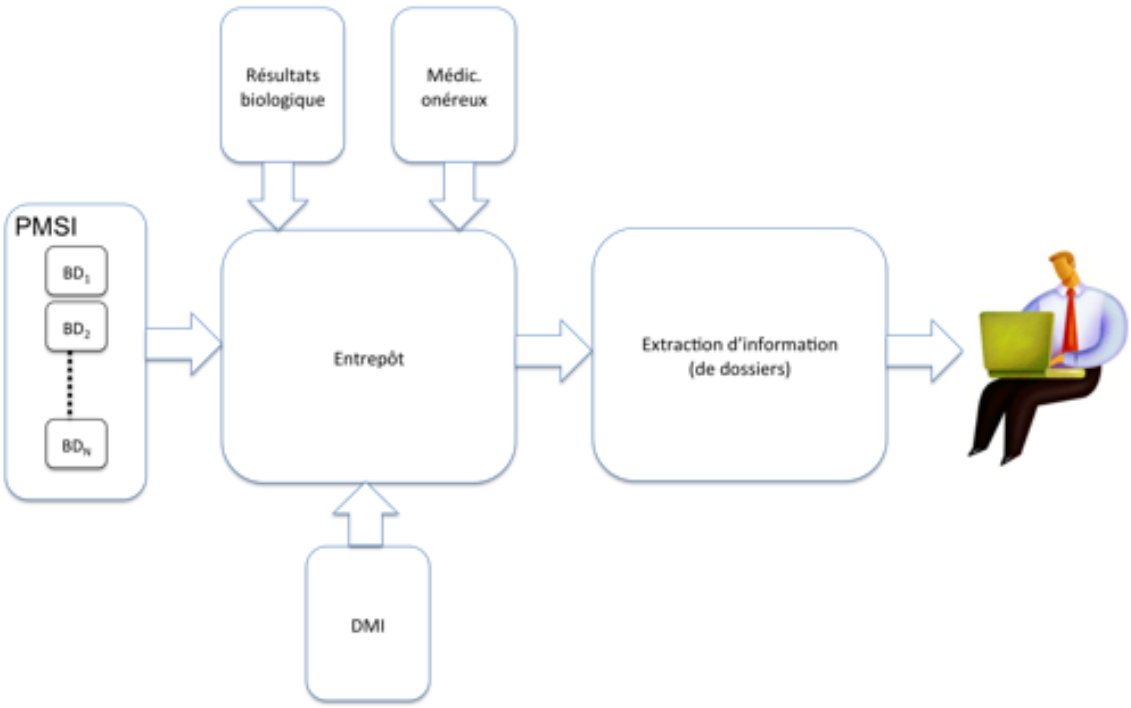


FIGURE 8.2: Une partie du système d'information médical

d'information. Les informations demandées par les utilisateurs en sortie du système d'information peuvent être vues comme des dossiers de patients ayant des caractéristiques communes, comme par exemple les patients qui ont subi un traitement particulier.

L'avantage d'utilisation d'un tel système est que l'utilisateur final a à sa disposition les informations nécessaires (complètes et concises) sans être obligé de faire des requêtes individuelles sur chaque base de données.

Dans le paragraphe 8.1, nous présentons les problèmes de qualité pour chacune des sources de données. Ensuite, dans le paragraphe 8.2, nous présentons une possible stratégie d'étude de la qualité des données de ce système d'information. Puis, dans le paragraphe 8.3, nous considérons un système d'aide au codage médical, partie du système d'information hospitalier, afin de valider notre méthodologie.

8.1 LA QUALITÉ DES SOURCES DE DONNÉES

Maintenant, une étude d'un point de vue qualitatif sera faite sur ce système d'information. L'objectif de cette étude est d'analyser la qualité des données et des informations circulant à l'intérieur de ce système. Dans un premier temps, les sources de données seront analysées.

Le PMSI contient plusieurs bases de données, issues de différents systèmes informatiques ayant des caractéristiques hétérogènes. Il est possible que ces données aient été enregistrées en utilisant des éditeurs différents au cours du temps. Malheureusement, il existe des situations pour lesquelles les évolutions des systèmes informatiques utilisés à l'intérieur des établissements n'ont pas été suivies avec une mise à jour des bases de données. Ainsi, dans ce genre de situation, les principaux problèmes de qualité des données sont représentés par l'apparition de *doublons*, d'*inconsistances* et d'*erreurs*. Les doublons et les inconsistances sont principalement dus à l'utilisation de plusieurs systèmes informatiques qui, à cause d'une mauvaise conception, travaillent d'une manière semi-indépendante générant des données avec des caractéristiques différentes. Un exemple d'une

telle inconsistance est l'apparition de patients ayant plusieurs numéros d'identification. Les erreurs pouvant apparaître sont principalement liées au processus d'enregistrement manuel de données, accompagné par une absence ou une définition inadaptée de la normalisation du modèle des données saisies. La principale conséquence des données erronées est l'apparition des incohérences dans différents enregistrements. Un exemple est le cas des patients ayant une date de fin de séjour précédent la date de début de séjour.

Le DMI est une base de données dans laquelle le plus grand problème est la traçabilité des dispositifs, c'est-à-dire la possibilité de connaître pour chaque patient le dispositif exact qui lui a été implanté. La traçabilité est principalement affectée par des erreurs de datage, entre l'envoi du dispositif et la date d'implantation de celui-ci chez le patient. Il existe de cas quand dans la base de données DMI il y a des implantations avant que le produit soit livré d'où des problèmes de cohérence.

La base de données correspondant aux **molécules onéreuses** est affectée par le même type de problème que dans le cas de DMI.

La base de données des **résultats biologiques** contient un grand nombre de données, chaque donnée ayant beaucoup d'attributs. Ce grand volume de données est issu de l'utilisation de capteurs de dernière génération capables d'enregistrer un spectre très large de mesures. En conséquence, le premier problème est lié au grand volume de données accessibles. Cette base de données est aussi influencée, comme le PMSI, par les différents changements au niveau de la technologie et de la modification permanente des noms des mesures prises. Ainsi, il peut y avoir des problèmes de consistance liés à un mauvais référencement : différentes échelles. De plus, comme toute donnée issue d'un capteur, les résultats biologiques sont aussi affectés par des imperfections liées aux caractéristiques du capteur : résolution, précision, etc.

8.2 La stratégie d'étude

Pour cette application, il existe plusieurs étapes de traitement des données qui doivent être analysées. Ces étapes sont identiques à celles présentées dans [Berti-Equille 06] :

- création des données ;
- collecte/import des données ;
- stockage des données ;
- intégration des données ;
- recherche et analyse des données.

Dans un premier temps, il faut que chaque base de données soit analysée d'un point de vue qualitatif afin de pouvoir lui associer des critères de qualité accompagnés par des mesures de qualité. L'analyse de la qualité des bases des données doit se faire en s'axant sur les problèmes d'inconsistance, d'incohérence, de données manquantes et erronées, du volume de données, d'accessibilité, de sécurité (Cf. chapitre 2)

Dans un deuxième temps, il faut que l'entrepôt soit analysé pour connaître exactement son mode de construction (quelle stratégie a été implémentée). La construction d'un entrepôt peut se faire suivant plusieurs modèles avec une implémentation en : étoile, relationnel ou encore en flocons [Di Ruocco 12]. Cette analyse devra nous indiquer comment l'intégration des données a été faite et si pour une donnée quelconque il existe un accès (une référence) à la (ou aux) base(s) de données qui l'a produite. Le processus d'intégration de données de différentes bases de données ayant des niveaux de qualité différents peut générer des pertes de données ou encore des inconsistances. À cause de cela, il est nécessaire d'évaluer au préalable la qualité des bases de données et ensuite, de calculer la qualité de données contenues dans l'entrepôt, car ce dernier sera utilisé dans le processus d'extraction d'information. Normalement, suite au processus d'intégration de plusieurs sources de données, le résultat est supposé posséder une meilleure qualité. L'augmentation du niveau de qualité de l'entrepôt est réalisée par un phénomène de combinaison de sources contenant des données représentant des aspects plus ou moins différents de la même réalité physique. Mais cette augmentation en niveau de la qualité est directement dépendante du niveau de cohérence

et de consistance entre les différentes bases de données. Ainsi, l'intégration des données dans l'entrepôt est, dans la plupart des cas, précédée par des prétraitements afin d'avoir des données mieux structurées et d'une meilleure qualité intrinsèque (ex. élimination des données manquantes et erronées).

Le suivi des changements de qualité entre les données contenues dans les bases de données et les mêmes données contenues dans l'entrepôt reste une tâche difficile à cause de la complexité et du volume très grand de données.

8.2.1 L'étude de la qualité des bases de données

Les problèmes de qualité des bases de données ont été introduits dans le paragraphe précédent. Ainsi, la qualité des données sera quantifiée par les critères de qualité suivants : la cohérence, la complétude, la consistance, l'inverse du niveau de données manquantes et erronées, l'inverse du volume de données, le niveau d'accessibilité et de sécurité. Par la suite, ces critères de qualité seront définis et évalués indépendamment l'un de l'autre.

1. *La cohérence* : Dans la première étape, il faut étudier la façon dont les bases de données ont été construites. Plus précisément, il faut regarder les contraintes logiques de création des tables. Pour que les entités contenues dans les bases de données soient cohérentes il faut que des contraintes soient définies et respectées, cf. paragraphe 2.1. Ainsi, pour chaque base de données il faut commencer par une analyse des contraintes nécessaires pour chaque attribut. Après cette analyse, il faut vérifier si, lors de la construction de cette base de données, ces contraintes ont été définies. Dans le cas positif, la base de données peut être considérée comme cohérente. Dans le cas contraire, quand la base de données n'impose pas ces contraintes, il faut voir à quel niveau les données sont affectées. Par exemple, dans le cas d'une base de données ayant comme attributs les dates de début et du fin du séjour il faut que la contrainte : DateDébut DateFin soit définie. Dans le cas contraire, il faut quantifier le nombre d'individus qui ne respectent pas cette règle.
2. *La consistance* : une entité peut avoir plusieurs enregistrements dans une base de données, soit des copies, soit des nouveaux enregistrements. Aussi, dans le cas de plusieurs bases de données, la même information peut être contenue dans plusieurs endroits suite à des problèmes de saisie ou de mises à jours, des inconsistances peuvent apparaître. Pour que la base de données soit consistante il faut que les attributs des mêmes entités aient les mêmes valeurs (nom, prénom, date de naissance, numéro de sécurité sociale, etc.). Pour le processus d'agrégation de plusieurs bases de données, la consistance est une dimension très importante. Sans elle, le processus d'agrégation ne peut pas se faire directement et il nécessite un prétraitement (ex. deux enregistrements de la même personne ayant deux adresses différentes. Pour réaliser l'agrégation on peut garder l'adresse de l'enregistrement le plus récent).
3. *La complétude* : des entités (que toutes les entités sont représentées dans le modèle de la base de données) ; des attributs (il faut que la liste d'attributs des entités soit exhaustive) ; des relations (l'ensemble des associations entre les entités doit être exhaustif) ; des occurrences (toutes les occurrences d'une entité doivent apparaître).
4. *Le volume de données* : quantifié par le nombre d'individus et d'attributs de chaque table. En plus, il faut évaluer le taux d'entrée de nouveaux individus dans la table. Ce critère de qualité est en corrélation négative avec le critère d'accessibilité, à cause du temps de traitement qui est directement dépendant du volume de données.
5. *Le niveau d'accessibilité* : c'est un critère qui doit être évalué par un expert. Il traduit les efforts nécessaires pour récupérer les données.
6. *Le niveau de sécurité* : c'est un critère de qualité impératif au sens où une base de données doit être construite de façon à ne permettre l'accès qu'aux utilisateurs voulus et après une étape d'identification. De plus, certaines données doivent être anonymisées avant de les traiter, afin de protéger les informations personnelles des patients.

Patient					
IPP	Nom	Prénom	Date_Naissance	Sexe	Adresse

TABLE 8.1: La table Patient du PMSI

Séjour					
No_Séjour	IPP	Date_Début_Séjour	Durée_Séjour	Diagnostiques	Actes_Codés

TABLE 8.2: La table Séjour du PMSI

Pour qu'un échange de données entre différentes applications soit possible, il faut que des standards soient utilisés dans la construction des bases de données. L'utilisation des standards permet non seulement de garantir une consistance des données, mais aussi une facilité d'intégration de plusieurs bases de données, pas de la construction de l'entrepôt de données.

a) L'étude des bases de données du PMSI

Cette base de données contient plusieurs tables organisées sous la forme d'un modèle relationnel. Dans notre cas d'étude, les tables les plus importantes sont les tables **Patient** et **Séjour**. Le format de ces deux tables est présenté dans les tableaux 8.1 et 8.2 :

La table « *Patient* » a pour clé primaire l'IPP (l'identifiant du patient). Un IPP unique est alloué pour chaque nouveau patient. La table « *Séjour* » a comme clé primaire l'attribut No_séjour et comme clé étrangère l'attribut IPP.

La première chose qui doit être faite est de définir les attributs devant impérativement prendre une valeur : nom, prénom, date de naissance, adresse, date du début du séjour, date de la fin du séjour, diagnostique principal. Si un de ces attributs n'a pas de valeur saisie, il faut essayer de le remplir en se basant sur les autres enregistrement appartenant au même patient. Exemple : si l'adresse n'est pas saisie, une recherche sur les patients ayant le même nom, prénom et date de naissance devrait fournir l'adresse. Après avoir rempli tous les champs, il faut vérifier qu'il n'y a pas de valeurs aberrantes. Quand des enregistrements avec des valeurs aberrantes sont retrouvés, il est suggéré d'essayer de retrouver d'autres enregistrements du même patient pour les corriger si possible. Pour cela, il faut établir des contraintes sur les attributs susceptibles de subir des saisies aberrantes. Exemples :

- le sexe peut prendre que deux valeurs M ou F. S'il y a des enregistrements avec des saisies genre masculin, masc., fém., féminin, m, f, etc., elles doivent être transformées en M ou F pour garder la consistance ;
- le jour de naissance doit être en entier compris entre 1 et 31, le mois en entier compris entre 1 et 12, l'année entre 1900 (pour une année inférieure il faut faire une vérification) et l'année actuel, que le mois de naissance peut avoir le nombre de jours indiqué dans le jour de naissance (exemple : le 30 février n'existe pas) ;

Après cette étape, une analyse sur la présence de doublons dans la base de données PMSI est entreprise. Ces doublons peuvent apparaître au début des séjours hospitaliers si le patient n'est pas détecté comme étant déjà enregistré dans le système informatique (exemple : le patient arrive aux urgences et il déclare que c'est la première fois qu'il vient à l'hôpital même si ce n'est pas le cas). La non identification du patient implique la création d'une nouvelle entité dans le système informatique avec un nouveau numéro d'identification. L'apparition des doublons implique que le suivi de l'historique des patients est remis en question et on peut s'interroger sur la véridicité des informations, extraites et fournies aux médecins, issues de l'utilisation des bases de données susceptibles de présenter des problèmes de qualité.

Résultats Biologiques				
No_Séjour	Mesures	Date_Prise_Mesures	Durée_Examen	Examen

TABLE 8.3: La table Résultats Biologiques

Les doublons peuvent aussi apparaître à cause d'erreurs de frappe ou d'utilisation d'abréviations (exemple : abréviation du prénom Jean en J.). De plus, les noms et les prénoms composés peuvent induire des inconsistances qui doivent être corrigées. Par exemple, il faut vérifier les fautes de genre Jean Luis au lieu de Jean-Louis ou encore Leroux au lieu de Le Roux. Afin d'identifier les doublons il est nécessaire de procéder à une comparaison des chaines de caractères en utilisant des mesures de similarité :

- **la distance de Hamming** : adaptée pour les attributs numériques (numéro de sécurité sociale, numéro d'identité national, numéro de rue, numéro de téléphone, code postale, âge, etc.) ;
- **la distance de Jaro** : étant données deux chaînes de caractères s_1 et s_2, de longueurs respectives L_1 et L_2, ayant C caractères en commun et T transpositions de caractères (adaptée pour des chaines courtes) :

$$Jaro(s_1, s_2) = \frac{1}{3}\left(\frac{C}{L_1} + \frac{C}{L_2} + \frac{2C - T}{2C}\right) \tag{8.1}$$

- **la distance de Jaro-Winkler** : Si P est la longueur du plus long préfixe commun entre s_1 et s_2 :

$$JWinkler(s_1, s_2) = Jaro(s_1, s_2) + max(P, 4) * k * (1 - Jaro(s_1, s_2)) \; k \in (0, 0.25) \tag{8.2}$$

- **Soundex** : Il s'agit de faire un encodage des chaines en gardant la première lettre suivie par 3 chiffres codant les consonnes qui suivent la première lettre :

$$B, F, P, V \rightarrow 1; C, G, J, K, Q, S, X, Z \rightarrow 2; D, T \rightarrow 3; L \rightarrow 4; M, N \rightarrow 5; R \rightarrow 6$$

b) L'étude des résultats biologiques

Cette base de données contient les attributs présentés dans le tableau 8.3. Cette table a pour clé primaire et aussi pour clé étrangère l'attribut No_Séjour.

Dans ce cas, la qualité de données est influencée par de possibles inconsistances de référencement. Ces problèmes apparaissent à cause des évolutions technologiques, c'est-à-dire les remplacements des anciens appareils de prise de résultats, par des appareils de nouvelles technologies. Une solution pour limiter les effets d'un mauvais référencement est d'enregistrer, pour chaque observation, le type de technologie dont elle est issue.

Aussi, comme il s'agit de mesures qui sont prises par des capteurs, il faut que chaque enregistrement soit accompagné d'une description de la précision du capteur. Cette précision peut être présentée sous la forme d'un intervalle de la forme $\pm x\%$.

Comme le nombre d'enregistrements de cette base de données est très important, il est important de quantifier le volume de données à traiter afin de pouvoir estimer le temps de traitement nécessaire. Pour cela, il faut connaître le nombre d'enregistrements et le nombre et le type (entier, réel, string, booléen, date, etc.) de chaque attribut. Aussi, comme il s'agit d'une base de données dynamique, avec des entrées quotidiennes, il faut savoir la fréquence moyenne d'apparition de nouveaux enregistrements.

b) L'étude du DMI et des Molécules Onéreuses

D.M.I.				
Nom	Code_Disp	Date_Sortie	Date_Impl	No_Séjour

TABLE 8.4: La table de la base de données des Dispositifs Médicaux Implantables

M.O.				
Nom	Code_MO	Date_Sortie	Date_Administration	No_Séjour

TABLE 8.5: La table de la base de données des Molécules Onéreuses

La base de données des dispositifs médicaux implantables est décrite dans le tableau 8.4. Cette table a pour clé primaire et aussi pour clé étrangère l'attribut No_séjour.

La base de données des molécules onéreuses est présentée dans le tableau 8.5. Cette table a pour clé primaire et aussi pour clé étrangère l'attribut No_séjour.

Comme le principal problème dans ce cas est l'inconsistance des dates de l'envoie, de la réception et de l'implémentation/administration des DMI ou des MO, il faut trouver ces inconsistances et les éliminer.

Les mesures qui devront été faites consistent en la comparaison des dates :

$$\text{DateEnvoi} > \text{DateReception} > \text{DateImplémentation} \tag{8.3}$$

8.2.2 L'étude de la qualité de l'entrepôt

La construction de l'entrepôt de données fait apparaître une intégration (fusion, agrégation) de plusieurs sources hétérogènes représentant des caractéristiques de la même entités physique qui, dans ce cas particulier est le patient. À la fin du processus d'intégration, il faut obtenir une représentation unique, consistante et sans erreurs de l'entité physique. Pour que cette intégration soit faite correctement, il faut commencer avec deux analyses :

- une analyse des attributs (schéma mapping) pour identifier les attributs en commun ;
- une analyse des objets (identification d'objets) pour trouver les objets communs.

Par conséquent, il est nécessaire de prendre en compte les relations entre les bases de données. Le principal problème de qualité dans le cas de l'entrepôt est l'incohérence entre les bases de données devant être intégrées.

Pour analyser la qualité de données contenues dans l'entrepôt, il faut connaître le processus de construction de celui-ci. D'habitude, l'intégration de données contenues dans plusieurs bases de données se fait grâce à différents processus de traitement P_{ij}, afin de réduire l'hétérogénéité des données et d'augmenter ainsi la qualité intrinsèque de données.

La construction de l'entrepôt correspondant à l'intégration des bases de données décrites au début de cette étude : le PMSI, les Dispositifs Médicaux Implantables (D.M.I), les Molécules Onéreuses (M.O.) et les Résultats Biologiques (R.B.) est présentée dans la figure 8.3. Parmi les traitements décrits par les processus P_{ij} il y a :

- la suppression des doublons ;
- la subdivision des attributs (séparation de l'adresse en numéro, rue, ville, code postale, du nom en nom et prénom, etc.) ;
- la vérifications des contraintes d'intégrité (la date de début du séjour doit précéder la date de fin du séjour, la suppression des valeurs aberrantes, etc.) ;
- le remplissage des champs qui n'ayant pas de valeur (en utilisant des enregistrements sur le même patient et qui contiennent cette valeur).

En analysant les transformations de données en sortie de chaque processus, les changements de qualité de données peuvent être étudiés. Par conséquent, l'analyse de la qualité de données

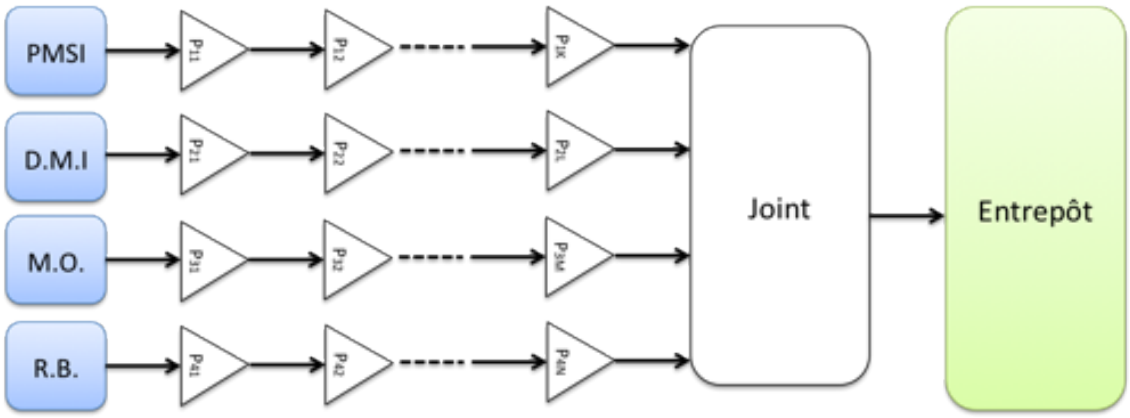


FIGURE 8.3: La construction de l'entrepôt à partir des plusieurs sources de données

contenues dans l'entrepôt sera faite à partir de la qualité des données contenues dans les quatre bases de données et en analysant les changements de qualité à travers les processus de traitement.

La construction de l'entrepôt de données doit fournir une augmentation de la complétude (en utilisant plusieurs sources de données représentant la même entité) et de la concision (en éliminant toutes les données redondantes).

Une dimension de qualité très importante est la complétude. Pour la mesurer, on peut faire une analogie avec la mesure de rappel (du couple précision/rappel). Ainsi la complétude décrit la quantité de données, en termes du nombre de n-uplets (complétude extensionnelle) et du nombre d'attributs (complétude intentionnelle) [Bleiholder 09] :

$$Compl_{Ext} = \frac{||\text{objets uniques dans le dataset}||}{||\text{tous les objets uniques dans l'Univers}||} \tag{8.4}$$

$$Compl_{Int} = \frac{||\text{attributs uniques dans le dataset}||}{||\text{tous les attributs uniques dans l'Univers}||} \tag{8.5}$$

En faisant une analogie avec la précision, la mesure de concision exprime l'*unicité* des objets représentés dans le dataset. Comme dans le cas de la mesure de complétude il existe une concision extensionnelle et une concision intentionnelle [Bleiholder 09] :

$$Concis_{Ext} = \frac{||\text{objets uniques dans le dataset}||}{||\text{tous les objets uniques dans le dataset}||} \tag{8.6}$$

$$Concis_{Int} = \frac{||\text{attributs uniques dans le dataset}||}{||\text{tous les attributs uniques dans le dataset}||} \tag{8.7}$$

Le problème le plus difficile à traiter dans la construction de l'entrepôt de données est la gestion des conflits et donc, implicitement, des inconsistances. Les conflits peuvent apparaître sous la forme de :

- **conflit schématique** : différents nom d'attributs, bases de données structurées différemment ;
- **conflits d'identité** : de la façon d'identifier les entités dans les sources de données ;
- **conflits de données** : valeurs différentes pour les attributs caractérisant la même entité. Ce type de conflit peut être divisé en deux catégories : incertitudes vis-à-vis de la valeur de l'attribut à cause du manque de données et contradictions à cause des différentes valeurs pour le même attribut.

La gestion des conflits entre les différentes bases de données qui doivent être intégrées peut se faire en adoptant une des ces stratégies : l'ignorance des conflits, la prévention des conflits ou la résolution des conflits.

8.2.3 L'étude de l'extracteur d'informations

L'avantage principal de l'utilisation du processus de fouille de données après la construction de l'entrepôt est que les processus de nettoyage et l'intégration de données sont déjà réalisés et donc la fouille de données est responsable seulement de l'extraction d'informations utiles, travaillant dans un environnement adapté.

Les informations issues du module d'extraction d'informations sont directement proposées à l'utilisateur final. Ainsi, la qualité de ces informations va se rapporter directement à la satisfaction de l'utilisateur final, c'est-à-dire le médecin. Par conséquent, la façon de communiquer entre le système et l'utilisateur (l'interface homme-machine) est très importante, mais elle ne sera pas traitée lors de cette thèse. En ce qui suit, les informations extraites par le module d'extraction d'informations seront analysées.

Dans le cas le plus général, l'extraction d'information est un processus de fouille de données chargé de fournir à l'utilisateur final des informations utiles. Le plus souvent, les informations demandées par l'utilisateur (le médecin) sont des informations sur un patient spécifique ou des informations sur les patients qui ont manifesté des symptômes spécifiques ou identiques.

Une première dimension de qualité est la précision temporelle, c'est-à-dire l'obsolescence, qui décrit l'exactitude de données par rapport à l'instant qu'elles sont censées représenter [Di Ruocco 12]. Ainsi, si par exemple la situation d'un patient est demandée, il faut que les informations fournies à l'utilisateur contiennent tout l'historique du patient sans oublier les dernières évolutions. Une autre demande peut être représentée par la situation médicale d'un patient à la fin du dernier séjour. Dans ce cas, il faut que seulement ces informations lui soient aussi présentées.

Un problème important pour les utilisateurs finaux est le temps d'attente pour obtenir une réponse. Cette deuxième dimension de qualité, d'*accessibilité*, décrit l'ergonomie du système d'information. Il est conseillé que l'utilisateur final ait accès aux informations voulues sans faire beaucoup de click de souris ou des saisies de données (ainsi que les enchaînements fastidieux d'écran sont à éviter).

La procédure classique d'évaluation des informations fournies à l'utilisateur est d'utiliser le couple de mesures : précision / rappel.

$$\text{Précision} = \frac{||DocRelevants \cap \text{DocRetrouvés}||}{||\text{DocRetrouvés}||} \tag{8.8}$$

$$Rappel = \frac{||DocRelevants \cap \text{DocRetrouvés}||}{||DocRelevants||} \tag{8.9}$$

8.3 SYSTÈME D'AIDE AU CODAGE DES ACTES MÉDICAUX ET DES DIAGNOSTIQUES

Une partie du système d'information hospitalier, correspondant à une application ayant pour objectif d'aider les médecins dans le processus de codage médical, est maintenant présentée. En effet, depuis l'adoption de l'ordonnance du 24 avril 1994 sur la réforme de l'hospitalisation en France, chaque acte et diagnostique médical doit être codé et associé à un patient. Cependant, ce processus de codage n'est pas évident à cause du nombre très grand de codes existants [1] et des ressemblances entre eux. Ainsi, très souvent, les médecins le considère comme ennuyeux [Lecornu 09a]. De plus, les erreurs de codage peuvent conduire à des statistiques faussées et, en conséquence, à des bilans

1. La classification commune des actes médicaux (CCAM) décrit plus de 7600 procédures médicales et le standard ICD-10 contient plus de 18000 diagnostique.

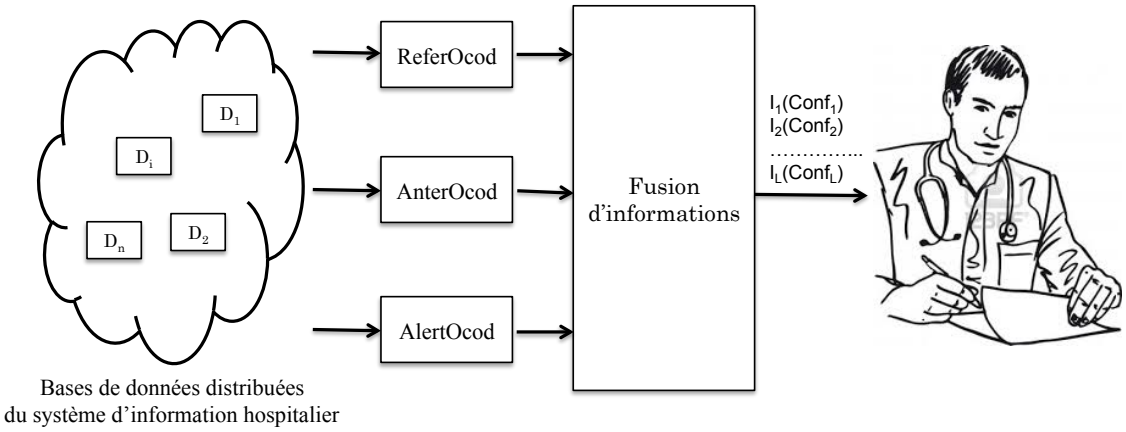


FIGURE 8.4: Architecture d'un système d'aide au codage des actes médicaux

financiers erronés. Ainsi, un système d'aide au codage médical est d'une très grande utilité pour tout médecin.

Comme dans le cas de l'application précédente, paragraphe 7.2, ce système va servir pour la validation de la méthodologie d'évaluation de la qualité de l'information. Par la suite, les trois étapes de cette méthodologie seront décrites.

8.3.1 Évaluation de la qualité locale

La première étape de la méthodologie, correspondant au chapitre 4, propose de décomposer le système en ses modules de traitement élémentaires. L'architecture simplifiée du système d'aide au codage médical est présentée dans la figure 8.4. Trois modules, ayant un comportement similaires aux classifieurs, sont utilisés pour l'extraction des informations sur les codes médicaux possibles. Puis, un module de fusion collecte les informations fournies par ces modules et fournit l'information finale à l'utilisateur. Par la suite, chacun des ces modules de traitement élémentaires sera décrit plus en détail en mettant l'accent sur l'évaluation de sa qualité locale.

8.3.1.1 ReferOcod

Le premier module d'extraction d'information est nommé **ReferOcod** [Lecornu 09a]. Il a pour rôle de prédire les codes diagnostiques d'un patient en utilisant comme données d'entrée l'âge, le sexe, la durée de son séjour hospitalier, les diagnostiques et les actes médicaux précédents. Ce données sont extraites à partir de bases de données volumineuses contenant les résumés standardisés et anonymisés de sortie de patients. En sortie, ce module de traitement fournit une liste de codes obtenus suite à des prédictions probabilistes [Lecornu 09a]. Figure 8.5 présente une interface, développée lors du projet Medidex [2], d'un exemple de résultat fourni par ReferOcod. C'est une liste de douze codes accompagnés par la description du diagnostique médical.

La qualité de données reçues à son entrée peut être évaluée par trois critères de qualité :

- *Précision* : mesurée par le taux entre le valeurs correctement enregistrées et le nombre total de valeurs ;
- *Consistance* : définie dans le paragraphe 8.2.1 ;
- *Obsolescence* : mesures pour ce critère de qualité ont été présentées lors du modèle de Wang et Strong (paragraphe 3.3.1).

2. Med1dex est un projet financé par l'Agence Nationale de la Recherche au travers du projet TECSAN, n° ANR-07-TECSAN-013-02 qui associe le CHRU de Brest, TELECOM-Bretagne et la société PRISMEDICA.

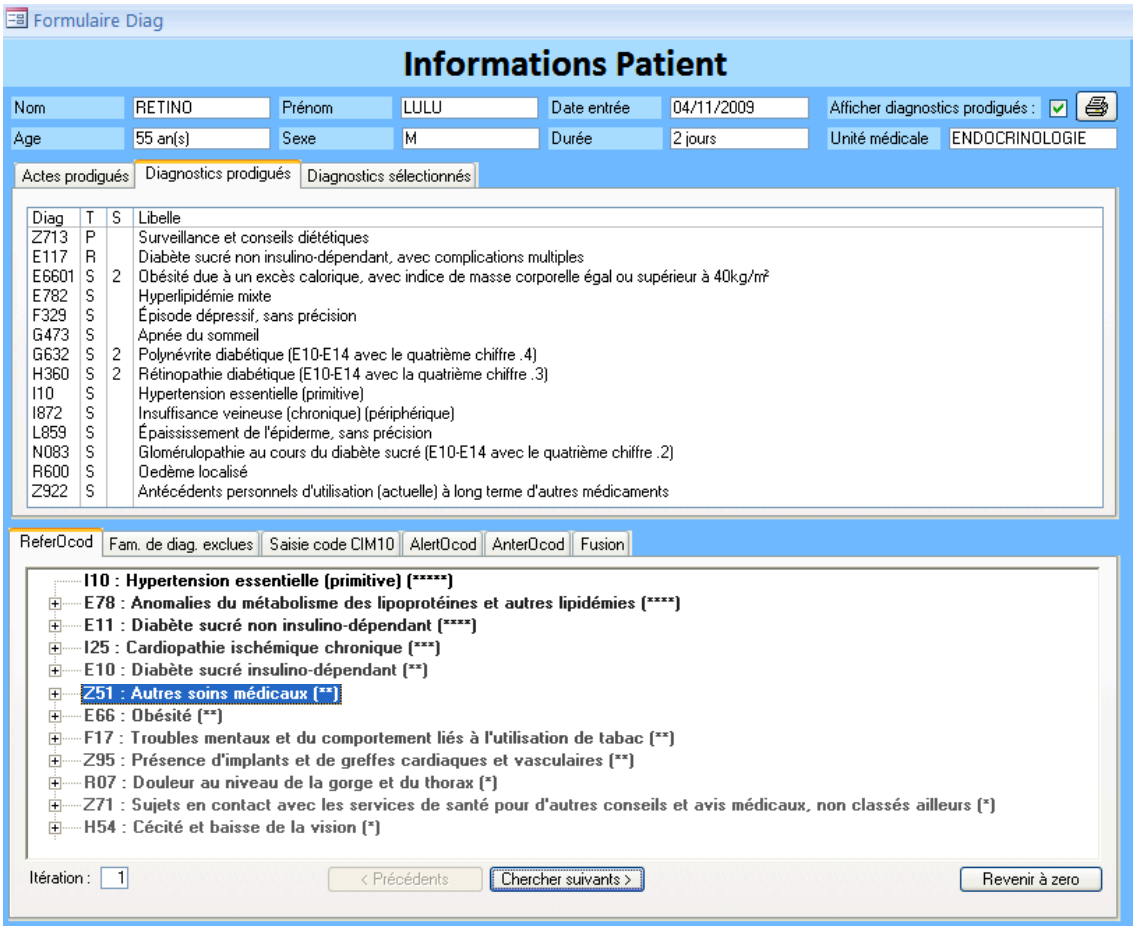


FIGURE 8.5: Exemple de liste de codes fournis par le module d'extraction d'information ReferOcod

La qualité en sortie du module ReferOcod est quantifiée par le critère de qualité *Confiance*, qui est mesuré par les probabilités des diagnostiques proposés ($\hat{P}(D_j)$).[Lecornu 09a] propose pour l'évaluation de ces probabilités d'utiliser quatre sources de données :

1. l'âge, le sexe et la durée du séjour : $\hat{P}(D_j|\text{âge, sexe, durée du séjour})$;
2. l'unité médicale : $\hat{P}(D_j|\text{UM})$;
3. les procédures médicales utilisées : $\hat{P}(D_j|\text{proc}_1, \dots \text{proc}_N)$;
4. les diagnostiques médicaux déjà codés : $\hat{P}(D_j|D_1, \dots D_M)$

En final, la probabilité du diagnostique j est donnée par la combinaison de ces quatre estimations :

$$\begin{aligned}\hat{P}(D_j) \quad &= \beta_1 \hat{P}(D_j|\text{âge, sexe, durée du séjour}) \\ &+\beta_2 \hat{P}(D_j|\text{UM}) \\ &+\beta_3 \hat{P}(D_j|\text{proc}_1, \dots \text{proc}_N) \\ &+\beta_4 \hat{P}(D_j|D_1, \dots D_M)\end{aligned} \tag{8.10}$$

avec $\beta_i \in [0, 1]$ et $\sum_{i=1}^{4} \beta_i = 1$. Les valeurs des poids β_i sont dépendantes des performances de chaque source de données et donc, elles sont estimées expérimentalement.

8.3.1.2 AnterOcod

Le deuxième module d'extraction d'information est **AnterOcod** [Lecornu 10]. Il réalise une estimation de la récurrence des maladies. Ainsi, il essaie trouver parmi les codes précédents, ceux qui sont susceptibles d'avoir une occurrence temporelle. C'est le cas des maladies relativement chroniques, qui ont déjà été très probablement codées lors des séjours précédents, et qui peuvent être proposées comme des codes pertinent pour le séjour actuel.

Ainsi, les données utilisées à son entrée sont identiques à celles du cas du module ReferOcod, c'est-à-dire les résumés standardisés et anonymisés de sortie de patients, stockés dans des bases de données spécifiques. Par conséquent, la qualité en entrée est décrite par les mêmes critères que dans le cas du module ReferOcod : la *Précision*, la *Complétude* et l'*Obsolescence*.

En sortie, ce module propose également une liste de codes. Ces codes sont estimés à l'aide de modèles mathématiques faisant appel aux courbes de survie actuarielles, comme par exemple l'estimateur de Kaplan-Meyer [Collett 97]. Grâce à ces outils mathématiques, à chaque maladie chronique est associé un taux de rappel, constituant une base de connaissances. De plus, les codes associés au patient dans les deux dernières années sont également utilisés, afin de construire la liste finale des codes proposés par le module AnterOcod. La qualité de ces informations est quantifiée par le critère *Confiance* et il est mesuré par le taux de rappel de chaque code [Lecornu 10].

8.3.1.3 AlertOcod

Le troisième module d'extraction d'information est **Alertocod** [Lecornu 09b]. Il utilise les données, issues d'examens de laboratoire et d'autres indicateurs de la condition du patient, se retrouvant dans des bases de données distribuées dans le cadre du système d'information hospitalier. La qualité de ces données est quantifiée par les mêmes critères que pour les autres deux modules d'extraction d'informations.

AlertOcod réalise trois fonctionnalités principales [Lecornu 09b] :

1. Extraction de règles décrivant la condition du patient et la nature des résultats de laboratoire la caractérisant. Grâce à celles-ci, une liste de codes est proposée.
2. Notification de l'utilisateur par des alertes indiquant qu'un groupe de faits soutient une règle appliquée pour un cas spécifique d'un patient. Ainsi, des codes sont identifiés et leur pertinence est expliquée.

3. Management automatique de l'interface entre les différentes sources de données.

Chaque code proposé par ce module est accompagné d'une valeur d'importance, qui est une mesure de qualité décrivant la confiance que l'utilisateur peut avoir dans ce code. La représentation choisie pour ces valeurs d'importance est dans un format linguistique, par exemple {*très rare, rare, souvent*}.

8.3.1.4 Fusion d'informations

Les trois modules d'extraction d'informations, ReferOcod, AnterOcod et Alertocod, fournissent trois listes différentes de codes. Comme chaque module d'extraction d'information est spécialisé dans l'identification de codes pour des cas spécifiques, chacune de ces listes contient une information partielle sur la vraie liste de codes. Ainsi, le module de fusion d'information a pour rôle de collecter ces informations partielles et de fournir une liste unique contenant les codes les plus *plausibles*.

Le tableau 8.6 présente un exemple de listes de codes à fusionner. Ainsi, on peut voir que les premières deux sources d'information fournissent des valeurs numériques dans l'intervalle unitaire mais avec des significations différentes : S_1 exprime la confiance dans les diagnostiques par des probabilités et S_2 par des taux de rappel. En même temps, la troisième source exprime la mesure de confiance à l'aide d'une valeur linguistique.

Source	**Signification**	**Diagnostiques/ValConf**		
ReferOcod	Probabilités	$D_1/0.5$	$D_2/0.4$	$D_3/0.8$
AnterOcod	Taux de rappel	$D_1/0.4$	$D_2/0.5$	$D_4/0.7$
AlertOcod	Val. Linguistiques	D_1/souvent	D_2/rare	

TABLE 8.6: Exemple de listes de codes fournis par les trois sources d'extraction d'informations

À cause de leur hétérogénéité, de signification et de représentation, la fusion d'informations présentées dans le tableau 8.6 n'est pas évidente. Une solution possible à ce problème de fusion est de réaliser une transformation des valeurs de confiance dans un domaine commun, c'est-à-dire essayer d'homogénéiser les données. Dans [Todoran 11], nous avons proposé de passer dans le domaine des possibilités. Ce choix est expliqué, dans un premier temps, par la facilité d'association des mesures de possibilité aux valeurs linguistiques. Dans un deuxième temps, cette théorie offre un cadre très flexible de construction d'opérateurs de fusion, voir l'annexe B.5 pour quelques exemples.

En conclusion, la confiance en sortie du module de fusion sera donnée par le résultat d'une fusion possibiliste.

La qualité en sortie du module de fusion est évaluée par les critères de qualité indiqués ci-dessous :

- *Correction* : proportion de codes qui sont corrects dans la liste finale :

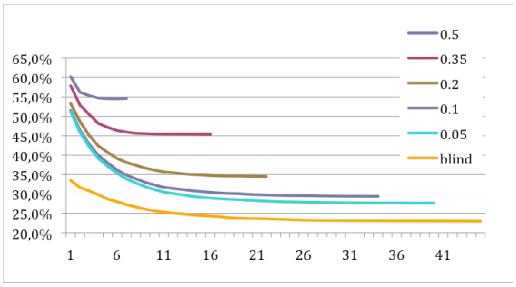
$$Cr = \frac{tp}{tp + fp} \tag{8.11}$$

- *Complétude* : proportion de codes qui sont corrects par rapport à l'ensemble des codes qui devraient être présentés :

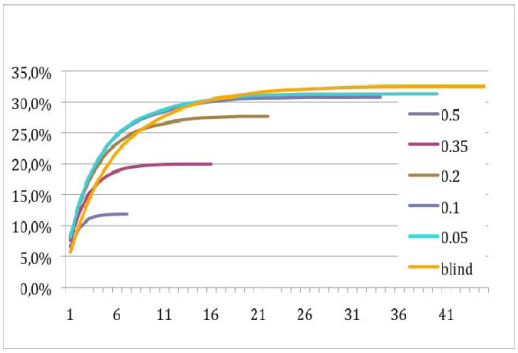
$$Cm = \frac{tp}{tp + fn} \tag{8.12}$$

- *Confiance* : mesurée par un degré de confiance propre à chaque code. Comme il a été expliqué ci-dessus, cette mesure de confiance est donnée par le résultat d'une fusion de valeurs de confiance des codes en entrée du module.

avec tp les codes correctement identifiés, fp les codes incorrectement identifiés et fn les codes correctes mais qui n'ont pas été proposés.



(a) Courbes de précision pour cinq seuil différent et le cas de proposition aveugle



(b) Courbes de rappel pour cinq seuil différent et le cas de proposition aveugle

FIGURE 8.6: Précision et rappel pour le module d'extraction d'informations ANTEROCOD

8

8.3.2 Construction de la fonction de transfert de la qualité

Dans ce paragraphe, la deuxième étape de la méthodologie d'évaluation de la qualité de l'information est présentée, correspondant au chapitre 5. Il s'agit de déterminer une fonction de transfert de la qualité pour chaque module de traitement élémentaire.

Les trois modules d'extraction d'informations ont été développés et validés en utilisant un jeu de données test issues de vrais résumés de patients standardisés et anonymisés. Ainsi, dans les articles de référence, leurs caractéristiques statistiques peuvent se retrouver. Nous présentons à titre d'exemple l'évaluation statistique du module AnterOcode dans la figure 8.6. Cette évaluation a été réalisée sous la forme des courbes précision-rappel dans le cas de cinq seuils différents. En effet, comme indiqué précédemment, chaque code est caractérisé par un taux de rappel en fonction de l'écart temporel entre la dernière occurrence de ce code et le séjour actuel. Lorsqu'un de ces taux dépasse le seuil imposé, il est ajouté à la liste de codes proposés au médecin. Toujours dans la figure 8.6, il est représenté le cas où tous les codes disponibles sont proposés sans aucun classement (proposition aveugle).

Grâce à ces études antérieures, il n'est pas nécessaire d'estimer la fonction de transfert de qualité pour les modules d'extraction d'informations, car elles ont déjà été déterminées pour chaque module.

L'étude de la qualité de l'information fournie par le module de fusion d'informations a été faite dans un autre article, [Puentes 13].

8.3.3 Évaluation de la qualité globale du système

Dans ce paragraphe, la troisième étape de la méthodologie d'évaluation de la qualité de l'information est présentée, correspondant au chapitre 6. Il s'agit de l'évaluation de la qualité globale du système en utilisant la qualité locale, définie dans le paragraphe 8.3.1.

L'étude complet du système d'aide au codage médical est présentée dans la figure 8.7. Comme dans le cas de l'application précédente, paragraphe 7.2, les valeurs de la qualité locale sont illustrées après chaque module de traitement élémentaire.

Après avoir défini la qualité locale et déterminé les fonctions de transfert de la qualité, chaque changement de qualité est directement propagé à la sortie du système et donc, la qualité globale est automatiquement estimée. Comme la quantité de données médicales continue d'augmenter, il est envisageable d'ajouter d'autres modules d'extractions d'informations, spécialisés dans d'autres caractéristiques (figure 8.7). Si ce nouveau module est compatible avec le module de fusion d'information, cf. paragraphe 4.1, sa qualité en sortie sera également définie par le critère de qualité

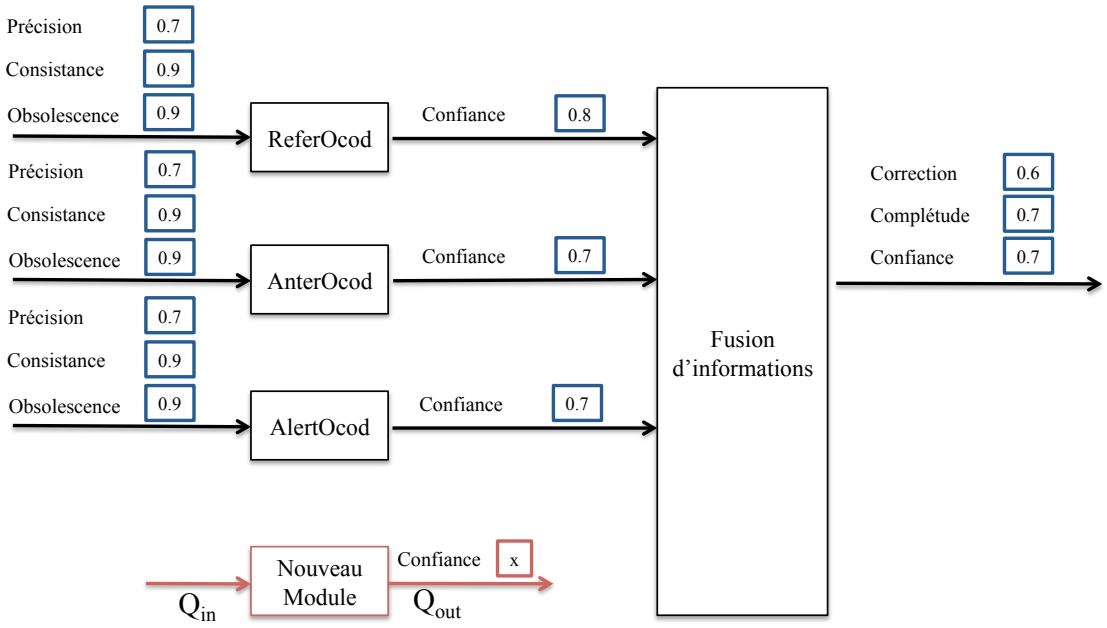


FIGURE 8.7: Analyse de la performance globale du système d'aide au codage d'actes médicaux

Confiance. Donc, il reste seulement à déterminer sa fonction de transfert de qualité et puis, l'évaluation de la qualité globale du système continuera à se réaliser d'une manière automatique.

8.4 CONCLUSION

Après une première application de notre méthodologie dans le cadre d'un système du domaine de la défense, dans ce chapitre nous avons choisi un système d'information hospitalier. À cause de sa grande complexité, nous n'avons pas procédé à l'étude de l'ensemble de ce système, mais nous avons choisi une application particulière, partie de ce système : un système d'aide au codage médical.

Ce système fait intervenir plusieurs sources de données, chacune étant hétérogènes par rapport aux autres. De plus, l'information finale est représentée par une liste de codes de taille variable en fonction du patient. Ainsi, la problématique générale est différente de celle du domaine de la défense.

La validation de notre méthodologie a été réalisée en appliquant les trois étapes de la méthodologie. Dans la première étape, la qualité locale pour chaque module élémentaire de traitement a été définie en sélectionnant les critères de qualité adaptés à l'utilisation. De plus, des mesures de qualité ont été proposées pour chacun des ces critères. Comme les performances de tous ces modules ont été analysées lors des études antérieures, les fonctions de transfert de qualité des modules n'ont pas dû être calculées.

Comme dans le cas du système du domaine de la défense, l'utilisation de la qualité locale et de la fonction de transfert de qualité ont permis d'estimer automatiquement la qualité des informations proposées au médecin, objectif de notre méthodologie.

« Trois années de thèse ne peuvent se conclure en une page, d'ailleurs, une thèse se conclue-t-elle vraiment un jour ? Ces travaux ne sont qu'une ouverture, un ensemble d'idées qui ne demandent qu'à survivre à leur auteur qui espère qu'elles ne s'endormiront pas au fond d'un vieux placard. »

Yannick Le Bras

QUATRIÈME PARTIE : CONCLUSION GÉNÉRALE ET PERSPECTIVES

TELECOM Bretagne

9 Conclusions et perspectives

9.1 Conclusions

Dans cette thèse, le travail de recherche réalisé est orienté vers la proposition d'une nouvelle méthodologie de définition et d'analyse de la qualité des données et de l'information d'un système d'information. À cause de la complexité de la notion de qualité, ainsi que de celles de données et d'information, ce travail se situe à un carrefour de plusieurs domaines de recherche très actifs : l'ingénierie des systèmes d'information, l'aide à la décision, traitement du signal et de l'image, la fusion d'informations, etc. La définition et donc, implicitement l'évaluation de la qualité des données et de l'information soulèvent également des questions philosophiques. En effet, des définitions pour les notions de données et d'information ont été proposées depuis l'Antiquité, sans cependant converger vers une définition généralement acceptable par la communauté scientifique. Par conséquent, les défis de ce doctorat ont été très importants.

Il n'est jamais évident de commencer un travail de recherche (ou d'une autre nature) sur un sujet qui connait une certaine maturité, induite par le fait que dans les dernières deux décennies un nombre important de questions (« pertinente » ou pas) ont été déjà posées, mais qui n'a pas connu de développements remarquables. Bien sûr, dans ce genre de situations il existe deux alternatives. La première est d'essayer restreindre les objectifs afin de pouvoir très clairement cadrer et identifier une question. Ensuite, des méthodes spécifiques sont identifiées et appliquées pour trouver une réponse, une solution à cette question. Malheureusement, l'inconvénient de cette approche est qu'elle n'assure pas que la solution trouvée est valide à l'extérieur du cadre de départ. La deuxième alternative est de commencer par le problème d'ensemble, en essayant de rester dans le cadre le plus général possible. L'avantage majeur de cette approche est l'applicabilité étendue des solutions proposées. Par contre, l'inconvénient est la difficulté de généraliser certains concepts et notions, avec l'interrogation permanente de savoir si une solution généralement applicable existe vraiment ou s'il s'agit d'une poursuite d'une fée Morgane [1].

Nous avons choisi la deuxième alternative. La raison principale de ce choix est la quasi-inexistence d'une telle approche dans la communauté scientifique concernée par les problèmes de qualité des données et/ou de l'information. En effet, la grande majorité des études de recherche sur ce sujet est partie d'un cadre restreint et se sont contentés de fournir des solutions à des problèmes provisoires. Ainsi, la plupart des solutions proposées ne sont plus d'actualité, car elles n'offrent des réponses que dans l'environnement, figé, dans lequel elles ont été développées, mais qui, naturellement, a évolué.

Il existe aussi des travaux de recherche traitant le sujet de la qualité des données et de l'information d'une perspective plus large. Le programme « Information Quality de MIT (MITIQ) » a débuté en ce sens au début des années 90. Malheureusement, la plupart des chercheurs impliqués dans ce programme sont consultants en management des systèmes d'information. Ainsi, la quasi-totalité des recherches ont été menées dans le cadre de ce type de système d'information.

1. Morgane, nom signifiant en breton « née de la mer » (Mor signifie « mer » et gane signifie « né ») est un personnage mythique et fabuleux du cycle arthurien, très souvent vue comme une sirène ayant des pouvoirs magiques et associée aux mirages

Néanmoins, le cadre générique de définition de la qualité des données et d'information proposé par Wang et Strong dans [Wang 96] est un résultat d'une valeur incontestable, utilisé depuis comme la principale référence sur ce sujet. Cependant, dans toutes les études partant du modèle de Wang et Strong, il manque l'utilisation d'outils mathématiques et même s'ils existent, ils ne sont pas rigoureusement développés. Dans le soutien de notre critique, nous citons [Borek 14] qui affirme catégoriquement que ces méthodes d'évaluation de la qualité sont proposées par des consultants en management n'ayant pas de solides connaissances en mathématique. De plus, il y a la recommandation surprenante de Thomas Redman, un des gourous du domaine de la qualité des données, qui propose (malheureusement un peu tard : 2013 !) d'appliquer la théorie de l'information de Shannon et la notion d'incertitude dans la définition et l'évaluation de la qualité des données et de l'information [Redman 13].

Ainsi, la conclusion est qu'il existe un écart important entre les différents domaines de recherche sur le sujet de la qualité de l'information. D'une part il y a la communauté du management qui propose des méthodologies suffisamment flexibles et exhaustives pour représenter presque tous les aspects de la qualité et de l'autre part il y a la communauté de chercheurs issus d'une formation en sciences et génie qui développent des outils mathématiques performants, mais pour un cadre restreint. Un exemple de ce dernier cas est la théorie de Shannon qui a été développée dans le cadre des communications numériques en 1946 et qui a pour objectif d'assurer une transmission des données en gardant une certaine qualité.

À partir de cette conclusion, également exprimée à la fin de la première partie de cette thèse, il nous est clair qu'il est impératif de rapprocher ces deux visions sur la qualité des données et de l'information. De plus, nous avons également identifié la nécessité de se situer dans un cadre agnostique par rapport au domaine d'application. Par conséquent, l'objectif de ce doctorat a été de développer une nouvelle méthodologie, toute en réutilisant les développements existants, d'évaluation de la qualité des données et de l'information dans un contexte indépendant du domaine d'application.

La première partie de ce doctorat a été consacrée à la définition des principaux concepts intervenant dans le développement de cette méthodologie dans le chapitre 1 et à l'état de l'art de la qualité des données et de l'information, respectivement, chapitres 2 et 3. Par rapport aux autres études, nous avons considéré que la notion de données n'était pas équivalente à celle d'information. En effet, comme la qualité est définie par rapport aux propriétés voulues de l'entité étudiée, nous sommes persuadés qu'il est nécessaire de discerner entre ces deux notions car leurs propriétés sont différentes. Par conséquent et comme une nouveauté par rapport aux autres études, nous avons commencé en proposant des définitions pour les notions de données, d'information et de connaissance. Les différences entre ces trois notions sont souvent très subtiles, raison pour laquelle nous avons opté pour une définition mettant en évidence leurs interconnexions et leurs propriétés sous la forme d'un enrichissement sémantique partant des données, en passant par l'information, pour finalement aboutir à la connaissance. Le contexte générale de définition de ces notions a été celui des systèmes d'information complexes utilisés dans le processus d'aide à la décision. Grâce à ces définitions nous fournissant les propriétés/dimensions à évaluer, la définition de la qualité des données et de l'information est rendue plus compréhensible et plus accessible.

L'état de l'art de la qualité des données et de l'information a été présenté dans deux chapitres différents pour bien les distinguer. En ce qui concerne la définition et l'évaluation de la qualité des données, nous avons pris la décision de ne considérer que les données stockées dans des bases de données relationnelles et les données issues de capteurs. Même si l'étude de la qualité des données est un sujet de recherche *in sine*, assez vieux, il était possible de ne le présenter que brièvement dans une annexe. Nous avons identifié deux visions complémentaires qui méritaient d'être regroupées :

- La première vision correspond au domaine du management des bases de données et représente la qualité des bases de données et de l'entrepôt de données ;
- La deuxième vision considère la qualité des données rapportée aux imperfections dont les données y sont susceptibles.

Quant à l'état de l'art sur la qualité de l'information, il a fait ressortir de nombreuses confusions

entre la qualité des données et celle de l'information au niveau des dimensions adaptées à chacune. De plus, la qualité de l'information n'est analysée qu'en sortie du système d'information, sans s'intéresser aux processus internes aux systèmes le générant. Ainsi, les systèmes d'information sont considérés comme des boîtes noires et la qualité est évaluée en entrée du système (qualité des données) et en sortie du système (qualité de l'information). Mais, pour l'utilisateur final, une information sans l'accompagnement d'un niveau *expliqué* de qualité n'est pas acceptable parce qu'il ne peut pas l'interpréter et l'incorporer dans son processus cognitif de prise de décision.

Par conséquent, comme solution à ce problème, nous avons proposé, dans la deuxième partie de cette thèse, une nouvelle méthodologie d'évaluation de la qualité de l'information. Proposer à l'utilisateur une qualité expliquée de l'information implique de connaître la provenance de cette qualité, c'est-à-dire les processus qui l'ont influencée. Pour cela, nous avons proposé d'exploiter l'architecture interne du système d'information afin de pouvoir analyser l'évolution de la qualité à travers du système. Suite à cette décomposition, la qualité peut être analysée à deux niveaux :

- **Qualité locale** : en entrée et en sortie de chaque module de traitement élémentaire ;
- **Qualité globale** : en sortie du système d'information, c'est-à-dire la qualité de l'information proposée à l'utilisateur ;

Notre méthodologie d'évaluation de la qualité se déroule en trois étapes :

1. Dans la première étape, la qualité locale en sortie de chaque module est définie en choisissant les critères et les mesures de qualité adaptés.
2. Dans la deuxième étape, l'influence sur la qualité de chaque module de traitement est modélisée et une *fonction de transfert de qualité*, individuelle à chacun de ces modules, est définie. Cette fonction a pour rôle de réaliser la liaison, sous la forme d'une cartographie, entre la qualité en sortie du module et celle en entrée.
3. Dans la troisième étape, grâce à la qualité locale et au concept de fonction de transfert de qualité, la qualité de l'information en sortie du système d'information, c'est-à-dire la qualité globale, est automatiquement estimée.

En conclusion, la méthodologie proposée est fondée sur le principe *divide et impera*. Ce principe signifie que dans un premier temps la qualité locale est évaluée, en sortie de chaque module et que cette évaluation sera ensuite utilisée pour automatiquement estimer la qualité globale. Également, nous pouvons affirmer que notre méthodologie est simple, intuitive et facile d'implémentation. Ces caractéristiques, conjointement avec l'agnosticisme vis-à-vis du domaine d'application, sont les vrais atouts de notre méthodologie.

Bien sûr, comme dans le cas de nouveau concept, cette méthodologie a dû être validée. La quatrième partie de cette thèse est consacrée à ce sujet. À cause de l'absence d'*a priori* sur le domaine d'application, la validation ne peut pas se faire qu'en considérant au moins deux systèmes d'information provenant de deux domaines d'application différents. Nous en avons choisi deux : le premier est un système multi-capteurs de reconnaissance automatique de cibles radar du domaine de la défense et le deuxième est un système d'aide au codage médical, donc venant du domaine hospitalier.

Dans les deux cas, il s'agit d'un système d'information complexe avec des contraintes fortes sur les niveaux de qualité acceptable pour leurs utilisateurs. Le premier système a servi comme support à une évaluation complète en employant les trois étapes de notre méthodologie. Le deuxième système a servi d'exemple à la situation pour laquelle il existe une analyse préalable de certains composants du système. Dans ce cas, nous avons montré que ces résultats peuvent être incorporés dans notre méthodologie. Dans les deux cas d'application, les résultats, pour le moment limités à des simulations, sont encourageants et nous pouvons affirmer que la première étape de validation de notre méthodologie a été faite avec succès. Bien sûr, par la suite, il sera nécessaire de procéder à une validation dans le monde réel.

L'utilité de notre méthodologie est, dans un premier temps, pour les utilisateurs d'un système d'information qui doivent comprendre le processus d'évaluation de la qualité de l'information. Nous sommes persuadés que l'implémentation de cette méthodologie permettra aux utilisateurs de

mieux gérer le processus de prise de décisions. De plus, notre méthodologie est également d'une réelle importance pour un analyste du système, car l'évaluation de la qualité locale lui permet d'analyser les performances des modules de traitement.

En final, nous récapitulons les principaux avantages de notre méthodologie par rapport aux existantes. Notre approche :

1. propose de décomposer le système d'information en modules élémentaires afin de pouvoir exploiter son architecture interne pour la traçabilité de la qualité à l'intérieur du système ;
2. propose une nouvelle vision de la notion de qualité : qualité locale vs. qualité globale ;
3. n'essaie pas de remplacer les méthodologies existantes, mais elle les utilise et essaie de les valoriser sous cette nouvelle vision ;
4. propose une définition formelle de la qualité des données et de l'information permettant une implémentation facile dans la pratique ;
5. introduit le concept novateur de « fonction de transfert de qualité » permettant de modéliser l'influence du module de traitement sur la qualité ;
6. propage la qualité à travers le système d'information, et après avoir défini la qualité locale, la qualité de l'information en sortie du système est automatiquement évaluée ;
7. permet à l'utilisateur d'avoir une qualité individuelle et expliquée pour chaque information qui lui est proposée ;
8. nécessite qu'un calcul d'une fonction de transfert lors d'une mise à jour d'un module ou d'un ajout d'un nouveau module. Par conséquent, l'évolution du système (mise à jours, suppressions/ajout des modules) est facilement prise en compte.

Ce travail de thèse se voit comme une ouverture de nombreuses voies de recherche à explorer dans des travaux futurs. Dans le paragraphe suivant, je présente, avec regret et soulagement comme disait [Le Bras 11], mais dans l'espérance qu'un jour quelqu'un s'en planchera sur les quelques idées et suggestions méritant d'être développées, mais qui, à cause des diverses contraintes, ne l'ont pas été lors de ce doctorat.

9.2 Perspectives

L'aspect principal qui mérite un approfondissement est la représentation des critères de qualité par rapport aux caractéristiques des données ou de l'information. Lors de notre définition de la qualité locale, nous avons proposé une liste de critères de qualité pour caractériser les données et une autre pour l'information. Le raisonnement à la base de cette classification des critères a été représenté par l'adéquation des critères proposés dans la littérature de quantifier le contenu et la sémantique (les deux caractéristiques que nous avons considérées) des données et de l'information. À notre avis cette classification a été bien faite, mais nous ne l'avons pas prouvé. Une idée pouvant être exploitée est la représentation de la qualité non pas sous la forme d'une liste mais sous la forme d'une ontologie. Si le contenu d'une donnée ou d'une information peut plus ou moins facilement être caractérisé (nous avons proposé d'utiliser la valeur et le type de la valeur), la notion de sémantique est un peu plus difficile à caractériser. Je propose comme point de départ de sa caractérisation les travaux de Carnap et Bar-Hillel, [Carnap 52].

Un deuxième aspect méritant attention est le développement des mesures de qualité dans les bonnes théories mathématiques. Tout au long de cette thèse nous avons insisté sur la nécessité de quantifier les critères de qualité par des mesures de qualité et nous avons fait référence aux mesures développées dans diverses théories mathématiques, comme par exemple la théorie des probabilités, la théorie de Dempster-Shafer ou la théorie des possibilités. Mais comme ces cadres mathématiques présente un grand nombre de mesures possibles, il est nécessaire de faire une étude sur le choix des mesures adaptées. À mon avis, dans le processus de sélection de ces mesures vont intervenir les caractéristiques de l'information et les aspects quantifiés par le critère de qualité, cf. figure 9.1.

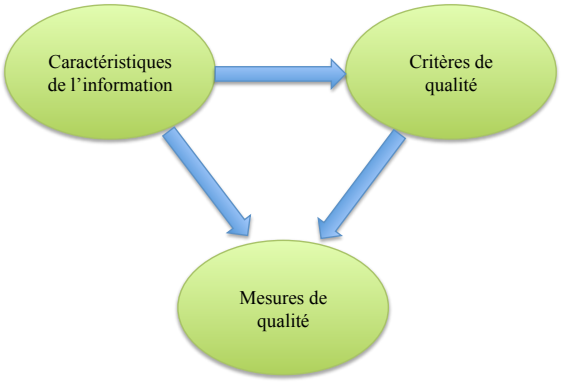


FIGURE 9.1: Les interactions entre les caractéristiques de l'information (ou équivalent des données), les critères de qualité et les mesures de qualité

Un troisième aspects méritant de futurs travaux et développements est la notion novatrice de fonction de transfert de la qualité. Lors de sa définition, nous avons proposé deux alternatives pour la déterminer : analytique ou non-analytique. Néanmoins, nous sommes persuadé qu'elle pourrait être mieux formalisée afin de la rendre plus facilement calculable.

Un dernier aspect qui n'a pas été traité lors de ce doctorat est celui de la visualisation de la qualité de l'information. La notion de qualité est complexe puisqu'elle est représentée par plusieurs critères de qualité, qui à leur tour peuvent être quantifiés par plusieurs mesures de qualité. Notre méthodologie permet la traçabilité de l'évolution entière de la qualité à travers le système. Par conséquent, il serait intéressant de pouvoir représenter la qualité dans un format visuel facilement compréhensible par l'utilisateur utilisant la sémiotique.

En final de cette thèse, je rappelle que l'évaluation de la qualité de l'information proposée aux utilisateurs est un problème multidisciplinaire. Par conséquent, un spécialiste dans ce domaine a besoin des connaissances transversales. Après le succès du programme sur la qualité de l'information de MIT, d'autres universités américaines ont commencé à proposer des programmes universitaires sur la qualité des données et de l'information. Un exemple récent est celui de l'Université d'Arkansas à Little Rock [2] proposant un programme de master sous la coordination de John Talburt. À l'Université Purdue il existe un autre programme orienté vers l'évaluation et l'amélioration de la qualité des données [Verykios 02]. Prenant comme exemple ces modèles, il existe des efforts, ces dernières années en Australie, pour former un groupe de recherche sur la qualité de l'information. Malheureusement en Europe, même si est reconnu comme important, la qualité de l'information, elle reste un sujet de niche quasi-inexistant, sans aucun programme de recherche ou de formation sur ce sujet. À mon avis, l'exemple américain doit être reproduit en Europe aussi.

2. Information Quality Graduate Program at University of Arkansas at Little Rock (UALR) : `http://ualr.edu/informationquality/masters/msiq-curriculum/`

Annexes

A

Les théories mathématiques de l'incertain

Soit Ω l'univers des ensembles. Dans le cas des ensembles classiques on peut définir une fonction caractéristique qui associe pour chaque élément $\omega \in \Omega$ l'appartenance de celui-ci au sous-ensemble $A \subseteq \Omega$:

$$\mu_A(\omega) = \begin{cases} 1 & \text{si} \quad \omega \in \Omega \\ 0 & \text{si} \quad \omega \notin \Omega \end{cases} \tag{A.1}$$

Ainsi, cette fonction peut prendre seulement des valeurs booléennes et les ensembles définis en utilisant cette fonction caractéristique sont appelés ensembles nets ou classiques.

> **Définition 22 :** Un ensemble flou est défini par une fonction ressemblante à une fonction caractéristique, appelée fonction d'appartenance. Chaque fonction d'appartenance définit un ensemble flou en désignant pour chaque élément le degré d'appartenance de ceux-ci dans l'ensemble flou. Pour un ensemble flou $\tilde{A} \subseteq \Omega$ la fonction d'appartenance se définit :
>
> $$\mu_{\tilde{A}} : \Omega \to [0, 1] \tag{A.2}$$

Observation : Chaque ensemble flou est complètement et uniquement défini par sa fonction d'appartenance. Ainsi il n'y a pas d'ambiguïté si on utilise pour la définition de la fonction d'appartenance la notation :

$$\tilde{A} : \Omega \to [0, 1] \tag{A.3}$$

L'utilisation des ensembles flous permet de caractériser les notions vagues, comme sont celles exprimées par le langage naturel (ex. petit, moyen, grand, etc.). Mais les notions exprimées en langage naturel sont dépendantes du contexte et donc il faut prendre des précautions dans le passage vers le numérique. En conclusion de l'utilité des ensembles flous on peut dire qu'ils permettent d'exprimer la transition de l'appartenance vers la non-appartenance d'une manière graduelle.

Une autre notion importante pour les ensembles flous sont les α-coupes [1]. Étant donné un ensemble flou $\tilde{A} \subseteq \Omega$ et $\alpha \in [0, 1]$, les α-coupes sont définies comme :

$${}^{\alpha}\tilde{A} = \{\omega \in \Omega | \tilde{A}(\omega) \geq \alpha\} \tag{A.4}$$

De cette définition des α-coupes on peut remarquer que les α-coupes sont des ensembles nets qui pour chaque ensemble flou forment un empilement d'ensemble :

$${}^{\alpha}\tilde{A} \subseteq^{\beta} \tilde{A} \quad ; \quad \text{pour} \quad \alpha > \beta \tag{A.5}$$

Chaque ensemble flou et uniquement représenté par l'empilement des α-coupes, plusieurs techniques pouvant être employées pour le passage α-coupes vers l'ensemble flou [Bloch 03], dont voici

1. En anglais α-cut

un exemple :

$$\tilde{A}(\omega) = \sup_{\alpha \in [0,1]} \min(\alpha, \tilde{A}(\omega)) \tag{A.6}$$

L'importance des α-coupes est le fait de pouvoir seuiller la fonction d'appartenance et de se retrouve avec un ensemble net qui pourra être utilisé dans une étape de prise de décision.

Une mesure monotone est définie sur une famille de sous-ensembles non-vides C appartenant à Ω : $g : C \to [0, \infty]$ et qui satisfait les propriétés :

- $g(\oslash) = 0$
- $\forall A, B \in C$, si $A \subseteq B \Rightarrow g(A) \leq g(B)$ (monotonicité)
- pour toute séquence croissante $A_1 \subseteq A_2 \subseteq \cdots \in C$

$$\text{si} \quad \bigcup_{i=1}^{\infty} A_i \in C \Rightarrow \lim_{i \to \infty} g(A_i) = g\left(\bigcup_{i=1}^{\infty} A_i\right)$$

- pour toute séquence décroissante $A_1 \supseteq A_2 \supseteq \cdots \in C$

$$\text{si} \quad \bigcap_{i=1}^{\infty} A_i \in C \Rightarrow \lim_{i \to \infty} g(A_i) = g\left(\bigcap_{i=1}^{\infty} A_i\right)$$

Pour tous deux sous-ensembles $A, B \in C$ avec $A \cap B = \oslash$ une mesure monotone peut exprimer les situations suivantes :

1. $g(A \cup B) > g(A) + g(B)$ traduisant une action constructive
2. $g(A \cup B) = g(A) + g(B)$ traduisant que A et B sont non-interactives vis-à-vis de la propriété mesurée. La théorie de probabilité peut exprimer seulement cette situation
3. $g(A \cup B) < g(A) + g(B)$ traduisant une incompatibilité entre A et B vis-à-vis de la propriété mesurée

Dans les situations où il y a une manque d'information, l'incertitude va générer une croyance et pas une connaissance. C'est pour cela qu'il faut pouvoir exprimer le type et la quantité d'incertitude dans un cadre mathématique bien défini.

Avec le développement des théories des ensembles flous et des mesures monotones le cadre pour la représentation de l'incertain est devenu plus général pouvant prendre en compte plusieurs types d'incertitudes. Ceci est devenu possible en remplaçant les ensembles nets par des ensembles flous et la théorie des mesures additives par la théorie de mesures monotones. Les deux changements ont les qualités d'être plus souples et moins restrictives.

A.1 La théorie des probabilités

La théorie des probabilités a été développée historiquement dans le cadre des ensembles nets, mais elle peut être formalisée à l'aide des axiomes. La notion centrale de la théorie des probabilités est la distribution de probabilité qui assigne une mesure de probabilité à chaque événement possible. Cette mesure de probabilité d'un événement $\omega \in \Omega$, sera notée par $P(\omega)$. La mesure de probabilité du même événement $\omega \in \Omega$ dans la présence d'une information (historique) H s'écrit $P(\omega|H)$. Elle s'exprime par un nombre et elle doit respecter les 3 axiomes suivants (convexité, additivité et multiplication) :

- $0 \leq P(\omega|H) \leq 1$, pour chaque événement $\omega \in \Omega$;
- $P(\omega_1 \cup \omega_2|H) = P(\omega_1|H) + P(\omega_2|H)$, pour chaque deux événements $\omega_1, \omega_2 \in \Omega$ qui sont mutuellement exclusifs ;
- $P(\omega_1 \cap \omega_2|H) = P(\omega_1|H) \cdot P(\omega_2|\omega_1 \cap H)$, où $P(\omega_2|\omega_1 \cap H)$ exprime la probabilité de production de l'événement ω_2 en sachant que l'événement ω_1 s'est produit. La probabilité $P(\omega_2|\omega_1)$ s'appelle probabilité conditionnelle de ω_2 en sachant ω_1.

Malheureusement cette suite d'axiomes n'indique pas quelle est la signification de la notion de probabilité. Ainsi on n'a pas une réponse sur quel type d'incertitude est représentée, sur le type et la taille de l'information H et sur la façon de déterminer la probabilité $P(\omega)$. Ces axiomes nous présentent, sous la forme de règles, la façon de combiner deux ou plusieurs événements incertains ou si des événements incertains sont consistants entre eux.

Dans l'histoire il y a eu plusieurs interprétations de la notion de probabilité, chacune présentant des moyens spécifiques de détermination des probabilités initiales. Parmi ces différentes visions, selon [Singpurwalla 02] les plus proéminent sont :

- la théorie classique (Cardano, Pascal, Fermat, Bernoulli, Bayes, Laplace, Poisson) ;
- la théorie de l'a priori ou la théorie logique (Keynes, Jeffreys, Ramsey) ;
- la théorie de la fréquence relative (Venn, Von Mises, Reichenbach) ;
- la théorie subjective (Borel, Ramsey, De Finetti, Savage).

Dans cet exposé, on va garder la théorie subjective des probabilités, pour les autres théories et pour des références voir [Singpurwalla 02]. Aussi dans cette référence il y a une comparaison entre la théorie basée sur la fréquence relative (probabilité objective) et la théorie des probabilités subjectives. L'idée centrale de cette théorie est l'hypothèse sur la nature subjective de la notion de probabilité. Le développement de cette théorie a été fait par De Finetti [Finetti 74]. La probabilité est vue comme un degré de croyance d'une personne donnée à un moment de temps donné. De Finetti a proposé une approche comportementale dans laquelle le degré de croyance s'exprime par une volonté de parier. Ainsi, comme exemple donné par De Finetti, la probabilité d'un événement est la quantité, p, qu'on est prêt à parier dans le cas d'un pari de deux variantes en échange d'un euro pour que l'événement se produit. Vice-versa on est disponible à parier $1-p$ pour que l'événement ne se produit pas. Comme dans cette théorie les valeurs de probabilités initiales sont évaluées par une personne, il n'y a pas des probabilités inconnues, correctes ou incorrectes. S'il existe des données *a priori* sur les événements d'intérêt (c'est-à-dire probabilités *a priori*), cette théorie va combiner ces données avec les données issues de l'expertise personnelle, les premières ayant pour rôle d'améliorer les dernières par les moyens de calcul des probabilités

L'utilité des probabilité subjective est très claire quand on doit utiliser des opinions d'experts, à cause d'une indisponibilité des données sur les événements d'intérêt. Ainsi, dans cette théorie l'expertise humaine se traduit par la détermination des probabilités. Dans le contexte actuel, les systèmes d'information fournissent des quantités importantes des données et d'informations, mais dans très peu de cas elles donnent directement les probabilités dont on a besoin. Par conséquent, un point de vue subjectif des probabilités peut engendrer un paradigme de quantification des incertitudes ensemble avec l'intégration des données et des informations pour aider l'utilisateur dans le processus de prise de décisions [Singpurwalla 02].

Une des méthodes les plus populaires de combinaison et de mise à jours des probabilités est la méthode de Bayes. Le théorème de Bayes permet de faire la liaison entre les probabilités de deux événements en termes de probabilités conditionnelles :

$$P(\omega_1|\omega_2 \cap H) = \frac{P(\omega_2|\omega_1 \cap H) \cdot P(\omega_1|H)}{P(\omega_2|H)} \tag{A.7}$$

Un autre théorème très important dans le calcul de probabilités est le théorème des probabilités totales. Cette théorème permet de calculer la probabilité d'un événement ω, comme la somme de toutes les probabilités conditionnelles de tout ensemble d'événements mutuellement exclusifs et exhaustifs :

$$P(\omega|H) = \sum_{(\forall)\ \omega_i} P(\omega|\omega_i \cap H) \cdot P(\omega_i|H) \tag{A.8}$$

Comme conséquence, le théorème de Bayes dans le cas d'un seul événement ω_i s'écrit :

$$P(\omega_i|\omega \cap H) = \frac{P(\omega|\omega_i \cap H) \cdot P(\omega_i|H)}{\sum_{(\forall)\ \omega_i} P(\omega|\omega_i \cap H) \cdot P(\omega_i|H)} \propto P(\omega|\omega_i \cap H) \cdot P(\omega_i|H) \tag{A.9}$$

À présent le théorème de Bayes a été définie dans le cas discret. Il peut l'être également défini dans le cas continu, faisant appel à des fonctions de densité de probabilité. Si on prend deux variables aléatoires X et Y de densité de probabilité $g(x;H)$ et respectivement $f(y;H)$, avec la densité de probabilité de Y dépendante de la variable X, le théorème de Bayes s'écrit :

$$g(x|y;H) = \frac{f(y|x;H)\cdot g(x;H)}{\int f(y|x;H)\cdot g(x;H)} \propto L(y|x;H)\cdot g(x;H) \tag{A.10}$$

Le dénominateur de cette équation est une constante de normalisation et comme conclusion on a que la densité de probabilité a posteriori est égale au produit entre la densité de probabilité a priori et une fonction appelée dans le domaine de la statistique, fonction de vraisemblance, $L(y|x;H)$. Ainsi, le théorème de Bayes réalise une combinaison de l'*a priori* et de la vraisemblance pour obtenir l'*a posteriori*.

A

A.1.1 Discussion sur la fonction de vraisemblance

Il est important de remarquer que la vraisemblance n'est pas une probabilité [Edwards 92]. Si $P(R|hyp)$ représente la probabilité d'avoir le résultat R dans le cas d'une hypothèse fixe hyp, la vraisemblance est définie sur des données fixes R et des hypothèses variant. La vraisemblance n'obéit pas aux axiomes d'une probabilité et la fonction de vraisemblance ne génère pas de distribution de probabilité (en particulier la condition de normalisation n'est pas nécessaire d'être respectée).

Dans la théorie bayésienne, la fonction de vraisemblance est un opérateur de pondération de l'*a priori* pour déterminer l'*a posteriori*. Si la fonction de vraisemblance est en concordance avec la distribution de probabilité *a priori*, la variance de la distribution de probabilité sera diminuée par rapport à une combinaison linaire entre la distribution a priori et la fonction de vraisemblance. Dans le cas quand la fonction de vraisemblance n'est pas en concordance avec la distribution de probabilité *a priori*, le résultat de la distribution de probabilité *a posteriori* sera indésirable parce qu'il ne sera pas soutenu ni par distribution de probabilité *a priori*, ni par la fonction de vraisemblance.

Dans [Singpurwalla 04] et [Singpurwalla 02] il a été montré que les méthodes bayésiennes peuvent être utilisées pour faire la liaison entre la théorie de probabilités et la théorie des ensembles flous. Le point commun entre ces deux théories est l'utilisation d'une fonction d'appartenance (A.2). La fonction d'appartenance d'un ensemble flou peut être vue comme une fonction de vraisemblance pour un ensemble fixe S. Cette dernière affirmation est justifiée par le fait que le processus de détermination des fonctions d'appartenance est par sa nature subjectif, réfléchissant l'opinion d'un sujet sur le degré d'appartenance dans l'ensemble S. Comme résultat de cette hypothèse, les distributions de probabilité *a priori* peuvent directement être combinées avec les fonctions d'appartenance, vue comme des fonctions de vraisemblance, pour déterminer la distribution de probabilité *a posteriori*.

A.2 LA THÉORIE DES POSSIBILITÉS

L'hypothèse de départ de cette théorie est d'avoir un ensemble d'alternatives mutuellement exclusives Ω, parmi lesquels une seule est la vraie. À cause du caractère incomplet de l'information disponible on n'est pas certain quelle alternative est la vraie, donc on va avoir une incertitude de classification. Si on suppose que l'information qu'on dispose nous permet d'identifier un ensemble $E \subseteq \Omega, \quad E \neq \oslash$ dans lequel on est sûr que la vraie alternative se trouve, on peut définir une mesure de possibilité :

$$Pos_E(\omega) = \begin{cases} 1 & \text{si} \quad \omega \in E \\ 0 & \text{si} \quad \omega \in E^c \end{cases} \qquad (\forall)\omega \in \Omega \tag{A.11}$$

En prenant maintenant l'ensemble de tous les sous-ensembles de Ω, noté par la suite avec 2^{Ω}, la mesure de possibilité peut s'appliquer pour un ensemble $A \in 2^{\Omega}$:

$$Pos_E(A) = \sup_{\omega \in A} Pos_E(\omega) \tag{A.12}$$

Une mesure de possibilité peut être également définie, de la même manière, pour une α-coupe (voir équation A.4) :

$${}^{\alpha}Pos_E(\omega) = \begin{cases} 1 & \text{si} \quad \omega \in {}^{\alpha}E \\ 0 & \text{si} \quad \omega \in {}^{\alpha}E^c \end{cases} \qquad (\forall)\omega \in \Omega \tag{A.13}$$

Et si nous sommes intéressés par la mesure de possibilité d'un ensemble $A \in 2^{\Omega}$:

$${}^{\alpha}Pos_E(A) = \sup_{\omega \in A} {}^{\alpha}Pos_E(\omega) \tag{A.14}$$

En suivant le même raisonnement que celui de l'équation A.6, il est montré dans la littérature que :

$$Pos_E(\omega) = \sup_{\alpha \in [0,1]} \min[\alpha,\ {}^{\alpha}Pos_E(\omega)] \tag{A.15}$$

Maintenant si on considère un ensemble flou $\tilde{A}$ et en utilisant la notation de A.6, on a la mesure de possibilité donnée par :

$$Pos_E(\tilde{A}) = \sup_{\omega \in \Omega} \min[\tilde{A}(\omega),\ Pos_E(\omega)] \tag{A.16}$$

Avec la mesure de possibilité définie, la mesure de nécessité se calcule directement utilisant la relation :

$$Nec_E(A) = 1 - Pos_E(A^c) \tag{A.17}$$

Ainsi la théorie des possibilités utilise deux mesures monotones et semi-continues, la mesure de possibilité étant continue par valeurs supérieures et la mesure de nécessité étant continue par valeurs inférieures. En prenant la mesure de possibilité, $Pos \ : \ 2^{\Omega} \rightarrow [0,1]$, elle a les propriétés suivantes [Klir 01] :

- $Pos(\oslash) = 0$;
- $Pos(\Omega) = 1$;
- pour toute famille $\{A_i \mid A_i \in 2^{\Omega},\ i \in I\}$ avec I un ensemble d'index arbitraire on a :

$$Pos\left(\bigcup_{i \in I} A_i\right) = \sup_{i \in I} Pos(A_i) \tag{A.18}$$

A.3 La théorie de Dempster-Shafer

Shafer, décrivant dans sa monographie [Shafer 76] les bases de la théorie mathématique de l'évidence, voit cette théorie comme une généralisation de la théorie des probabilités subjectives, cette dernière étant introduite par De Finetti [Finetti 74]. Mais, en même temps, il précise que cette théorie ne doit pas être considérée comme supérieure, mais comme une alternative. Ainsi, chacune de ces deux théories va trouver des situations pour lesquelles leur emploi est adapté. Toujours selon son point de vue, la théorie de probabilité est applicable lorsqu'il est possible de connaître les probabilités des réponses possibles à une question, tandis que la théorie de l'évidence est applicable lorsqu'il est possible de connaître les probabilités des réponses possibles à une question relative.

Dans la théorie de l'évidence, la combinaison de plusieurs sources d'évidence se fait en supposant leur indépendance. Cette indépendance initiale pourrait être enlevée (par exemple dans un cas conflictuel quand seule une des sources peut être confiante) et d'arriver à une situation de conditionnement probabiliste.

En plus de l'indépendance des sources, Shafer dans sa définition de la théorie de l'évidence, fait les hypothèses d'exhaustivité et d'exclusion mutuelles des réponses, avec une seule réponse possible, c'est-à-dire une autre hypothèse d'unicité.

La théorie de l'évidence utilise deux mesures duales floues, appelées *croyance* Cr et *plausibilité* Pl. Ces deux mesures, par rapport à la mesure de probabilité qui est additive, sont supra-additives et respectivement sous-additives :

$$Cr(A \cup B) \geq Cr(A) + Cr(B) - Cr(A \cap B) \tag{A.19}$$

$$Pl(A \cap B) \leq Pl(A) + Pl(B) - Pl(A \cup B) \tag{A.20}$$

Ces deux mesures permettent la modélisation de l'incertitude, de l'imprécision et de l'ignorance et en plus, elles peuvent prendre en compte les éventuelles situations de conflits et d'ambiguïtés dans le cas d'un système multi-sources.

Les fonctions de croyance sont définies sur des sous-ensembles, une différence importante par rapport à la définition des probabilités, définies sur des singletons. En général, ces fonctions s'expriment en utilisant une autre fonction m, qui s'appelle fonction de masse. Cette fonction de masse exprime le degré de croyance dans la proposition faite par une source et strictement dans cette proposition.

Ainsi si nous considérons l'ensemble de tous les événements possibles, Ω et l'ensemble de tous les parties A de Ω, on a :

$$m : 2^{\Omega} \longrightarrow [0,1] \quad \sum_{A \subset \Omega} m(A) = 1 \tag{A.21}$$

$$Cr(A) = \sum_{B \subseteq A} m(B) \tag{A.22}$$

$$Pl(A) = \sum_{B \cap A \neq \oslash} m(B) = 1 - Cr(\overline{A}) \tag{A.23}$$

Selon [Smets 94a], chaque modèle de croyance a deux composantes : la première est une composante statique décrivant l'état actuel de la croyance et la deuxième est une composante dynamique exprimant la façon dont la croyance va se modifier avec l'intégration de nouvelles informations.

Ainsi, dans [Smets 94a], un modèle fondé sur deux niveaux est proposé :

- *le niveau crédal* : niveau où les croyances sont calculées, combinées et mises à jour ;
- *le niveau pignistique* : niveau où les croyances sont utilisées dans le processus de prise de décisions.

Toutes les difficultés de mise en pratique de cette théorie sont liées à la détermination des fonctions de masse et implicitement aux fonctions de croyances et de plausibilités.

A.4 La théorie de l'information généralisée

Quand on peut avoir des connaissances *a priori*, on peut utiliser la théorie des probabilités pour prendre en compte les incertitudes possibles sous la forme de probabilités *a priori*. En plus si des opinions des experts sont aussi disponibles, elles sont dans la plupart des cas exprimées sous des formes linguistiques et donc l'utilisation des fonctions d'appartenance floues sont plus adaptées ([Singpurwalla 04] montré que ces fonctions peuvent également être vues comme des fonctions de

vraisemblance et donc en appliquant le théorème de Bayes on peut combiner ces deux incertitudes pour construire la distribution *a posteriori*).

Comme on a pu voir dans les trois paragraphes précédents, dans chaque théorie, l'incertitude est exprimée à l'aide d'une fonction, qu'on peut l'appeler *fonction d'incertitude*, qui attribue à chaque alternative un nombre dans l'intervalle $[0, 1]$ traduisant le degré d'évidence (la vraisemblance, la probabilité, la croyance, la possibilité, etc.). En fonction de la théorie de l'incertitude utilisée, la fonction d'incertitude peut être représentée par la mesure de probabilité ou par la mesure de possibilité ou par la mesure de croyance.

Ces différents types de fonctions d'incertitude ont des caractéristiques spécifiques qui les différencient entre elles. Ainsi, par exemple, il y a des différences entre les opérations utilisées. Dans la théorie des probabilités, les opérations utilisées sont l'addition et la multiplication suivant leur définition de l'algèbre linéaire. Dans la théorie de l'information généralisée, l'opérateur d'addition sera vu comme une disjonction généralisée et l'opérateur de multiplication comme une conjonction généralisée. En plus de ces deux opérateurs, d'autres opérateurs peuvent être également utilisés.

Dans chaque théorie de l'incertitude et pour chaque type d'incertitude, il peut se définir **une mesure d'incertitude**. Cette mesure d'incertitude est une fonctionnelle qui attribue pour chaque fonction d'incertitude une valeur numérique non-négative [Klir 01] :

$$u : U(\mu) \rightarrow \mathbb{R}^{+} \tag{A.24}$$

Concernant les mesures d'incertitudes (théorie des probabilités, théorie des possibilités, théorie de l'évidence, théorie des probabilités imprécises), elles ont été définies et construites pour des **variables disjointes**. Une variable disjointe a la propriété d'avoir à chaque instant **une seule valeur** et l'incertitude est exprimée vis-à-vis de cette valeur. **Les variables conjonctives** ont la propriété d'être caractérisées par des valeurs multiples. Une théorie rigoureuse dans ce cas n'a pas été développées jusqu'à présent. Un premier effort dans cette direction a été fait dans l'étude [Yager 87a].

Pour être acceptée comme une mesure de la quantité de l'incertitude d'un certain type dans une certaine théorie, une mesure de l'incertitude doit avoir quelques propriétés axiomatiques [Klir 01] :

1. **Sous-additivité** : pour une évidence conjointe la quantité d'incertitude ne doit pas être plus grande que la somme des quantités d'incertitude des évidences marginales ;
2. **Additivité** : la quantité d'incertitude d'une évidence conjointe est égale à la somme des quantités d'incertitude des évidences marginales seulement dans le cas où les représentations marginales de l'évidence sont non-interactives ;
3. **L'intervalle** : de l'incertitudes est entre $[0, M]$. La valeur de 0 correspond au cas de la certitude totale et la valeur de M dépend du cardinal de l'ensemble universel et de la métrique choisie ;
4. **Continuité** : la mesure d'incertitude doit être continue ;
5. **Expansibilité** : l'ajout d'alternatives caractérisées par une évidence nulle, ne doit pas influencer la quantité de l'incertitudes ;
6. **Consistance** : l'utilisation de différents techniques (de bon sens) de calcul de l'évidence doit aboutir aux mêmes résultats.

En plus de ces six demandes axiomatiques d'une mesure d'incertitude, il peut y avoir des demandes particulières qui doivent s'appliquer pour certaines théories d'incertitudes seulement. Un exemple est dans le cas de la théorie des possibilités, dans laquelle l'évidence peut être ordonner :

– **Monotonicité** : la mesure d'incertitude doit préserver l'ordre.

Comme dans une théorie de l'incertitude plusieurs types d'incertitudes peuvent coexister, il doit y avoir une mesure d'incertitude globale qui peut couvrir la combinaison individuelle de ces types d'incertitude.

A.5 Information basée sur l'incertitude

Si on regarde la définition de la notion d'incertitude dans le contexte d'un système de traitement de l'information, on peut dire que l'incertitude est due à un manque d'information. Donc les notions d'incertitude et d'information sont liées, une baisse du niveau d'incertitude valant une augmentation du niveau de l'information. Par conséquent, le gain en information peut être mesuré comme la différence entre l'incertitude *a priori* et l'incertitude *a posteriori*. Mais cette mesure indirecte de la quantité d'information est dépendante de la capacité de caractériser les incertitudes et plus précisément de caractériser les différents types de l'incertitude dans des cadres mathématiques différents. Cette information déterminée suite à la réduction de l'incertitude a été appelée par [Klir 88] *information fondée sur l'incertitude*[2].

Cette définition de l'information est très attractive pour la caractérisation d'un système. Ainsi si on considère un système S et on veut mesurer la quantité d'information d'une réponse à une question Q (demande de prédiction, diagnostique, etc.) obtenue dans le cadre expérimental du système CE_S, la formule de calcule est immédiate :

$$\text{Information}(R_S \mid S, Q) = \text{Incertitude}(R_{CE_S} \mid CE_S, Q) - \text{Incertitude}(R_S \mid S, Q) \tag{A.25}$$

où R_S désigne la réponse à la question Q obtenue par le système S et où R_{CE_S} désigne la réponse à la question Q obtenue seulement dans le cadre expérimental du système CE_S.

Comme observation, on peut dire que cette mesure d'information fondée sur l'incertitude ne peut pas englober des aspects sémantiques ou pragmatiques de l'information et donc elle n'est pas adaptée pour la quantification de l'information issue des communications humaines ou encore des informations cognitives.

Dans leur étude, Klir et Folger [Klir 88] ont séparé deux termes pour la caractérisation de l'incertitude : *le flou* et *l'ambiguïté*. En général, le flou est présente quand il est impossible de faire des distinctions nettes et précises entre les alternatives. L'ambiguïté est présente quand le choix parmi deux ou plusieurs alternatives n'est pas spécifiée. À partir de ces définitions, on peut observer que le concept d'ensemble flou donne le cadre mathématique pour traiter le flou, le vague. En même temps, la mesure floue nous permet d'avoir le cadre mathématique pour traiter l'ambiguïté.

Trois types d'ambiguïtés peuvent être identifiés [Klir 88] :

1. **La non-spécificité :** caractérise la taille des sous-ensembles pour lesquels une mesure floue les donne comme des locations potentiellement vraies. Plus la taille est grande, plus la non-spécificité est importante ;
2. **La dissonance** ou encore **le conflit :** typique pour les situations où une mesure floue donne des sous-ensembles disjoints comme des locations potentiellement vraies ;
3. **La confusion :** est associée avec le nombre de sous-ensembles contenus dans Ω et qui sont soutenus comme des locations potentiellement vraies.

A.6 La mesure du flou

Le cardinal d'un ensemble flou est donné par :

$$|\tilde{A}| = \sum_{\omega \in \Omega} \mu_{\tilde{A}}(\omega) \tag{A.26}$$

où avec $\tilde{A}$ a été représenté un ensemble flou.

Une mesure du flou[3] est une fonction $f : \tilde{P}(\Omega) \to \mathbb{R}$, où avec $\tilde{P}(\Omega)$ a été noté l'ensemble de tous les sous-ensembles de Ω. Dans [Klir 88] trois propriétés axiomatiques ont été présentées, liées à la notion de degré de flou, qu'une mesure du flou doit respecter :

2. De l'anglais : "Uncertainty-based information"
3. en anglais "measure of fuzziness"

1. $f(\tilde{A} = 0 \Leftrightarrow A$ est un ensemble net ;
2. S'il y a la relation : $\tilde{A}$ *est moins flou (ou plus net) que* $\tilde{B}$:

$$\tilde{A} \prec \tilde{B} \Rightarrow f(\tilde{A}) \leq f(\tilde{B})$$

3. Le maximum de degré de flou est atteint seulement pour les ensembles flous catalogués comme étant "les plus flous".

Dans la littérature a été proposée une classe des mesures du flou ayant la forme :

$$f(\tilde{A}) = h\left(\sum_{\omega\in\Omega} g_\omega(\mu_{\tilde{A}}(\omega))\right) \tag{A.27}$$

où g_ω est une fonction $g_\omega : [0,1] \to \mathcal{R}^+$ monotone croissante sur $[0, \frac{1}{2}]$ et monotone décroissante sur $[\frac{1}{2}, 1]$, avec un seul point de maximum en $\frac{1}{2}$ et de valeur nulle en 0 et 1. La fonction $h : \mathbb{R}^+ \to \mathbb{R}^+$ est une fonction monotone croissante.

Les choix les plus citées dans la littérature pour les fonction g_ω et h sont :

– En prenant pour comme fonction h la fonction identité et pour la fonction g_ω :

$$g_\omega(\mu_{\tilde{A}}(\omega)) = -\mu_{\tilde{A}}(\omega)\log_2\mu_{\tilde{A}}(\omega) - [1-\mu_{\tilde{A}}(\omega)]\log_2[1-\mu_{\tilde{A}}(\omega)]\ ;\ (\forall)\omega\in\Omega$$

En remplaçant dans la formule A.27 on obtient pour la mesure du flou la forme :

$$f(\tilde{A}) = -\sum_{\omega\in\Omega}\{-\mu_{\tilde{A}}(\omega)\log_2\mu_{\tilde{A}}(\omega) + [1-\mu_{\tilde{A}}(\omega)]\log_2[1-\mu_{\tilde{A}}(\omega)]\} \tag{A.28}$$

– En prenant comme fonction $h(x) = x^{1/n}$ pour $n \in [1,\infty]$ et avec :

$$g_\omega(\mu_{\tilde{A}}(\omega)) = \begin{cases} \mu^n_{\tilde{A}}(\omega) & \text{si} \quad \mu_{\tilde{A}}(\omega) \in [0, \frac{1}{2}] \\ 1-\mu^n_{\tilde{A}}(\omega) & \text{si} \quad \mu_{\tilde{A}}(\omega) \in [\frac{1}{2}, 1] \end{cases} \qquad (\forall)\omega\in\Omega$$

En remplaçant dans la formule A.27 on obtient pour la mesure du flou la forme :

$$f(\tilde{A}) = \left(\sum_{\omega\in\Omega}|\mu_{\tilde{A}}(\omega) - \mu_C(\omega)|^n\right)^{\frac{1}{n}} \tag{A.29}$$

Dans cette dernière équation la mesure du flou se calcule en utilisant une distance, faisant partie de la classe des distances de type Minkowski, entre l'ensemble flou $\tilde{A}$ et un ensemble net C caractérisé par :

$$\mu_C(\omega) = \begin{cases} 0 & \text{si} \quad \mu_{\tilde{A}}(\omega) \leq \frac{1}{2} \\ 1 & \text{si} \quad \mu_{\tilde{A}}(\omega) > \frac{1}{2} \end{cases} \qquad (\forall)\omega\in\Omega$$

Une autre façon naturelle de quantifier le flou d'un ensemble est d'utiliser le manque de distinction entre l'ensemble flou et son complément.

Maintenant si on s'intéresse à la différence d'un ensemble flou et d'un ensemble classique, on peut définir une mesure de distance pour la quantifier :

$$Z(\tilde{A}) = \sum_{\omega\in\Omega}|\mu_{\tilde{A}}(\omega) - c(\mu_{\tilde{A}}(\omega))| \tag{A.30}$$

Dans la formule précédente, l'opérateur c désigne le complément flou. Si pour le complément flou, on prend comme formule de calcul $c(\tilde{A}(\omega)) = 1 - \tilde{A}(\omega)$ l'équation A.30 peut se réécrire :

$$Z(\tilde{A}) = \sum_{\omega\in\Omega}(1 - |2\mu_{\tilde{A}}(\omega) - 1|) \tag{A.31}$$

Ces deux équations quantifient l'éloignement d'un ensemble flou vis-à-vis d'un ensemble net, plus la distance Z est grande plus les différences, en terme de flou, entre les deux ensemble sont importantes.

Dans [Klir 88], il est montré que si on veut avoir comme unité de mesure pour le flou d'un ensemble « le bit », on peut utiliser la formule suivante :

$$f_c(\tilde{A}) = |\tilde{A}| - \sum_{\omega \in \Omega} |\mu_{\tilde{A}}(\omega) - c(\mu_{\tilde{A}}(\omega))| \quad \text{[en bits]} \tag{A.32}$$

A.7 Les mesures de la non-spécificité

Comme présenté précédemment, la non-spécificité est caractérisée pour les situations quand plusieurs alternatives sont possibles, donc pour les ensembles d'alternatives avec un cardinal plus grand que 1.

Le rôle de l'utilisation des sources multiples d'information est d'augmenter le niveau d'information globale et de réduire la non-spécificité, ou de manière équivalente d'augmenter la spécificité. Ainsi, dans le cas où les informations fournies par les différentes sources ne sont pas conflictuelle, la fusion de ces information va rendre un niveau d'information plus grand que le niveau d'informations individuelles des sources. Dans le cas où le conflit est présent, l'opération de fusion d'information peut entrainer une diminution de la quantité de l'information par rapport à des informations individuelles. Dans ce type de situations, une solution est de ne pas prendre en compte toutes les informations, donc de faire un filtrage avant l'opérateur de fusion.

A.7.1 La mesure de spécificité

La notion de spécificité est inversement reliée à la notion d'entropie et elle est employée pour mesurer la qunatité d'information contenue dans une affirmation. Dans [Yager 05] une mesure de la spécificité à été proposée : Étant donné A un ensemble flou du domaine X et x^* tel que $A(x^*) = max(A(x))$, la spécificité du A est donnée par

$$Sp(A) = A(x^*) - \hat{A} \tag{A.33}$$

où avec $\hat{A}$ a été notée la valeur moyenne de la fonction d'appartenance du A sur le domaine $X - \{x^*\}$.

A.7.2 L'information Hartley

Avant que la théorie des ensembles flous soit développée, Hartley en 1928 et Shannon en 1948 ont proposé deux théories pour la mesure de l'information et donc implicitement de l'incertitude.

Pour définir une mesure de l'information, Hartley a considéré **un ensemble fini** Ω constitué de n éléments. Quand on veut construire une séquence d'éléments $\omega_i \in \Omega$ par des sélections successives, a priori il va y avoir une ambiguïté proportionnelle avec le nombre d'éléments $|\Omega| = n$. Le nombre total des séquences possibles obtenues par s sélections est n^s et la quantité d'information associée avec ces sélections est égale à :

$$I(n^s) = s \log_2 n = \log n^s \quad \text{[en bits]} \tag{A.34}$$

Et si on note avec N le nombre total d'alternatives possibles, le moyen de sélection étant indifférent, l'information de Hartley peut encore s'exprimer sous la forme :

$$I(N) = \log_2 N \quad \text{[en bits]} \tag{A.35}$$

Si on prend deux ensembles X et Y et on considère qu'on peut définir une relation $R \subset X \times Y$ qui exprime la corrélation entre la sélection dans un ensemble et les sélections faites dans l'autre ensemble, trois catégories d'informations de type Hartley peuvent être définies :

1. *L'information marginale :*

$$I(X) = \log_2 |X|$$
$$I(Y) = \log_2 |Y|$$

2. *L'information conjointe :*

$$I(X,Y) = \log_2 |R|$$

3. *L'information conditionnelle :*

$$I(X|Y) = \log_2 \frac{|R|}{|Y|} = \log_2 |R| - \log_2 |Y|$$
$$I(Y|X) = \log_2 \frac{|R|}{|X|} = \log_2 |R| - \log_2 |X|$$

Une autre fonction, appelée fonction de transmission de l'information, peut être définie pour servir comme indicateur de la corrélation entre les deux ensembles :

$$T(X,Y) = I(X) + I(Y) - I(X,Y) \tag{A.36}$$

Cette fonction peut être généralisée pour le cas de n ensembles :

$$T(X_1, X_2, \ldots, X_n) = \sum_{i=1}^{n} I(X_i) - I(X_1, X_2, \ldots, X_n)$$

A.7.3 L'information de Hartley

L'information de Hartley a été définie seulement pour le cas des ensembles finis, mais une extension pour le cas des espaces euclidiens de dimension n existe aussi [Klir 06b]. En considérant que les alternatives sont des points dans un espace $\mathbb{R}^n$ et que l'évidence qu'on dispose nous amène à un sous-ensemble $E \subset \mathbb{R}^n$ borné et convexe, l'information HL a la forme :

$$I_{HL}(E) = \min_{t \in T} \log_2 \left[\prod_{i=1}^{n} (1 + \mu(E_{i_t})) + \mu(E) - \prod_{i=1}^{n} \mu(E_{i_t}) \right] \tag{A.37}$$

Dans cette dernière équation, avec μ a été notée la mesure de Lebesgue, avec T l'ensemble de toutes les transformations d'un système de coordonnés à un autre et avec E_{i_t} la $i^{\text{ème}}$ projection unidimensionnelle du E dans le système de coordonnés.

A.7.4 La non-spécificité dans la théorie de l'évidence

La mesure d'information de Hartley permet de quantifier la non-spécificité dans le cas de la théorie des possibilités. Dans ce paragraphe, les mesures permettant de quantifier la non-spécificité seront présentées dans le cadre de la théorie de l'évidence (de Dempster-Shafer). Dans cette théorie la non-spécificité est présente quand la masse d'évidence est distribuée pour des sous-ensembles A_i

qui contiennent plus d'un seul élément. Une méthode pour quantifier la non-spécificité est d'adapter la mesure de Hartley pour ce cas :

$$H_e(m) = \sum_{i=1}^{n} m(A_i) \log_2(|A_i|) \tag{A.38}$$

où $m = \{m(A_1), m(A_2), \cdots, m(A_n)\}$ est la masse d'évidence.

A.7.5 La non-spécificité pour les possibilités graduelles

Quand tous les éléments focaux sont emboîtés : $\Omega = \{\omega_1, \omega_2, \cdots, \omega_n\}$ avec une relation d'ordre pour les valeurs des possibilités : $\pi(\omega_i) \geq \pi(\omega_{i+1})$, la mesure possibiliste de la non-spécificité dévient :

$$H_p(\pi) = \sum_{i=2}^{n} [(\pi(\omega_i) - \pi(\omega_{i+1})) \log_2 i] \tag{A.39}$$

Si la théorie des possibilités est regardée du point de vue des ensembles flous et si on utilise les α-coupes, la mesure de la non-spécificité peut s'écrire sous la forme [Higashi 83] :

$$U(A) = \int_0^1 \log_2(|^{\alpha}A|) d\alpha \tag{A.40}$$

Toujours dans [Higashi 83] il a été proposée une distance pour mesurer le rapprochement en niveau d'information dans le cas des distributions de possibilité. Cette distance a été définie en partant des distributions de possibilité partiellement ordonnées. Deux distributions de possibilité π_1 et π_2 sont partiellement ordonnées, relation noté par la suite par $\pi_1 \preceq \pi_2$ si et seulement si $\pi_1(\omega) \leq \pi_1(\omega)\ \forall \omega \in \Omega$. Pour toute paire π_1, π_2 le gain d'information obtenu en remplaçant π_2 par π_1 est quantifié par :

$$g(\pi_1, \pi_2) = U(\pi_2) - U(\pi_1) = \int_0^1 \log_2 \frac{|^{\alpha}\pi_2|}{|^{\alpha}\pi_1|} d\alpha \tag{A.41}$$

Comme $\pi_1 \preceq \pi_2 \Rightarrow^{\alpha} \pi_1 \subseteq^{\alpha} \pi_2\ \forall \alpha \in [0,1]$, le gain d'information est une quantité positive $g(\pi_1, \pi_2) \geq 0$. Une autre observation est que dans le cas où $\pi_1 \preceq \pi_2 \preceq \pi_3$ le gain d'information est additif, c'est-à-dire $g(\pi_1, \pi_3) = g(\pi_1, \pi_2) + g(\pi_2, \pi_3)$. De ces deux observations, on a comme conséquence que $g(\pi_1, \pi_3) \geq g(\pi_1, \pi_2)$. A l'aide de cette mesure de gain d'information, Higashi et Klir [Higashi 83] ont proposé une distance métrique basée sur le rapprochement de l'information, dans le cadre de la théorie de possibilité :

$$G(\pi_1, \pi_2) = g(\pi_1, \max\{\pi_1, \pi_2\}) + g(\pi_2, \max\{\pi_1, \pi_2\}) \tag{A.42}$$

La construction d'une distance d'information ayant des propriétés d'une métrique dans le cadre des distributions de probabilité de la même façon que dans le cadre de la théorie de possibilités n'a pas été possible, car pour les distribution de probabilité on ne peut pas avoir une ordonnance partielle entre deux distributions de probabilité.

Pour les nombres flous et pour les intervalles flous, la mesure de non-spécificité s'écrit :

$$HL(A) = \int_0^1 \log_2(1 + \mu(^{\alpha}A)) d\alpha \tag{A.43}$$

A.7.6 La fusion des informations en intégrant la notion de spécificité

Cette partie est faite d'après [Yager 05]. En considérant qu'on a deux informations sur une variable V de la forme $V\ is\ A$ et $V\ is\ B$, le résultat de la fusion (dans la théorie des possibilités) sera donné par $V\ is\ D$ avec $D = A \cap B$ et $D(x) = Min\{A(x), B(x)\}$. Maintenant si on considère qu'on a n informations de la forme $V\ is\ A_i$ avec $i = \overline{1,n}$, le résultat de la fusion sera donné par $V\ is\ D$ avec $D = \bigcap_i^n A_i$ et $D(x) = Min\{A_i(x)\}$. Comme conclusion on peut voir que plus d'informations on a en entrée, plus spécifique sera le résultat de la fusion.

Observations :

- Si les informations d'entrée (qui peuvent être vues comme des représentations floues normalisées : $((\exists)x \quad \text{pour lequel} \quad A_i(x) = 1; \quad \forall i)$) ne sont pas conflictuelles, la fusion va rendre une information d'une spécificité plus grande :

$$Sp(D) \geq Sp(A_i) \quad \forall i \quad \text{et où} \quad D = \bigcap_i^n A_i \tag{A.44}$$

- Dans le cas des informations conflictuelles, on peut être dans la situation où le résultat de la fusion, l'ensemble flou D, n'est plus normalisé et donc l'information de sortie peut être plus confuse que les informations en entrée. Dans ce cas, des mesures de précaution doivent être prises : ne pas prendre pour l'opération de fusion toutes les informations disponibles, mais seulement un sous-ensemble (pour limiter le degré de conflit) suffisamment grand pour que la crédibilité du résultat de fusion soit acceptable.

Par la suite sera notée avec P_i la relation $V\ is\ A_i$ et avec $P = \{P_1, P_2, \cdots, P_n\}$. Aussi on définit une mesure $\mu : 2^{|P|} \to [0,1]$ qui pour chaque sous-ensemble $B \in P$, $\mu(B)$ désigne la crédibilité qu'on a dans le résultat de la fusion des éléments de B.

Si on considère $B \in P$, on va noter avec $D_B = \bigcap_i^n A_i$ la fusion des connaissances contenues par B, donc on va avoir l'affirmation $V\ is\ D_B$ (avec une crédibilité $\mu(B)$). Pour évaluer la qualité de l'affirmation $V\ is\ D_B$ deux mesures de qualité peuvent être utilisées :

1. **L'informativité** : directement liée à la notion de spécificité

$$\text{Inf}(B) = Sp(B) = D_B(x^*) - \hat{D_B} \tag{A.45}$$

2. **La crédibilité** : mesurée par $\text{Cred}(B) = \mu(B)$

Comme les deux mesures de qualité ont les valeurs exprimées dans l'intervalle unitaire, la qualité globale peut s'obtenir en utilisant différents opérateurs. Dans [Yager 05] a été choisi d'exprimer la qualité globale sous la forme du produit des deux qualités :

$$\text{Qual(B)} = \text{Inf(B)} \cdot \text{Cred(B)} \tag{A.46}$$

A.7.7 Sur la confiance

La construction des mesures de confiance n'est pas évidente et dans beaucoup de cas, elle se fait en fonction du besoin de l'application courante. D'après [Yager 05] il peut y avoir :

- **Mesures de confiance basées sur la notion de cardinal** : Dans ce cas, il n'y a pas de différences de crédibilité entre les différentes pièces d'informations (alternatives) proposées. La seule chose qui compte c'est le nombre. Pour donner une qualité globale des propositions, dans ce cas elle peut se faire en utilisant des termes linguistiques.
- **Mesures de confiance additives** : Chaque pièce d'information P_i a une crédibilité propre α_i et en plus la somme des crédibilités individuelles est normée : $\sum_i \alpha_i = 1$.

– **Mesures de confiance basées sur une collection de sous-ensembles de P** : En notant avec $G_k,\ k = \overline{1,K}$ une collection de sous-ensembles de P on peut construire une mesure de confiance de B sur le principe : $\mu(B) = 1$ ssi B contient des pièces d'informations de chaque G_k et $\mu(B) = 0$ au cas contraire. Sur cette idée, on peut construire une mesure de confiance de la forme :

$$\mu(B) = \sum_{k=1}^{K} g_k \frac{|B \cap G_k|}{|B|} \quad \text{avec} \quad g_k \in [0,1] \tag{A.47}$$

Observation : Cette idée pourrait être utilisée pour l'évaluation d'un système d'aide à la décision quand on ne connait pas le nombre *a priori* d'alternatives à proposer. Ainsi, on peut construire le G_k avec l'aide de l'utilisateur et après évaluer notre système en fonction des alternatives proposées (le B).

– **Mesures de confiance contenant des pièces d'informations nécessaire** : Utile dans les cas où une pièce d'information P_i est impérative pour l'utilisateur : $\mu(B) = 0$ si $P_j \notin B$.

A

A.8 LES MESURES BASÉES SUR L'ENTROPIE

Dans le cas de la théorie des probabilités deux sources d'incertitudes peuvent être identifiées : la non-spécificité et le conflit. Le premier type d'incertitude, la non-spécificité, peut être quantifié en utilisant la mesure de Hartley. Pour quantifier le conflit, on peut utiliser l'entropie de Shannon. Cette entropie a été également utilisée pour mesurer l'incertitude dans la théorie de l'évidence.

A.8.1 L'entropie de Shannon pour les distribution des probabilités

L'entropie de Shannon est une mesure de l'incertitude d'une variable aléatoire. Pour une distribution de probabilité d'une variable aléatoire discrète X prenant les valeurs dans l'alphabet χ, l'entropie de Shannon s'écrit :

$$\begin{aligned} S(X) &= -\sum_{x\in\chi} p(x)\log_2(p(x)) \\ &= -\sum_{x\in\chi} p(x)\log_2\left(1 - \sum_{y\neq x}(p(y))\right) \end{aligned} \tag{A.48}$$

Dans la deuxième forme de l'équation A.48 il peut s'observer que l'entropie de Shannon exprime le conflit entre $p(x)$ et $\sum_{y\neq x} p(y)$. L'entropie de Shannon quantifie la valeur moyenne de conflit parmi les évidences exprimées utilisant des distributions des probabilités pour un ensemble fini d'alternatives mutuellement exclusives. Ainsi par rapport à la mesure de Hartley, qui mesure la non-spécificité, l'entropie de Shannon mesure le conflit.

Pour le cas quand $\sum_{x\in\chi} p(x) < 1$ on parle d'un système incomplète et la mesure d'entropie de Shannon s'écrit :

$$S(X) = -\frac{\sum_{x\in\chi} p(x)\log_2(p(x))}{\sum_{x\in\chi} p(x)} \tag{A.49}$$

Dans ce dernier cas, il est possible de calculer l'entropie d'un seul événement.

A.8.2 L'entropie de Rényi pour les distribution des probabilités

L'entropie de Shannon peut être vue comme une moyenne arithmétique pondérée, avec les probabilités des événements, des quantités $-\log p(x)$. Comme il a été remarqué dans [Aczel 06], la moyenne arithmétique n'est pas la seule moyenne intéressante. L'entropie de Rényi d'ordre α est définie par :

$$ {}^{\alpha}H_n(p_1, p_2, \ldots, p_n) = \frac{1}{1-\alpha} \log_2 \sum_{k=1}^{n} p_k^{\alpha} \tag{A.50} $$

Quand $\alpha \to 1$ l'entropie de Rényi est équivalente avec l'entropie de Shannon. Pour encore généraliser l'entropie de Shannon, on peut généraliser la moyenne arithmétique pondérée par [Aczel 06] :

$$ f^{-1}\{H_n(p_1, p_2, \ldots, p_n)\} = f^{-1}\left\{\sum_{k=1}^{n} p_k f(-\log_2 p_k)\right\} \tag{A.51} $$

où la fonction f est continue et strictement monotone ($f(x) = x$, $f(x) = \log(x)$, $f(x) = e^x$, $f(x) = x^c$, etc.).

A.8.3 La mesure de divergence

Lorsqu'un expert fait une estimation $P(X)$ d'une fonction distribution de probabilité inconnue $S(X)$, on peut utiliser une mesure de divergence pour quantifier la différence entre la vraie distribution et celle estimée [Klir 06b] :

$$ S_D(S, P) = -\sum_{x \in \chi} s(x) \log_2 \left(\frac{s(x)}{p(x)}\right) \tag{A.52} $$

Cette mesure peut être utile pour apprécier les opinions obtenues d'un ensemble d'experts qui sont mis dans les mêmes conditions. De plus, cette mesure peut exprimer la surprise d'un expert (Cooke1991).

A.8.4 Les mesures basées sur l'entropie pour la théorie de l'évidence

1. **La mesure de dissonance :**

$$ D(m) = -\sum_{i=1}^{n} m(A_i) \log_2(Pl(A_i)) \tag{A.53} $$

   où $\{A_1, A_2, \cdots, A_n\}$ est un ensemble d'éléments focaux, $m(A_i)$ est la masse d'évidence pour l'élément A_i et $Pl(A_i)$ est la mesure de plausibilité.

2. **La mesure de confusion :**

$$ C(m) = -\sum_{i=1}^{n} m(A_i) \log_2(Cr(A_i)) \tag{A.54} $$

   où $Cr(A_i)$ est la mesure de croyance.

A.8.5 Agrégation des incertitudes dans le cadre de la théorie de l'évidence

Le but de cette section est de combiner les deux types d'incertitudes : la non-spécificité et le conflit dans le cadre de la théorie de l'évidence. L'agrégation de l'incertitude est une fonction définie sur l'ensemble de mesures de croyance dans le $\mathbb{R}^+$. Cette mesure d'agrégation est donnée par [Klir 06b] :

$$AU(Cr) = \max_{P_{Cr}} \left[-\sum_{x \in \chi} p(x) \log_2(p(x)) \right] \tag{A.55}$$

où P_{Cr} est l'ensemble de toutes les distributions de probabilité $p(x)$ qui satisfassent les conditions suivantes :

- $p(x) \in [0,1]$ $(\forall)$ $x \in \chi$ et $\sum_{x \in \chi} p(x) = 1$
- $Cr(A) \leq \sum_{x \in A} p(x) \leq 1 - Cr(A^c)$

Dans [Klir 06b] est décrit un algorithme itératif pour le calcul de l'équation A.55.

A.9 MÉTHODOLOGIES POUR LES TRAITEMENTS AVEC DES INCERTITUDES

Dans le paragraphe A.4, les éléments de base de la théorie de l'information généralisée (TIG) ont été présentés. Dans ce paragraphe, quelques méthodologies pour les traitements avec des incertitudes seront présentées selon [Klir 06a].

Si on regarde plus attentivement la théorie de l'information généralisée, on peut remarquer deux choses :

1. Dans la TIG, il existe plusieurs théories de l'incertain et donc on peut observer une *diversité* en ce sens. Cette diversité est très importante car les axiomes de départ de chaque théorie de l'incertain sont différents et donc, en fonction de l'application, il y a des théories qui se prêtent mieux que les autres. En plus à cette diversité, on peut ajouter la notion de dynamique car des nouvelles théories peuvent être construites ;
2. Toutes les théories de la TIG ont des propriétés communes et donc, on peut observer une *unicité* en ce sens (un exemple étant la généralisation des mesures de Hartley et de Shannon pour d'autres théories).

A.9.1 Le principe de minimum d'incertitude

Ce principe est aussi appelé le principe de perte minimale d'information. Comme le nom l'indique, ce principe facilite la sélection d'alternatives plausibles en minimisant la perte d'information pour résoudre un problème. Deux exemples de problèmes à résoudre sont le problème de simplification et le problème avec conflit.

A.9.1.1 Le problème de simplification

Toute simplification d'un système est d'habitude accompagnée par une perte d'information. Ainsi le but est de réduire cette perte d'information ou de limiter l'augmentation de l'incertitude. En général les simplifications d'un système sont faites pour réduire la complexité et pour cela trois catégories de simplification peuvent être identifiées [Klir 06a] :

- En éliminant des entités : variables, sous-systèmes, etc. ;
- En combinant des entités : variables, états, etc. ;

– Décomposition du système en plusieurs sous-systèmes.

Pour des exemples concernant le problème de simplification voir [Klir 06a].

A.9.1.2 Le problème avec du conflit

Le conflit peut apparaître sous la forme des inconsistances locales lors de l'intégration des sous-systèmes dans un seul système. Pour les variables appartenant à plusieurs sous-systèmes, l'inconsistance locale du système peut être caractérisée par des différences au niveau des projections des fonction d'incertitude en fonction de ces variables. Ainsi cette inconsistance doit être résolue par la modification des fonctions d'incertitudes des sous-systèmes en minimisant la perte d'information. La perte totale d'information peut être calculée comme étant la somme des pertes en information de chaque sous-système.

Si on note avec $I({}^s u, {}^s \hat{u})$ la perte d'information due à la modification de la fonction d'incertitude ${}^s u$ avec ${}^s \hat{u}$, l'élimination de l'inconsistance peut être formulée comme un problème d'optimisation :

– Étant donnée une famille de sous-systèmes $\{{}^s Z | s \in \mathbb{N}_n\}$avec les fonctions d'incertitudes ${}^s u$ formalisées dans une certaine théorie de l'incertain inconsistantes localement, trouvez les fonction d'incertitudes ${}^s \hat{u}$ pour lesquelles $\sum_{s=1}^{n} I({}^s u, {}^s \hat{u})$ est minimales sous les contraints des axiomes de la théorie de l'incertain utilisée et sous la condition d'avoir la consistance locale avec les fonction d'incertitude ${}^s \hat{u}$.

A.9.2 Le principe du maximum d'incertitude

Ce principe nous permet de développer des procédures de traitement pour un grand nombre de problèmes qui nécessitent un raisonnement ampliatif.

A.9.2.1 Le principe du maximum d'entropie

C'est le principe utilisé dans la théorie de l'information classique, donc dans le cadre de la théorie des probabilités. La formulation de ce principe peut être faite comme un problème d'optimisation :

– déterminer une distribution de probabilité $p(x)$ qui maximise l'entropie de Shannon sous les contraintes $c_i,\ i = \overline{1,n}$ qui expriment les axiomes de la théorie des probabilités, mais aussi les informations partielles sur la distribution inconnue (des moments d'une (ou des) variable aléatoire, distributions de probabilités marginales, etc.).

Pour des exemples d'application de ce principe voir [Klir 06a] - estimation de la distribution de probabilité sachant le moment d'ordre 1 (l'espérance mathématique).

Une situation très utile pour la pratique est quand on dispose d'une distribution de probabilité $p^{'}(x)$ et on veut déterminer la distribution de probabilité $p(x)$ suite à l'arriver de nouvelle évidence. En ce cas on peu utiliser **le principe de minimum d'entropie croisée**. L'entropie croisée est définie par :

$$\hat{S}(p(x), p^{'}(x) = \sum_{x \in \chi} p(x) \log_2 \left(\frac{p(x)}{p^{'}(x)} \right) \tag{A.56}$$

La nouvelle évidence qu'on dispose va bien sûr diminuer l'incertitude et donc l'incertitude exprimée par $p(x)$ est plus petite que l'incertitude exprimée par $p^{'}(x)$. Ainsi le principe de minimum d'entropie croisée nous permet de réduire l'incertitude de $p^{'}(x)$ avec la plus petite quantité nécessaire pour satisfaire les nouvelles contraintes données par la nouvelle évidence.

A.9.2.2 Le principe du maximum de non-spécificité

Dans les cas où le seul type d'incertitude est la non-spécificité (dans la théorie classique des possibilités quand la non-spécificité est quantifiée par la mesure de Hartley A.34 ou encore dans la théorie de l'évidence quand la mesure de la non-spécificité est donnée par A.38), on peut énoncer un principe de maximisation de la non-spécificité (dans une manière similaire que le principe du maximisation d'entropie et donc qui s'applique pour les situations de raisonnement ampliatif). Ainsi, ce principe peut être vu comme un problème d'optimisation, nécessitant la maximisation de la fonctionnelle fondée sur la mesure de Hartley sous les contraintes des axiomes de la théorie des possibilités classiques et de la nouvelle évidence qui est disponible.

Ce principe, par rapport au principe de maximum d'entropie qui est applicable seulement dans le cadre de la théorie des probabilités, est applicable pour toutes les autres théories et donc ces deux principes sont complémentaires. Pour plus d'informations et exemples voir [Klir 06a] et [Klir 88].

A.9.2.3 Le principe du maximum d'incertitude dans la TIG

Dans la plupart des situations, sauf dans le cas des théorie classiques, plusieurs types d'incertitude coexistent dans la cadre de la même théorie. Ainsi, il arrive le problème de maximisation d'incertitude dans ces cas.

Le principe de maximisation de l'incertitude dans la TIG peut se réaliser en quatre façon différentes, en fonction de la fonctionnelle à maximiser dans le processus d'optimisation :

- la mesure de Hartley généralisée ;
- la mesure de Shannon généralisée ;
- la mesure agrégée de l'incertitude totale (problème : cette fonctionnelle est très peu sensitive aux changements de l'évidence) ;
- la mesure désagrégée de l'incertitude totale.

Le choix de la fonctionnelle à maximiser est dépendante du contexte de chaque application. Malheureusement ces problèmes d'optimisation ne sont pas bien développés à l'heure actuelle, des exemples d'application pour des situations simples pouvant être retrouvés dans [Klir 06a].

A.9.3 Le principe de la généralisation exigée

La base de ce principe est de ne pas faire de choix *a priori* vis-à-vis de la théorie de l'incertitude utilisée, mais de bien analyser l'application, de connaître les besoins et seulement après de choisir la théorie qui se prête le mieux à ces exigences.

Ce principe est devenu possible avec le développement de la TIG. À l'heure actuelle il existe beaucoup de théories mathématiques traitant l'incertain. Comme décrit précédemment, toutes ces théorie de l'incertain peuvent être caractérisées par une certaine unicité et donc, on peut travailler avec la TIG dans sa globalité et non plus avec une seule théorie qui fait partie de la TIG. Dans ce cas, on peut passer d'une théorie à une autre pendant le traitement d'une application quelconque. Les motifs pour changer la théorie sont :

- la théorie actuelle n'est pas suffisamment générale pour englober les différents types d'incertitude présentes. Ainsi c'est nécessaire d'utiliser une théorie plus générale (un exemple étant le passage de la théorie des probabilités à la théorie de l'évidence).
- la théorie utilisée n'est plus adaptée à l'étape courante de l'application (par exemple quand la complexité de calcul est très grande).

Une observation peut être faite lorsqu'on veut passer à une théorie plus générale. Le choix de la théorie plus générale que celle actuelle n'est pas optionnelle, mais elle dépend de l'incertitude qui doit être intégrée. Ce principe a été introduit par Klir dans [Klir 06a] et sauf quelques exemples simples pour lesquels il a été testé, son application dans la pratique reste discutable.

B Introduction aux techniques de fusion d'informations

B.1 Le vote

C'est une technique de fusion des données intuitive et simple à mettre en place. Elle est employée surtout dans les cas de fusion de décisions, chaque source proposant pour fusionner un ensemble fini de décisions. Ces décisions sont, en général, accompagnées par un indice de fiabilité, de confiance. L'opération de combinaison de toutes ces décisions est basée sur un système de vote. En entrée du module de fusion sont présentées l'ensemble de toutes les décisions proposées et en sortie une liste de décisions sera générée, chaque décision ayant un poids égal au nombre total de fois qu'elle a été proposée par les sources.

Les inconvénients de cette technique sont les situations semblables au cas présenté dans le tableau 8.6. Ainsi, dans cette situation, le module va mettre en première position les décision D_1 et D_2, avec trois votes et sur les positions deux et trois les décisions D_3 et D_4 avec un vote. Par conséquent, cette technique n'a pas une bonne tolérance aux conflits entre les sources et ne prends pas en compte les imperfections des données. Un autre inconvénient est l'impossibilité de fusionner des données hétérogènes. Par exemple, les décisions exprimées sous une forme linguistique sont difficile à être transformées pour les fusionner avec d'autres décisions utilisant le vote.

B.2 Exploitant le réseau des sources

Dans [Barecke 11] une nouvelle approche de fusion a été proposée, prenant en compte les relations d'affinité et d'hostilité entre les sources. Pour la modélisation un graphe non-orienté est construit, avec des noeuds les sources et les arcs représentant les deux relations possibles entre deux sources.

Après la construction du graphe il va être partionné en sous-graphe avec les contraintes : à l'intérieur d'un sous-graphe toutes les sources soient en relation d'amitié et les sous-graphes soient liés entre elles par des liaisons d'hostilité. Avec ce partionnement fait et tenant compte de l'hypothèse que les sources en relation d'hostilité disant la même chose apporte plus d'information que les sources amies en accord, la fusion est faite en deux étapes : fusion intra-groupe et après fusion inter-groupes.

Mais cette fusion a quelques points négatifs. Premièrement l'association des relations entre les sources doit se faire par un expert et elle reste statique : si un changement apparaît (nouvelle source, changement de relations, etc.), il est difficile à prendre en compte. Deuxièmement, les relations entre les sources peuvent être différentes en fonction de la décision à prendre. Troisièmement, le partitionnement ne peut pas se faire toujours respectant toutes les contraintes.

B.3 La fusion probabiliste (bayésienne)

Cette technique de fusion de données utilise la théorie des probabilités, une théorie mathématique rigoureux et beaucoup analysé. Dans le cadre de cette théorie les imperfections des données, les incertitudes, sont modélisées par des distributions de probabilités.

La puissance de cette technique réside dans l'utilisation des probabilités conditionnelles, $p(D_i|S_j)$ et le pouvoir d'intégrer des connaissances *a priori* en utilisant la formule de Bayes :

$$p(D_i|S_j) = \frac{p(S_j|D_i)p(D_i)}{p(S_j)} \tag{B.1}$$

Ces probabilités peuvent être déduites en utilisant une approche fréquentielle dans le domaine discret et en utilisant des mélanges de distributions de probabilités, comme des gaussiennes par exemple, pour approximer la vraie distribution $p(D_i|S_j)$.

Étant donné que chaque source S_j accompagne chaque décision D_i par une probabilité de la forme $p(D_i|S_j)$, toutes ces probabilités vont être mises sous une forme vectorielle :

$$p_j = (p(D_1|S_j), p(D_2|S_j), \quad \ldots \quad , p(D_n|S_j)) \tag{B.2}$$

Pour fusionner les décisions proposées par chaque source, une matrice ayant comme lignes les vecteurs de probabilités (B.2) sera construite. Comme opérateurs de combinaison, actionnant sur les colonnes de cette matrice, peuvent être utilisés le maximum, le minimum, la moyenne, la médiane, l'oracle, etc, voir [Xu 92] pour plus des détails.

L'inconvénient de cette technique est le besoin d'avoir une grande base d'apprentissage pour bien estimer les probabilités. De plus, l'emploi de cette approche probabiliste implique l'existence d'un ensemble de décisions exclusives et exhaustives, ce qui entraine une impossibilité de modéliser les imprécisions des données. Un autre inconvénient est la difficulté de modéliser des connaissances qui ne peuvent pas être exprimées sous une forme probabiliste ou de modéliser l'absence des connaissances.

Soit les décisions à discriminer $D = \{D_1, D_2, \ldots, D_N\}$ étant données les mesures faites : $\{\mu_1, \ldots, \mu_M\}$. La règle de Bayes nous dit qu'il faut maximiser :

$$P(d_i|\mu_1, \ldots, \mu_M) = \frac{\left[\prod_j P(\mu_j|d_i)\right] P(d_i)}{\sum_k \left\{\prod_j P(\mu_j|d_k)P(d_k)\right\}} \tag{B.3}$$

Après le calcul de toutes ces probabilités *a posteriori*, le choix final de la décision optimale est donné en maximisant :

$$d_i^* = arg\,\{max_i[P(d_i|\mu_1, \ldots, \mu_M)]\} \tag{B.4}$$

C'est une technique intéressante grâce à :
- sa simplicité : calcul des N probabilités et après maximisation ;
- bien adaptée pour le traitement des mesures.

Les inconvénients de cette technique :
- besoin de connaître tous les modèles de densité de probabilité ;
- nécessité d'avoir des connaissances *a priori* ;
- travail sur des singletons et donc, impossibilité de modélisation des objets qui appartiennent à des classes différentes ;
- impossibilité de prendre en compte les imprécisions, les ambigüités, etc. ;
- une des axiomes de la théorie probabiliste est : $P(A) + P(A^c) = 1$. Ainsi si $P(A)$ augmente, $P(A^c)$ va diminuer. Cette caractéristique est une propriété de base pour la théorie probabiliste : une seule décision est vraie.

Le filtrage de Kalman peut être aussi vu comme un algorithme de fusion bayesien mais dans lequel la fusion est faite dans le temps.

En sortie du module de fusion chaque décision sera caractérisée par un coefficient de fiabilité, de confiance. Ainsi ce que nous cherchons est de retrouver ce coefficient pour chaque décision possible. Par la suite la valeur d'une décision va signifier la valeur de ce coefficient de fiabilité.

Maintenant on fera une hypothèse : les N sources fournissent à leurs sorties des informations, pour chaque décision, sous la forme d'un coefficient de confiance, qui à son tour est exprimé par un intervalle de confiance I_i, accompagné d'un coefficient de confiance, β_i ($0 < i \leq N$ pour la $i^{\text{ème}}$ source). Pour simplifier l'exposé, la fusion pour une seule décision sera traitée. Ainsi, en sortie de chaque source i nous avons accès à :

- son intervalle de confiance : $I_i = [a_i, b_i], \quad 0 \leq a_i \leq b_i \leq 1, \quad i \leq N$
- le coefficient de confiance de cet intervalle : $\beta_i = P_i(D \in I_i), \quad i \leq N$

Dans [Zhu 03] une technique d'estimation d'un paramètre utilisant un système multi-sources est décrite. Cette technique est adaptée par la suite pour le cas traité : proposition d'un ensemble de décisions accompagnées par des coefficients de confiance.

Ainsi la vraie valeur de la décision d est soit dans I_i soit dans $I_i^c \triangleq [0, a_i) \cup (b_i, 1]$, qui peut être vu comme un intervalle avec un coefficient de confiance donné par :

$$1 - \beta_i = P_i(d \in I_i^c), \quad i \leq N \tag{B.5}$$

Il faut remarquer que ce n'est pas possible de savoir quel est le coefficient de confiance de chaque intervalle qui compose I_i^c. Ainsi chaque source nous fournit en sortie deux intervalles :

$$I_i^1 \triangleq I_i, \qquad I_i^0 \triangleq I_i^c \tag{B.6}$$

avec les coefficients de confiance :

$$\beta_i^1 \triangleq \beta_i, \qquad \beta_i^0 \triangleq 1 - \beta_i \tag{B.7}$$

Avec ceux-ci définis, le module de fusion va recevoir les informations suivantes :

- une liste d'intervalles fournie par les N sources : $I = \{I_1^{r_1}, \dots I_N^{r_N}\}, r_i \in \{0, 1\}$
- une liste de coefficients de confiance : $\beta = \{\beta_1^{r_1}, \dots \beta_N^{r_N}\}, r_i \in \{0, 1\}$

Sous l'hypothèse d'indépendance entre les sources, la fusion sera le résultat de :

1. **Combinaison des intervalles** :

   La combinaison des intervalles au niveau du module de fusion est le résultat de toutes les intersections possibles et de toutes les réunions connectées possibles (qui ne donnent pas l'ensemble vide) entre les intervalles de la liste I :

$$I_{\{1^{r_1}, \dots, N^{r_N}\}} = \bigcap_{i=1}^{N} I_i^{r_i}, \quad r_i \in \{0, 1\} \tag{B.8}$$

$$I_{\cup_{r=1}^{R}\{1^{r_1}, \dots, N^{r_N}\}} = \bigcup_{r=1}^{R} \bigcap_{i=1}^{N} I_i^{r_i}, \quad r_i \in \{0, 1\} \tag{B.9}$$

2. **Combinaison des coefficients de confiance** :

   Les coefficients de confiance suite à l'opération de combinaison, pour les intervalles non-vides, sont donnés par :

$$P(d \in I_{\{1^{r_1}, \dots, N^{r_N}\}} | \mathcal{C}) = \frac{1}{c} \prod_{i=1}^{N} \beta_i^{r_i}, \quad r_i \in \{0, 1\} \tag{B.10}$$

   où le symbole $\mathcal{C}$ indique les ensembles non-vides :

$$\mathcal{C} = \{I_{\{1^{r_1}, \dots, N^{r_N}\}} : I_{\{1^{r_1}, \dots, N^{r_N}\}} \neq \varnothing\} \tag{B.11}$$

et le paramètre c donné par :

$$c = \sum_{I_{\{1^{r_1},\ldots,N^{r_N}\}} \neq \varnothing} \prod_{i=1}^{N} \beta_i^{r_i}, \quad r_i \in \{0,1\} \tag{B.12}$$

et avec :

$$P(D \in I_{\{1^{r_1},\ldots,N^{r_N}\}} | \mathcal{C}) = 0 \quad \text{si} \quad I_{\{1^{r_1},\ldots,N^{r_N}\}} = \varnothing \tag{B.13}$$

3. **La sommabilité des coefficients de confiance**
   Comme les intervalles I_i^0 et I_i^1 sont disjoints $(\forall)i$, toutes les intersections non-vides sont aussi disjointes et le coefficient de confidence pour une telle intersection non-vide est donné par :

$$P(D \in \bigcup_{i=1}^{N} I_{\{1^{r_1},\ldots,N^{r_N}\}} | \mathcal{C}) = \frac{1}{c} \sum_{r=1}^{R} \prod_{i=1}^{N} \beta_i^{r_i}, \quad r_i \in \{0,1\} \tag{B.14}$$

Exemple de fusion probabiliste des intervalles

Considérons le cas des trois sources qui fournissent pour la même décision les informations du tableau B.1.

	S_1	S_2	S_3
I_i^1	[0.2,0.5]	[0.3,0.6]	[0.4,0.7]
β_i^1	0.8	0.83	0.85
I_i^0	$[0,0.2) \cup (0.5,1]$	$[0,0.3) \cup (0.6,1]$	$[0,0.4] \cup (0.7,1]$
β_i^0	0.2	0.17	0.15

TABLE B.1: Exemple de fusion probabiliste des intervalles

Pour cet exemple, les intersections des intervalles sont :

$$\begin{aligned}
&I_{\{1^{r_1},\ldots,N^{r_N}\}} = \{ \\
&[0.2,0.5] \cap [0.3,0.6] \cap [0.4,0.7] \Rightarrow [0.4,0.5] \\
&[0.2,0.5] \cap [0.3,0.6] \cap ([0,0.4] \cup (0.7,1]) \Rightarrow [0.3,0.4) \\
&[0.2,0.5] \cap ([0,0.3) \cup (0.6,1]) \cap [0.4,0.7] \Rightarrow \varnothing \\
&[0.2,0.5] \cap ([0,0.3) \cup (0.6,1]) \cap ([0,0.4] \cup (0.7,1]) \Rightarrow [0.2,0.3) \\
&([0,0.2) \cup (0.5,1]) \cup [0.3,0.6] \cup [0.4,0.7] \Rightarrow (0.5,0.6] \\
&([0,0.2) \cup (0.5,1]) \cup [0.3,0.6] \cup ([0,0.4] \cup (0.7,1]) \Rightarrow \varnothing \\
&([0,0.2) \cup (0.5,1]) \cap ([0,0.3) \cup (0.6,1]) \cap [0.4,0.7] \Rightarrow (0.6,0.7] \\
&([0,0.2) \cup (0.5,1]) \cap ([0,0.3) \cup (0.6,1]) \cap ([0,0.4] \cup (0.7,1]) \Rightarrow [0,0.2) \cup (0.7,1]\}
\end{aligned}$$

En appliquant la formule (B.12), nous obtenons que $c = 0.8595$ et les coefficients des nouveaux intervalles obtenus par fusion sont présentés dans le tableau B.2 (le calcul faisant appel à la formule (B.10)) :

Intervalle	$[0,0.2) \cup (0.7,1]$	$[0.2,0.3)$	$[0.3,0.4)$	$[0.4,0.5]$	$(0.5,0.6]$	$(0.6,0.7]$
Coeff. confiance	0.0059	0.0237	0.1159	0.6567	0.1642	0.0336

TABLE B.2: Les intervalles fusionnés

D'après la formule (B.14), les intervalles avec leurs coefficients de confiance sont sommables et ainsi, pour les intervalles initiaux proposés par les sources, on a :
$[0.2, 0.5] : 0.7963$; $[0.3, 0.6] : 0.9368$; $[0.4, 0.7] : 0.8545$

De ce résultat, nous pouvons remarquer que le coefficient de confiance pour l'intervalle $[0.3, 0.6]$ a augmenté suite à l'opération de fusion, tandis que pour l'intervalle $[0.2, 0.5]$ le coefficient de confiance a diminué face à celui proposé par la première source.

Maintenant, la situation avec des sources qui proposent des intervalles avec leur intersection vide sera traitée ($I_{\{1^{r_1},\ldots,N^{r_N}\}} = \varnothing$). Prenons l'exemple du tableau B.3. Dans ce cas la première source propose un intervalle qui ne se chevauche pas avec les autres deux intervalles proposés par les sources S_2 etS_3.

	S_1	S_2	S_3
I_i^1	[0.1,0.3]	[0.4,0.6]	[0.5,0.8]
β_i^1	0.8	0.83	0.85
I_i^0	$[0, 0.1) \cup (0.3, 1]$	$[0, 0.4) \cup (0.6, 1]$	$[0, 0.5) \cup (0.8, 1]$
β_i^0	0.2	0.17	0.15

TABLE B.3: Exemple de fusion probabiliste des intervalles disjoints

Dans cette situation, les nouveaux intervalles suite à l'opération de fusion avec leurs coefficients de confiance sont présentés dans le tableau B.4

Intervalle	$[0, 0.1) \cup (0.3, 0.4) \cup (0.8, 1]$	$[0.1, 0.3)$	$[0.4, 0.5)$	$[0.5, 0.6]$	$(0.6, 0.8]$
Coeff. confiance	0.0231	0.0926	0.1130	0.6402	0.1311

TABLE B.4: Les intervalles disjoints fusionnés

À partir des résultats du tableau B.4, nous pouvons obtenir d'autres intervalles avec leurs coefficients de confiance, mais cette fois-ci ces coefficients étant caractérisés par leur borne inférieure (par exemple : $[0.1, 0.6] \geq 0.8458$).

Parmi les avantages de l'utilisation du module de fusion sont :

1. Plusieurs intervalles de couverture variable peuvent être obtenus accompagnés avec leurs coefficients de confiance. Ainsi le module de fusion offre la possibilité d'avoir plusieurs résolutions.
2. Le fusion multi-sources permet d'augmenter ou de diminuer la confiance dans le résultat d'une certaine source.

Observation : La formule de combinaison donnée par (B.10) a comme idée de base la règle de combinaison de Dempster-Shafer dans le cadre de la théorie de l'évidence. La seule différence avec la règle de combinaison de Dempster-Shafer est que cette dernière travaille avec des coefficients de confiance sous la forme de masse d'évidence et elle n'est pas additive. Pour plus d'informations sur la théorie de Dempster-Shafer voir le paragraphe A.3 et sur la règle de Dempster-Shafer voir le paragraphe B.4.1.

Maintenant si nous revenons à la situation du tableau B.2 et deux critères d'optimisation sont imposés : minimisation de la couverture de l'intervalle sous la contrainte du coefficient de confiance et la maximisation du coefficient de confiance sous la contrainte de la couverture de l'intervalle de confiance, le module de fusion peut fournir en sortie des résultat de la forme présentée dans le tableau B.5

Contrainte coeff. de confiance	≥ 0.6	≥ 0.8	≥ 0.9
Contrainte couverture intervalle	≤ 1	≤ 2	≤ 3
L'intervalle optimal	$[0.4, 0.5]$	$[0.4, 0.6]$	$[0.3, 0.6]$
Coefficient de confiance	0.6567	0.8209	0.9368
La longueur de l'intervalle	1	2	3

TABLE B.5: Les résultats de la fusion d'intervalles sous des contraintes

Une autre situation à discuter est celle où le module de fusion n'a qu'une connaissance parmi les N sources et il y a au plus $K < N$ qui ne sont pas fiables mais sans connaître lesquelles. Avec cette hypothèse, les équations (B.10)-(B.13) deviennent :

$$P(D \in I_{\{1^{r_1},\ldots,N^{r_N}\}}|\mathcal{C}_2) = \frac{1}{c_0}\prod_{i=1}^{N}\beta_i^{r_i}, \quad r_i \in \{0,1\} \tag{B.15}$$

$$\mathcal{C}_2 = \{I_{\{1^{r_1},\ldots,N^{r_N}\}} : I_{\{1^{r_1},\ldots,N^{r_N}\}} \neq \varnothing, \sum_{i=1}^{N} r_i \geq N-K \tag{B.16}$$

$$c_0 = \sum_{\substack{I_{\{1^{r_1},\ldots,N^{r_N}\}}\neq\varnothing \\ \sum_{i=1}^{N} r_i \geq N-K}} \prod_{i=1}^{N}\beta_i^{r_i}, \quad r_i \in \{0,1\} \tag{B.17}$$

$$P(D \in I_{\{1^{r_1},\ldots,N^{r_N}\}}|\mathcal{C}_2) = 0 \quad \text{si} \quad I_{\{1^{r_1},\ldots,N^{r_N}\}} = \varnothing \quad \text{ou} \quad \sum_{i=1}^{N} r_i < N-K \tag{B.18}$$

Afin d'exemplifier la fusion avec des sources non-fiables, le cas du tableau B.1 sera de nouveau traité. Pour $K = 1$ nous avons :

$$\begin{aligned}
&I_{\{1^{r_1},\ldots,N^{r_N}\}} = \{ \\
&[0.2, 0.5] \cap [0.3, 0.6] \cap [0.4, 0.7] \Rightarrow [0.4, 0.5] \\
&[0.2, 0.5] \cap [0.3, 0.6] \cap ([0, 0.4] \cup (0.7, 1]) \Rightarrow [0.3, 0.4) \\
&([0, 0.2) \cup (0.5, 1]) \cup [0.3, 0.6] \cup [0.4, 0.7] \Rightarrow (0.5, 0.6]\}
\end{aligned}$$

les autres intersections étant vides. Ainsi, pour le paramètre de normalisation c_0, nous avons $c_0 = 0.5644 + 0.0996 + 0.1411 = 0.8051$. Dans le tableau B.6 les résultats pour $K = 1, 2, 3$ sont présentés.

f	$[0, 0.2) \cup (0.7, 1]$	$[0.2, 0.3)$	$[0.3, 0.4)$	$[0.4, 0.5]$	$(0.5, 0.6]$	$(0.6, 0.7]$
3	0.0059	0.0237	0.1159	0.6567	0.1642	0.0336
2	0	0.0239	0.1166	0.6606	0.1651	0.0338
1	0	0	0.1237	0.7010	0.1753	0

TABLE B.6: Les intervalles fusionnés

Ainsi si K est petit, les degrés de confiances vont se repartir sur un nombre plus petit d'intervalles, en rendant ces intervalles plus crédibles. Les mêmes contraintes sur la couverture des intervalles et sur la valeur minimale du coefficient de confiance peuvent s'imposer pour avoir en sortie la décision la plus crédible.

L'intérêt de cette technique de fusion est qu'elle peut s'appliquer dans le cadre des sources qui s'expriment utilisant des variables linguistiques. Par exemple pour le cas d'une source qui s'exprime par : *impossible, très faible, faible, moyen, fort, très fort, sûr* les intervalles de confiance peuvent s'exprimer sous la forme :

impossible, très faible, faible, moyen, fort, très fort, sûr
$\Rightarrow [0, 0.2], [0.1, 0.3], [0.2, 0.4], [0.35, 0.65], [0.6, 0.8], [0.7, 0.9], [0.8, 1]$
Observations :

1. dans tous les exemples traités, l'hypothèse que toutes les sources donnent une information complète est implicitement considérée comme vraie. Ainsi, quand une source ne donne pas d'informations sur une décision cela implique que pour cette décision la source fournie avec un coefficient de confiance égal à 1 que 0 est la valeur caractérisant la décision. Mais si l'hypothèse que cette source a une vision partielle est considérée, le cas où cette source ne s'exprime pas vis-à-vis d'une décision implique une représentation numérique de la forme : $I_i = [0, 1]$ avec un coefficient de confiance égal à 1. Par conséquent, la source est dans un état d'ignorance totale, pouvant ne pas être considérée pour le problème de fusion.

B.4 La fusion des croyances dans la théorie de Dempster-Shafer

Maintenant plusieurs modalités de fusion des fonctions de croyances issues des sources considérées indépendantes seront présentées (pour une description plus détaillée de ces techniques de combinaison et pour des comparaison entre elles voir [Smarandache 04]) :

B.4.1 La règle de Dempster

$$m(.) = [m_1 \oplus m_2](.) = \begin{cases} m(\oslash) & = 0 \\ m(A) & = \dfrac{\sum\limits_{X \cap Y = A} m_1(X) m_2(Y)}{1 - \sum\limits_{X \cap Y = \oslash} m_1(X) m_2(Y)} \end{cases} \qquad \forall (A \neq \oslash) \in 2^{\Theta} \tag{B.19}$$

La quantité $k_{12} = \sum_{X \cap Y = \oslash} m_1(X) m_2(Y)$ s'appelle degré de conflit entre le deux sources et il a le rôle d'éliminer les parties d'informations en contradiction. Cette règle a un bon comportement sauf les cas où les sources sont en fort conflit et peut être facilement généralisée pour la combinaison de N sources. Comme cas particulier dans lequel cette règle n'est pas fiable c'est la situation présentée dans le tableau 8.6.

B.4.2 La règle de disjonctive (Dubois et Prade 1986)

$$\begin{cases} m_{\cup}(\oslash) & = 0 \\ m_{\cup}(A) & = \sum\limits_{X \cup Y = A} m_1(X) m_2(Y) \end{cases} \qquad \forall (A \neq \oslash) \in 2^{\Theta} \tag{B.20}$$

Cette règle, proposée dans [Dubois 88], réalise un consensus disjonctif et elle est employée dans des cas où une des sources n'est pas fiable mais sans savoir laquelle.

B.4.3 La règle de Murphy

$$Bel_M(A) = \frac{Bel_1(A) + Bel_2(A)}{2} \tag{B.21}$$

[Murphy 00] propose une moyenne arithmétique des fonctions de croyances. Il s'agit d'une règle commutative mais pas associative.

B.4.4 La règle de Smets

$$\begin{cases} m_S(\oslash) & = \sum_{X \cap Y = \oslash} m_1(X) m_2(Y) = k_{12} \\ m_S(A) & = \sum_{X \cap Y = A} m_1(X) m_2(Y) \qquad \forall (A \neq \oslash) \in 2^\Theta \end{cases} \tag{B.22}$$

[Smets 94b] propose une règle commutative et associative qui ressemble beaucoup à la règle de Dempster sauf que dans cette situation il n'y a pas de normalisation. Comme particularité, l'ensemble vide peut avoir une masse non-nulle, le conflit entre les deux sources lui étant transféré.

B.4.5 La règle de Yager

$$\begin{cases} m_Y(\oslash) & = 0 \\ m_Y(A) & = \sum_{X \cap Y = A} m_1(X) m_2(Y) \quad \forall (A \notin \{\oslash, \Theta\}) \in 2^\Theta \\ m_Y(\Theta) & = m_1(\Theta) m_2(\Theta) + \sum_{X \cap Y = \oslash} m_1(X) m_2(Y) \quad A = \Theta \end{cases} \tag{B.23}$$

Par rapport à la règle de Smets, cette nouvelle règle [Yager 87b] traite le conflit entre les deux sources de manière différente le transférant à l'ensemble total (de tous les singletons, Θ qui représente l'ignorance totale.

B.4.6 La règle de Dubois et Prade

$$\begin{cases} m_{DP}(\oslash) & = 0 \\ m_{DP}(A) & = \sum_{X \cap Y = A; X \cap Y \neq \oslash} m_1(X) m_2(Y) + \sum_{X \cup Y = A; X \cap Y \neq \oslash} m_1(X) m_2(Y) \quad \forall (A \neq \oslash) \in 2^\Theta \end{cases} \tag{B.24}$$

Cette formule de combinaison [Dubois 88] traduit le fait que deux sources sont considérées comme fiable quand elles ne sont pas en conflit (la première somme) et dans le cas de conflit, une d'entre elles est fiable (deuxième somme).

B.4.7 La règle de Dezert et Smarandache (DSm)

Toutes les règles présentées jusque maintenant étaient construites sur le modèle de Shafer supposant un ensemble fini, exhaustif et exclusif ; le complément de chaque objet A appartient à l'ensemble de tous les sous-ensembles : $2^\Theta \equiv (\Theta, \cup)$ et ayant comme point de départ la règle de combinaison de Dempster. Dans [Smarandache 04], Dezert et Smarandache ont proposé une nouvelle théorie mathématique pour résoudre le problème de fusion en éliminant les trois hypothèses de départ de la théorie de Dempster-Shafer. Ainsi les nouvelles hypothèses de départ sont :

– ensemble *fini* et *exhaustive*. L'hypothèse d'exclusivité est éliminée en gardant celle d'exhaustivité (hypothèse peu limitante car tout ensemble d'hypothèses non-exhaustives peut être fermé en lui ajoutant une hypothèse de fermeture) ;
– l'utilisation du l'ensemble de tous les sous-ensembles, 2^Θ est remplacée par l'utilisation de la lattice de Dedekind : $D^\Theta \equiv (\Theta, \cup, \cap)$;
– $A \in D^\Theta \not\Longrightarrow A^c \in D^\Theta$ Mais si cette condition devra être possible le D^Θ peut se remplacer par $S^\Theta \equiv (\Theta, \cup, \cap, c(.))$.

Comme dans les autres cas, il est très important de bien définir le cadre Θ du problème, mais cette fois comme la règle DSm est une généralisation de celle de DS, pour la modélisation il faut choisir l'ensemble dans lequel les fonctions de croyance vont se définir : 2^Θ ou D^Θ ou S^Θ. Comme règle de fusion Dezert et Smarandache propose :

– la règle libre DSm

$$m_{\mathcal{M}^f(\Theta)}(A) = \sum_{X \cap Y = A} m_1(X) m_2(Y) \quad ; \quad A, X, Y \in D^\Theta \tag{B.25}$$

cette règle de combinaison a deux grands avantages : est simple à mettre en place et il n'y a pas de normalisation. Cette règle peut se généralisée facilement pour le cas avec n sources à combiner :

$$m_{\mathcal{M}^f(\Theta)}(A) = \sum_{\substack{X_1,\ldots,X_k \in D^\Theta \\ X_1 \cap \ldots \cap X_k = A}} \prod_{i=1}^{k} m_i(X_i) \tag{B.26}$$

– la règle hybride DSm :

$$m_{\mathcal{M}(\Theta)}(A) = \Phi(A)[S_1(A) + S_2(A) + S_3(A)] \tag{B.27}$$

avec les notations : $\boldsymbol{\phi} = \{\boldsymbol{\phi}_\mathcal{M}, \phi\}$; $\mathcal{U} = u(X_1) \cup \ldots \cup u(X_k)$ où $u(X)$ est l'union de tous les singletons θ_i qui compose X et avec l'ignorance totale $\mathcal{I}_t = \theta_1 \cup \ldots \cup \theta_n$ les termes de la règle hybride DSm (B.27) s'écrivent sous la forme :

$$\Phi(A) = \begin{cases} 1 & \text{si } A \notin \boldsymbol{\phi} \\ 0 & \text{si } A \in \boldsymbol{\phi} \end{cases} \tag{B.28}$$

$$S_1(A) = \sum_{\substack{X_1,\ldots,X_k \in D^\Theta \\ X_1 \cap \ldots \cap X_k = A}} \prod_{i=1}^{k} m_i(X_i) \tag{B.29}$$

$$S_2(A) = \sum_{\substack{X_1,\ldots,X_k \in \boldsymbol{\phi} \\ (\mathcal{U}=A) \vee [(\mathcal{U} \in \boldsymbol{\phi}) \wedge (A = \mathcal{I}_t)]}} \prod_{i=1}^{k} m_i(X_i) \tag{B.30}$$

$$S_3(A) = \sum_{\substack{X_1,\ldots,X_k \in D^\Theta \\ X_1 \cup \ldots \cup X_k = A \\ X_1 \cap \ldots \cap X_k \in \boldsymbol{\phi}}} \prod_{i=1}^{k} m_i(X_i) \tag{B.31}$$

La règle libre DSm peut être employée quand il n'y a pas des contraintes (exclusivité, appartenance à l'ensemble vide, etc.) sur les éléments de l'ensemble D^Θ. Mais il peut arriver que des nouveaux éléments devront être inclus dans l'ensemble de définition et que certains éléments de D^Θ ne sont plus possibles et seront passés à l'ensemble vide $\boldsymbol{\phi}$; si nous voulons prendre en compte cette dynamique du problème la règle hybride DSm est envisageable.

	S1	S2
D_1	0.99	0
D_2	0	0.99
D_3	0.01	0.01

TABLE B.7: Exemple de données à fusionner

	d_1	d_2	d_3	$d_1 \cap d_2$	$d_1 \cap d_3$	$d_2 \cap d_3$	$d_1 \cap d_2 \cap d_3$
m_{DSm}	0	0	0.0001	0.9801	0.099	0.099	0

TABLE B.8: Application de la règle classique DSm

B

B.4.8 Fusion des évidences imprécises dans le cadre de la théorie DSm

La théorie DSm peut être généralisée au cas où les sources d'évidence peuvent fournir la masse d'évidence que sous la forme d'un intervalle (ou plus généralement d'un ensemble d'intervalles) sous-unitaire, $m^I()$. Ces intervalles doivent respecter la condition d'admissibilité, c.à.d. :
$\forall X \in D^\Theta \quad (\exists) m(X) \in m^I(X) \quad \text{t.q.} \quad \sum_{X \in D^\Theta} m(X) = 1$

Pour la définition des opérations avec des intervalles voir ch. 6 de [Smarandache 04]. Avec ces opérateurs les deux règles de fusion DSm s'écrivent sous la même, mais avec la somme et le produit qui s'appliquent à des intervalles de masse d'évidence. Par exemple la règle classique DSm s'écrit :

$$m^I(A) = \boxed{\sum}_{\substack{X_1,\ldots,X_k \in D^\Theta \\ X_1 \cap \ldots \cap X_k = A}} \boxed{\prod}_{i=1}^{k} m_i(X_i) \tag{B.32}$$

Exemplification de la fusion d'évidence Avec toutes ces règles de combinaison prenons l'exemple du tableau B.7 avec l'hypothèse d'exclusivité entre les décisions. Si la règle Dempster-Shafer B.19 est appliquée les deux premières décisions vont avoir une masse nulle en sortie et la troisième décision sera celle choisie. En appliquant maintenant la règle de Dezert-Smarandache B.25 nous obtenons les résultats du tableau B.8.

De ce tableau, nous pouvons tirer la conclusion que la règle de combinaison de Dezert-Smarandache sort comme gagnantes à la fois les décisions D_1 et D_2. C'est un résultats important, qui pourrait être exploité si l'objectif du système n'est de proposer en sortie une seule décision. Néanmoins, dans la théorie DSm la masse d'évidence $A \cap B$ est vue comme un conflit partielle, car cette théorie est faite pour trouver *la meilleure décision*.

Une autre observation, toujours dans le cas de l'hypothèse d'exclusivité, est que la règle de combinaison DSm B.26 sort une masse non-nulle pour au maximum n décisions ensemble $D_1 \cap D_2 \cap \ldots \cap D_n$ et pour toute combinaison de $n+1$ décisions la masse sera nulle.

B.5 FUSION DANS LA THÉORIE DES POSSIBILITÉS

Les opérateurs de combinaison pour ce type de fusion ont l'avantage, face aux autres deux : probabiliste et d'évidence, qu'ils sont très souples et permettent de fusionner des informations de natures différentes et dans des contextes différents.

Normes et conormes triangulaires

Une norme triangulaire (t-norme - comportement sévère) est une fonction $t : [0,1] \times [0,1] \to [0,1]$ vérifiant les propriétés :

- commutativité : $\forall(x,y) \in [0,1]\times[0,1], t(x,y) = t(y,x)$
- associativité : $\forall(x,y,z) \in [0,1]\times[0,1]\times[0,1], t(t(x,y),z) = t(x,t(y,z))$
- 1 est l'élément neutre : $\forall x \in [0,1], t(x,1) = t(1,x) = x$
- croissante : $\forall(x,x^{'},y,y^{'}) \in [0,1]^4, (x \leq x^{'}, y \leq y^{'}), t(x,y) \leq t(x^{'},y^{'})$
- 0 est un élément absorbant : $\forall x \in [0,1], t(x,0) = t(0,x) = 0$

Exemples des t-normes classiques :

$$min(x,y), \quad xy, \quad max(0,X+y-1), \quad 1-min\left\{1,[(1-x)^p+(1-y)^p]^{\frac{1}{p}}\right\}$$

Ce dernier étant sous une forme paramétrique qui permet de construire des opérateurs en fonction de nos besoins. Par exemple pour $p=1$ l'opérateur $max(0,X+y-1)$ est obtenu.

Une t-conorme (comportement indulgent) est un autre opérateur obtenu à partir d'une t-norme et utilisant l'opérateur de complémentation (par exemple $\forall x \in [0,1], c(x) = 1-x$). La forme d'une T-conorme est donnée par $\forall(x,y) \in [0,1]\times[0,1], T(x,y) = c[t(c(x),c(y))]$. Les propriétés d'une t-conorme sont :

- commutativité : $\forall(x,y) \in [0,1]\times[0,1], T(x,y) = T(y,x)$
- associativité : $\forall(x,y,z) \in [0,1]\times[0,1]\times[0,1], T(t(x,y),z) = T(x,t(y,z))$
- 0 est l'élément neutre : $\forall x \in [0,1], t(x,0) = t(0,x) = x$
- croissante : $\forall(x,x^{'},y,y^{'}) \in [0,1]^4, (x \leq x^{'}, y \leq y^{'}), T(x,y) \leq T(x^{'},y^{'})$
- 1 est un élément absorbant : $\forall x \in [0,1], T(x,1) = T(1,x) = 0$

Exemples des t-conormes classiques : $max(x,y), \quad x+y-xy, \quad min(1,x+y)$.

Opérateurs de moyenne - comportement prudent

Un opérateur de moyenne est une fonction $m : [0,1]\times[0,1] \to [0,1]$ et qui vérifie :

- le résultat de cet opérateur est toujours compris entre : $\forall(x,y) \in [0,1]\times[0,1], min(x,y) \leq m(x,y) \leq max(x,y)$
- commutativité : $\forall(x,y) \in [0,1]\times[0,1], T(x,y) = T(y,x)$
- croissante : $\forall(x,x^{'},y,y^{'}) \in [0,1]^4, (x \leq x^{'}, y \leq y^{'}), m(x,y) \leq m(x^{'},y^{'})$

Exemple d'opérateurs de médianne :

- la médiane :

$$\forall(x,y,\alpha) \in [0,1]^3 m(x,y) = med(x,y,\alpha) = \begin{cases} x & \text{si} \quad y \leq x \leq \alpha \quad \text{ou} \quad \alpha \leq x \leq y \\ y & \text{si} \quad x \leq y \leq \alpha \quad \text{ou} \quad \alpha \leq y \leq x \\ \alpha & \text{si} \quad y \leq \alpha \leq x \quad \text{ou} \quad x \leq \alpha \leq y \end{cases}$$

- la moyenne : $\forall(x,y) \in [0,1]^2 m(x,y) = k^{-1}\left[\frac{k(x)+k(y)}{2}\right]$ avec k une fonction strictement croissante de $[0,1]$ dans $[0,1]$ Par exemple pour $k(x) = x^{\alpha}$ et $\alpha = 2$ la moyenne quadratique est obtenue : $m(x,y) = \sqrt{\frac{x^2+y^2}{2}}$.

Opérateurs adaptifs

Ce type d'opérateurs ont un comportement différent en fonction du conflit entre les sources. Ainsi si les sources fournissent des informations (sous la forme des distributions de possibilité) consonantes cet opérateur va actionner comme un *min* et si le conflit entre les deux sources est important le comportement de cet opérateur sera décrit par un *max*. Donc, ce type d'opérateur est bien adapté pour les cas des sources qui fournissent des informations partielles, qui est le cas de notre problème. Une mesure de conflit (suivant le modèle de Dempster-Shafer dans le cadre de la théorie de l'évidence) entre deux sources peut être décrite par :

$$h(\pi_1,\pi_2) = 1 - max_{c\in\Omega} min(\pi_1(c),\pi_2(c)) \tag{B.33}$$

Comme opérateurs de combinaison, il y a par exemple (pour plus d'opérateurs voir par exemple [Bloch 04]) :

$$\pi_{12} = max\left\{\frac{t(\pi_1(s),\pi_2(s))}{h(\pi_1,\pi_2)}, 1-h(\pi_1,\pi_2)\right\} \tag{B.34}$$

$$\pi_{12} = max\left\{\frac{min(\pi_1, \pi_2)}{h}, min[max(\pi_1, \pi_2), 1 - h]\right\} \tag{B.35}$$

Le premier opérateur (B.34) prend comme distribution finale le maximum entre un opérateur conjonctif normalisé et une constante fonction du degré de conflit. Le deuxième opérateur (B.35) par contre permet de passer d'une combinaison strictement conjonctive $\frac{min(\pi_1,\pi_2)}{h}$ quand le conflit est nul à une combinaison strictement disjonctive quand le conflit est très grand (l'opérateur max).

Maintenant si nous avons des connaissances sur la fiabilité des sources d'informations (par exemple la source délivrant π_1 est plus fiable que celle délivrant π_2), un bon compromis sera de prendre un opérateur conjonctif si les deux sources sont concordante et de ne pas prendre en compte la source délivrant π_2 dans le cas contraire :

$$\pi_{12} = min\{\pi_1, max[\pi_2, h(\pi_1, \pi_2)]\} \tag{B.36}$$

L'opérateur B.36 ne nécessite que de connaître une relation d'ordre entre les sources vis-à-vis de leur degré de fiabilité. En plus si des valeurs numériques sont disponibles caractérisant la fiabilité de chaque source, la distribution de possibilité pourrait être transformer dans une autre distribution représentant la fiabilité équivalente :

$$\pi_j^* = max\{\pi_j, 1 - \omega_j\} \tag{B.37}$$

Ainsi de cette formule, B.37 si la source j est totalement fiable, $\omega_j = 1$ et la nouvelle distribution prend la valeur de la distribution de possibilité $\pi_j^* = \pi_j$. Par contre si la source n'est pas fiable, $\omega_j \ll 1$, la distribution devient constante est égale à 1 représentant l'indifférence totale, tout élément pouvant être possible.

Références bibliographiques

[Aamodt 95] A. Aamodt & M. Nygård. *Different roles and mutual dependencies of data, information, and knowledge - an AI perspective on their integration.* Data and knowledge engineering, vol. 16, pages 191–222, 1995. 15, 17

[Aczel 06] J. Aczel. *Entropies, characterizations, applications and some history.* In Modern information processing : From theory to applications, eds. B. Bouchon-Meunier, G. Coletti et R.R. Yager, pages 3–10. Elsevier, 2006. 161

[Adriaans 96] P. Adriaans & D. Zantinge. Data mining. Addison-Wesley, 1996. 31

[Appriou 01] A. Appriou. *Situation assessment based on spatially ambigous multisensor measurements.* International Journal of Intelligent Systems, vol. 16, pages 1135 – 1166, 2001. 54, 110

[Atoyan 10] H. Atoyan, J-M. Robert & J-R. Duquet. *Human Factors Analysis of Different Types of Uncertainties in Complex Systems.* In Shahbazian E. & Rogova G., editeurs, Human Systems Integration to Enhance Maritime Domain Awareness for Port/Harbour Security, pages 61 – 68. IOS Press, 2010. 40, 54

[Baldwin 00] C. Y. Baldwin & K. B. Clark. Design rules : The power of modularity. MIT Press, Cambridge, MA, 2000. 7

[Ballou 95] D. Ballou & H. Pazer. *Designing Information Systems to Optimize the Accuracy-Timeliness Tradeoff.* Information Systems Research, vol. 6, no. 1, pages 51–72, 1995. 49

[Ballou 98] D. Ballou, R. Wang, H. Pazer & G. Tayi. *Modeling Information Manufacturing Systems to Determine Information Product Quality.* Management Science, vol. 44, no. 4, pages 462–484, April 1998. 48

[Ballou 03] D. Ballou & H. Pazer. *Modeling completeness versus consistency tradeoffs in information decision contexts.* IEEE Transactions on Knowledge and Data Engineering, vol. 15, no. 1, pages 240 – 243, jan.-feb. 2003. 48, 49, 69

[Barbara 92] D. Barbara, H. Garcia-Molina & D. Porter. *The management of probabilistic data.* Knowledge and Data Engineering, IEEE Transactions on, vol. 4, no. 5, pages 487–502, Oct 1992. 35, 36

[Barecke 11] T. Barecke, M-J. Lesot, H. Akdag & B. Bouchon-Meunier. *Stratégie de fusion d'informations exploitant le réseau des sources.* In 8ème Atelier sur la Fouille de Données Complexes, EGC2011, Brest, 2011. 165

[Batini 06] C. Batini & M. Scannapieco. Data quality : Concepts, methodologies and techniques. Springer-Verlag, 2006. 44

[Berkan 97] R. Berkan & S. Trubatch. Fuzzy systems design principles : Building fuzzy if-then rule bases. IEEE Press, 1997. 19

[Berti-Equille 06] L. Berti-Equille. *Qualité des données.* Techniques de l'ingénieur, 2006. 123

[Bisdikian 07] C. Bisdikian. *Quality of information trade-offs in the detection of transient phenomena.* In E. Carapezza, editeur, Unattended Ground, Sea, and Air Sensor Technologies and Applications IX, 2007. 91, 110

[Blasch 07] E. Blasch & S. Plano. *Evaluation of Information Fusion techniques Part 1 : System Level Assessment.* In Advances and Challenges in Multisensor Data and Information Processing, volume Sub-Series D : Information and Communication Security 8 of *NATO Security through Science Series*, pages 366–374. IOS Press, Amsterdam, 2007. 54

[Blasch 10] E. Blasch, P. Valin & É. Bossé. *Measures of effectiveness for high-level fusion.* International conference on information fusion (FUSION10), 2010. vii, 64, 65

[Bleiholder 09] J. Bleiholder & F. Naumann. *Data fusion.* ACM Comput. Surv., vol. 41, no. 1, pages 1 :1–1 :41, January 2009. 128

[Bloch 03] I. Bloch, editeur. Fusion d'information en traitement du signal et des images, chapitre Théorie des ensembles flous et des possibilités. Lavoisier, 2003. 147

[Bloch 04] I. Bloch. Fusion d'information en traitement du signal et des images. Lavoisier, Paris, 2004. 175

[Bonnisone 85] P. Bonnisone & R.M. Tong. *Editorial : reasoning with uncertainty in expert systems.* International Journal of Man-Machine Studies, vol. 22, pages 241–250, 1985. 57

[Borek 14] A. Borek, A. Kumar Parlikad, P. Woodall & M. Tomasella. *A risk based model for quantifying the impact of information quality.* Computers in Industry, vol. 65, no. 2, pages 354 – 366, 2014. 65, 140

[Borysowich 11] C. Borysowich. *Systems Design : Principles of Hierarchical Decomposition*, 2011. `http://it.toolbox.com/blogs/enterprise-solutions/systems-design-principles-of-hierarchical-decomposition-48204`, consulté le 12 septembre 2014. 74

[Bosc 96] P. Bosc & H. Prade. *An introduction to the fuzzy sets and possibility theory-based treatment of flexible queries and uncertain or imprecise databases.* In A. Motro & P. Smets, editeurs, Uncertainty Management in Information Systems, pages 285–324. Kluwer Academic Publishers, 1996. 37

[Bossé 06] É. Bossé, A. Guitouni & P. Valin. *An essay to characterise information fusion systems.* 9th Conference on Information Fusion (FUSION), 2006. 53

[Boulos 05] J. Boulos, N.N. Dalvi, B. Mandhani, S. Mathur, C. Re & D. Suciu. *MYSTIQ : a system for finding more answers by using probabilities.* In Proceedings of SIGMOD, page 891–893, 2005. 33

[Buckles 82] B. P. Buckles & F. E. Petry. *A fuzzy representation of data for relational databases.* Fuzzy Sets and Systems, vol. 7, pages 213–226, 1982. 36

[Buckles 84] B. P. Buckles & F. E. Petry. *Extending the fuzzy database with fuzzy numbers.* Information Sciences, vol. 34, no. 2, page 145–155, 1984. 37

[Calvo 02] T. Calvo, A. Kolesárová, M. Komorníková & R. Mesiar. *Aggregation Operators : Properties, Classes and Construction Methods.* In Tomasa Calvo, Gaspar Mayor & Radko Mesiar, editeurs, Aggregation Operators, volume 97 of *Studies in Fuzziness and Soft Computing*, pages 3–104. Physica-Verlag HD, 2002. 98

[Carnap 52] R. Carnap & Y. Bar-Hillel. *An outline of a theory of semantic information.* Rapport technique 247, Massachusetts Institute of Technology, Research Laboratory of Electronics, Cambridge, Massachusetts, 27 octobre 1952. 142

RÉFÉRENCES BIBLIOGRAPHIQUES

[Caseau 07] Y. Caseau, D. Krob & S. Peyronnet. *Complexité des systèmes d'information : une famille de mesures de la complexité scalaire d'un schéma d'architecture*. Génie logiciel, vol. 82, pages 23–30, 2007. 7, 73

[Charatan 99a] F. Charatan. *Family Compensated for Death after Illegible Prescription*. BMJ : British Medical Journal, vol. 319, no. 7223, page 1456, 1999. 40

[Charatan 99b] F. Charatan. *Medical errors kill almost 100000 Americans a year*. BMJ : British Medical Journal, vol. 319, no. 7224, page 1519, 1999. `http://www.ncbi.nlm.nih.gov/pmc/articles/PMC1117251/` consulté le 31 mai 2014. 39

[Chen 00] Guanling Chen & David Kotz. *A Survey of Context-Aware Mobile Computing Research*. Rapport technique, Dartmouth College, Hanover, NH, USA, 2000. 9

[Chen 01] Z. Chen. Data mining and uncertainty reasoning. John Wiley & Sons, 2001. 121

[Cheng 05] R. Cheng, S. Singh & S. Prabhakar. *U-DBMS : A database system for managing constantly-evolving data*. In Proceedings of VLDB, page 1271–1274, 2005. 33

[Christiansen 04] H. Christiansen & D. Martinenghi. *Simplification of Database Integrity Constraints Revisited : A Transformational Approach*. In M. Bruynooghe, editeur, Logic Based Program Synthesis and Transformation, volume LNCS 3018, pages 178–197. Springer-Verlag Berlin Heidelberg, 2004. 27

[Codd 70] E. F. Codd. *A Relational Model of Data for Large Shared Data Banks*. Communications of the ACM, vol. 13, no. 6, pages 377–387, June 1970. 26

[Collett 97] D. Collett. Modelling survival data in medical research. Chapman & Hall, Londres, 1997. 132

[Coskun 10] A. Coskun. Quality management and six sigma. Scyio, 2010. 68

[Costa 12] P. Costa, K. Laskey, E. Blasch & A-L. Jousselme. *Towards unbiased evaluation of uncertainty reasoning : The URREF ontology*. International conference on information fusion (FUSION12), 2012. 64

[Cykana 96] P. Cykana, A. Paul & M. Stern. *DoD guidelines on data quality management*. In Proceedings of the Conference on Information Quality (ICIQ), Cambridge, MA, page 154–171, 1996. 66, 67

[De Amicis 04] R. De Amicis & .C Batini. *A methodology for data quality assessment on financial data*. In Studies Commun. Sci. SCKM, 2004. 44

[Decker 08] H. Decker & D. Martinenghi. *Classifying Integrity Checking Methods with Regard to Inconsistency Tolerance*. In Proceedings of the 10th International ACM SIGPLAN Conference on Principles and Practice of Declarative Programming, PPDP '08, pages 195–204, 2008. 28

[Decker 09] H Decker & D. Martinenghi. *Modeling, measuring and monitoring the quality of information*. In C.A. Heuser & G. Pernul, editeurs, Advances in Conceptual Modeling - Challenging Perspectives, volume LNCS 5833, pages 212–221. Springer-Verlag Berlin Heidelberg, 2009. 25, 26, 28

[Deming 82] W. E. Deming. Quality, productivity and competitive position. Massachusetts Institute of Technology, Center for Advanced Engineering Study, Cambridge, Massachusetts, USA, 1982. 22, 43

[Desai 90] B. Desai. An introduction to database systems. West Publishing Company, St. Paul, MN, USA, 1990. 29

[Dey 01] A. Dey. *Understanding and Using Context.* Personal Ubiquitous Comput., vol. 5, no. 1, pages 4–7, January 2001. 8

[Di Ruocco 12] N. Di Ruocco, Scheiwiller, J-M. & A. Sotnykova. *La qualité des données : concepts de base et technique d'amélioration.* In La qualité et la gouvernance des données au service de la performance des entreprises, Eds. Berti-Equille L., Hermes Lavoisier, Paris, 2012. 123, 129

[Dong 09] X.L. Dong & F. Naumann. *Data Fusion : Resolving Data Conflicts for Integration.* Proc. VLDB Endowment, vol. 2, no. 2, pages 1654–1655, 2009. ix

[Dragos 13] V. Dragos. *An ontological analysis of uncertainty in soft data.* In 16th International Conference on Information Fusion, pages 1566–1573, July 2013. 65

[Dubois 85] D. Dubois & H. Prade. *A review of fuzzy set aggregation connectives.* Information Sciences, vol. 36, no. 1–2, pages 85 – 121, 1985.

[Dubois 88] D. Dubois & H. Prade. *Representation and combination of uncertainty with belief functions and possibility measures.* Computational Intelligence, vol. 4, pages 244–264, 1988. 171, 172

[Economist 10] The Economist. *Data, Data Everywhere.* Special report : Managing information, 25 février 2010. `http://www.economist.com/node/15557443`, consulté le 31 mai 2014. ix

[Edwards 92] A. W. F. Edwards. Likelihood (expended edition). Johns Hopkins University Press, 1992. 150

[Ehikioya 99] S.A. Ehikioya. *A characterization of information quality using fuzzy logic.* In 18th International Conference of the North American Fuzzy Information Processing Society (NAFIPS), pages 635–639, 1999. 77

[English 99] L. English. Improving data warehouse and business information quality : Methods for reducing costs and increasing profits. John Wiley & Sons, New York, 1999. 44

[English 09] L. English. Information quality applied : Best practices for improving business information, processes and systems. John Wiley & Sons, Indianapolis, 2009. x, 40

[Ethiraj 04] S. Ethiraj & D. Levinthal. *Modularity and Innovation in Complex Systems.* Management Science, vol. 50, no. 2, page 159 ?173, 2004. 7

[Feigenbaum 91] A. Feigenbaum. Total quality control (3ème édition). McGraw-Hill, New York, NY, USA, 1991. 22

[Finetti 74] B. De Finetti. Theory of probability, volume 1. John Wiley, New York, 1974. 149, 151

[Fisher 01] C.W. Fisher & B.R. Kingma. *Criticality of Data Quality as Exemplified in Two Disasters.* Information & Management, vol. 39, pages 109–116, 2001. 39

[Fisher 13] C.W. Fisher, E. Lauria, S. Chengalur-Smith & R. Wang. Introduction to information quality. An MITIQ Publication, AuthorHouse, 2013. 49

[Floridi 09] L. Floridi. *Philosophical Conceptions of Information.* In Formal Theories of Information, pages 13–53. Springer Berlin Heidelberg, 2009. 14

[Gader 04] P. Gader, A. Mendez-Vasquez, K. Chamberlin, J. Bolton & A. Zare. *A graphical interpretation of the Choquet integral.* IEEE International Geoscience and Remote Sensing Symposium, IGARSS '04, vol. 3, pages 1605 – 1608, 2004. 102

RÉFÉRENCES BIBLIOGRAPHIQUES

[Gardyn 97] E. Gardyn. *A Data Quality Handbook For A Data Warehouse.* In Proceedings of the Conference on Information Quality (ICIQ), Cambridge, MA, page 267–290, 1997. 66

[Gates 99] B. Gates. Business @ the speed of thought : Succeeding in the digital economy. Grand Central Publishing, 1999. ix

[Godfrey 13] P. Godfrey. *Architecting Complex Systems in New Domains and Problems : Making Sense of Complexity and Managing Its Unintended Consequences.* In M. Aiguier, Y. Caseau, D. Krob & A. Rauzy, editeurs, Complex Systems Design & Management, pages 41–51. Springer-Verlag Berlin Heidelberg, 2013. 7

[Grabisch 95] M. Grabisch. *Fuzzy integral in multicriteria decision making.* Fuzzy Sets and Systems, vol. 69, pages 279 – 298, 1995. 99, 100, 102

[Grabisch 97] M. Grabisch. *k-order additive discrete fuzzy measures and their representation.* Fuzzy Sets and Systems, vol. 92, pages 167 – 189, 1997. 102

[Grabisch 00] M. Grabisch. *A graphical interpretation of the Choquet integral.* IEEE Transactions on Fuzzy Systems, vol. 8, no. 5, 2000. 101, 102, 103

[Grant 06] J. Grant & A. Hunter. *Measuring inconsistency in knowledgebases.* Journal of Intelligent Information Systems, vol. 27, no. 2, pages 159–184, 2006. 28

[Gustavsson 05] P. M Gustavsson & T. Planstedt. *The road towards multi-hypothesis intention simulation agents architecture - fractal information fusion modeling.* In IEEE Proceedings of the Winter Simulation Conference, December 2005. 11

[Hall 92] L. Hall & A. Kandel. *The evolution from expert systems to fuzzy expert systems.* In A. Kandel, editeur, Fuzzy Expert Systems, pages 3–21. CRC Press, Boca Raton, FL, USA, 1992. 6

[Higashi 83] M. Higashi & G. Klir. *Measures of uncertainty and information based on possibility distributions.* International Journal of General Systems, vol. 9, no. 1, pages 43–58, 1983. 158

[Hunt 92] D. Hunt. Quality in america : How to implement a competitive quality program. Business One Irwin, Homewood, Ill., USA, 1992. vii, 22, 23

[INCOSE 04] INCOSE. System engineering handbook. International Council on Systems Engineering, 2004. 4

[James 13] J.T. James. *A New, Evidence-Based Estimate of Patient Harms Associated with Hospital Care.* Journal of Patient Safety, vol. 6, no. 3, pages 122–128, 2013. 39

[Juran 89] J. M. Juran. Juran on leadership for quality. Free Press, New York, 1989. vii, 22

[Kanaracus 14] C. Kanaracus. *IT spending growth to be slower than expected in 2014 due to pricing pressure, Gartner says.* PCWorld, 30 juin 2014. x

[Kim 02] E. Kim, W. Kim & Y. Lee. *Classifier fusion using local confidence.* LNAI 2366, 2002. 59

[Kim 13] J.K. Kim & J. Shao. Statistical methods for handling incomplete data. Chapman and Hall/CRC, 2013. 31

[King 78] W. King & J. Rodriguez. *Evaluating Management Information Systems.* MIS Quarterly, vol. 2, no. 3, pages 43–51, 1978. 43

[Klein 97] B. Klein, D. Goodhue & G. Davis. *Can Humans Detect Errors in Data ? Impact of Base Rates, Incentives, and Goalss.* MIS Quarterly, vol. 21, no. 2, page 169–194, 1997. 40

[Klein 14] L. Klein. *Multisensor data fusion for object detection, classification and identification (Tutorial).* In SPIE Defense, Security+Sensing, 6 Mai 2014. vii, 113, 114

[Klir 88] G. Klir & T.A. Folger. Fuzzy sets, uncertainty and information. Prentice-Hall International, Englewood Cliffs, New-Jersey, 1988. v, 36, 66, 154, 156, 164

[Klir 01] G. Klir & R.M. Smith. *On measuring uncertainty and uncertainty-based information : Recent development.* Annals of Mathematics and Artificial Intelligence, pages 5–33, 2001. 151, 153

[Klir 06a] G. Klir. Uncertainty and information : Foundations of generalized information theory. John Wiley & Sons, Hoboken, New Jersey, 2006. 65, 162, 163, 164

[Klir 06b] G. Klir & B.M. Ayyub. Uncertainty modeling and analysis in engineering and the science. Chapman & Hall/CRC, Boca Raton, 2006. 157, 161, 162

[Knight 08] S-A. Knight. *User Perceptions of Information Quality in World Wide Web Information Retrieval Behaviour.* PhD thesis, Edith Cowan University, Australia, 2008. 44

[Kovac 97] R. Kovac, Y.W. Lee & L. Pipino. *Total Data Quality Management : the case of IRI.* In Proceedings of the International Conference on Information Quality (ICIQ), Cambridge, MA, page 63–79, 1997. 66

[Kurtz 03] C. Kurtz & D. Snowden. *The New Dynamics of Strategy : sense making in a complex and complicated world.* IBM Systems Journal, vol. 42, no. 3, page 462–483, 2003. 7

[Kwan 96] S. Kwan, F. Olken & D. Rotem. *Uncertain, incomplete and inconsistent data in scientific and statistical databases.* In A. Motro & P. Smets, editeurs, Uncertainty Management in Information Systems, pages 127–153. Kluwer Academic Publishers, 1996. 6

[Laudon 11] K. Laudon & J. Laudon. Management information systems : Managing the digital firm. Pearson Custom Publishing ; Global ed of 12th revised edition, 2011. 41

[Le Bras 11] Y. Le Bras. *Contribution à l'étude des meures de l'intérêt des règles d'association et à leurs propriétés algorithmiques.* PhD thesis, Télécom Bretagne, 2011. 142

[Lecornu 09a] L. Lecornu, C. Le Guillou, P.J. Garreau, P. Saliou, H. Jantzem, J. Puentes & J-M. Cauvin. *REFEROCOD : A probabilistic method to medical coding support.* In 2009 Annual International Conference of the IEEE Engineering in Medicine and Biology (EMBC), 2009. 129, 130, 132

[Lecornu 09b] L. Lecornu, C. Le Guillou, G. Thillay, P.J. Garreau, H. Jantzem & J-M. Cauvin. *C2i : A tool to gather medical indexed information.* In 9th International Conference on Information Technology and Applications in Biomedicine (ITAB), 2009. 132

[Lecornu 10] L. Lecornu, C. Le Guillou, F. Le Saux, M. Hubert, J. Puentes & J-M. Cauvin. *ANTEROCOD : Actuarial survival curves applied to medical coding support for chronic diseases.* In 2010 Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), pages 1158–1161, 2010. 132

RÉFÉRENCES BIBLIOGRAPHIQUES

[Lee 02] Y.W. Lee, D. Diane M. Strong, R. Kahn & R. Wang. *AIMQ : a methodology for information quality assessment*. Information & Management, vol. 40, pages 133–146, 2002. 44, 65

[Lefebvre 07] E. Lefebvre, M. Hadzagic & É. Bossé. Advances and challenges in multisensor data and information processing, chapitre On quality of information in multi-source fusion environments. IOS Press, 2007. v, 58, 59, 60

[Lions 96] J.L. Lions. *ARIANE 5 Flight 501 Failurer*. Rapport par l'Inquiry Board, Paris, 19 juillet 1996. `http://sspg1.bnsc.rl.ac.uk/Share/ISTP/ariane5r.htm` consulté le 31 mai 2014. 39

[Llinas 08] J. Llinas. Handbook of multisensor data fusion : theory and practice (2nd edition), chapitre 25 : Assessing the performance of multisensor fusion processes, pages 655–675. CRC Press, 2008. 64

[Loshin 04] D. Loshin. Enterprise knowledge management - the data quality approach. Morgan Kaufmann, 2004. 44

[Ma 06] Z. Ma. Fuzzy database modeling of imprecise and uncertain engineering information. Springer-Verlag Berlin Heidelberg, 2006. 37

[Mandke 97] V.V. Mandke & M.K. Nayar. *Information integrity–a structure for its definition*. In Proceedings of the International Conference on Information Quality (ICIQ), Cambridge, MA, page 314–338, 1997. 66, 67

[Mason 78] R. Mason. *Measuring information output : A communication systems approach*. Information & Management, vol. 1, no. 1, pages 219–234, 1978. ix

[Matlin 77] G. Matlin. *How to Survive a Management Assessment*. MIS Quarterly, vol. 1, no. 1, pages 11–17, 1977. 43

[Mayer-Schonberger 13] V. Mayer-Schonberger & K. Cukier. Big data : A revolution that will transform how we live, work and think. Houghton Mifflin Harcourt, Boston, MA, 2013. ix

[Meyen 97] D.M. Meyen & M.J. Willshire. *A data quality engineering framework*. In Proceedings of the Conference on Information Quality (ICIQ), Cambridge, MA, page 95–116, 1997. 66, 67

[Mingers 96] J. C. Mingers. *An evaluation of theories of information with regard to the semantic and pragmatic aspects of information systems*. Systems Practice, vol. 9, no. 3, pages 187–209, 1996. 12

[Motro 95] A. Motro. *Imprecision and Uncertainty in Database Systems*. In P. Bosc & J. Kacprzyk, editeurs, Fuzziness in Database Management Systems, pages 3–22. Springer-Verlag Berlin Heidelberg, 1995. 25, 31, 37

[Motro 96] A. Motro & Smets P. Uncertainty management in information systems : From needs to solutions. Kluwer Academic Publishers, 1996. 30, 60

[Murphy 00] C.K. Murphy. *Combining belief functions when evidence conflicts*. Decision Support Systems, vol. 29, pages 1–9, 2000. 172

[Mutsuzaki 07] M. Mutsuzaki, M. Theobald, A. de Keijzer, J. Widom, P. Agrawal, O. Benjelloun, A.D. Sarma, R. Murthy & T. Sugihara. *Trio-One : Layering uncertainty and lineage on a conventional DBMS*. In Proceedings of CIDR, page 269–274, 2007. 33

[Naumann 01] F. Naumann. *From Databases to Information Systems - Information Quality Makes the Difference*. In Proceedings of the International Conference on Information Quality (ICIQ), pages 244–260. MIT, 2001. 50

[Naumann 02] F. Naumann. Quality-driven query answering, chapitre Information quality criteria, pages 29–50. LNCS 2261, Springer-Verlag, 2002. vii, 50, 51

[NetCraft 14] NetCraft & Internet Live Stats, 2014. `http://www.internetlivestats.com/total-number-of-websites`, consulté le 28 août 2014. v, 51

[O'Brien 11] J. O'Brien & G. Marakas. Management information systems, 10ème édition. McGraw-Hill/Irwin, New York, NY, USA, 2011. 6

[OMG 08] Object Management Group OMG. *UML Profile for Modeling Quality of Service and Fault Tolerance Characteristics and Mechanisms Specification, Version 1.1*, 2008. `http://www.omg.org/spec/QFTP/1.1/`, consulté le 28 août 2014. 50

[Papin 08] P. Papin. *Ce qu'il faut savoir sur les dispositifs médicaux implantables [DMI].* Revue de Chirurgie Orthopédique et Réparatrice de l'Appareil Moteur, vol. 94, no. 6, Supplement, pages 91–95, 2008. 121

[Parsons 01] S. Parsons. Qualitative methods for reasoning under uncertainty. MIT Press, Cambridge, Massachussetts, 2001. 57

[Pautke 02] R. Pautke & T. Redman. Information and database quality, chapitre The organisation's most important data issues, pages 1–12. Kluwer Academic Publishers, 2002. x

[Pearson 05] R.. Pearson. Mining imperfect data : Dealing with contamination and incomplete records. The Society for Industrial and Applied Mathematics (SIAM), 2005. 30, 31

[Peralta 06] V. Peralta. *Data Quality Evaluation in Data Integration Systems.* PhD thesis, Université de Versaille Saint-Quentin-en-Yvelines et Universidad de la Republica (Uruguay), 2006. 48

[Perry 04] W. Perry, D. Signori & J. Boon. *Exploring information superiority : A methodology for measuring the quality of information and its impact on shared awareness.* Rapport technique MR-1467, RAND Corporation, 2004. 60

[Pipino 02] L. Pipino, Y.W. Lee & R. Wang. *Data quality assessment.* Communications of the ACM, vol. 45, no. 4, page 211–218, 2002. 44

[Puentes 13] J. Puentes, J. Montagner, L. Lecornu & J-M. Cauvin. *Information quality measurement of medical encoding support based on usability.* Computer Methods and Programs in Biomedicine, vol. 112, no. 3, pages 329–342, 2013. 134

[Ramakrishnan 03] R. Ramakrishnan & J. Gehrke. Database management systems,. McGraw-Hill, New York, 2003. 25, 26, 28

[Ran 03] S. Ran. *A Model for Web Services Discovery with QoS.* SIGecom Exch., vol. 4, no. 1, pages 1–10, March 2003. 50

[Redman 92] T. Redman. Data quality : Management and technology. Bantam Books, 1992. 66, 67, 68

[Redman 96] T. Redman. Data quality for the information age. Artech House, Boston, Massachusetts, 1996. 40

[Redman 98] T. Redman. *The Impact of Poor Data Quality on the Typical Enterprise.* Communications of the ACM, vol. 41, no. 2, pages 79–82, 1998. 40

[Redman 13] T. Redman. *Data Quality Management Past, Present, and Future : Towards a Management System for Data.* In Handbook of Data Quality : Research and Practice, eds. Sadiq, S., pages 15–40. Springer-Verlag Berlin Heidelberg, 2013. 140

RÉFÉRENCES BIBLIOGRAPHIQUES

[Rempt 01] R. Rempt. *The Navy in the 21st Century, Part II : Theater Air and Missile Defense*. Johns Hopkins APL Technical Digest, vol. 22, no. 1, 2001. v, 63

[Rogova 04] G. Rogova & V. Nimier. *Reliability in information fusion : Literature survey*. 7th Conference on Information Fusion (FUSION), 2004. 58

[Rogova 10] G. Rogova & É. Bossé. *Information quality in information fusion*. 13th Conference on Information Fusion (FUSION), pages 1–8, 2010. v, 59, 60, 61, 62

[San Segundo 02] R. San Segundo. *A new concept of knowledge*. Online Information Review, vol. 26, no. 4, pages 239–245, 2002. 15

[Scannapieco 04] M. Scannapieco, A. Virgillito, C. Marchetti, M. Mecella & R. Baldoni. *The DaQuinCIS Architecture : A Platform for Exchanging and Improving Data Quality in Cooperative Information Systems*. Information Systems, vol. 29, no. 7, pages 551–582, September 2004. 44

[Schilit 94] B. Schilit, N. Adams & R. Want. *Context-Aware computing applications*. In Proceedings of the 1994 First Workshop on Mobile Computing Systems and Applications, WMCSA '94, pages 85–90, Washington, DC, USA, 1994. IEEE Computer Society. 8

[Schuck 10] T. Schuck & E. Blasch. *Description of the Choquet integral for tactical knowledge representation*. 13th Conference on Information Fusion (FUSION), 2010. 103

[Sebastian-Coleman 13] L. Sebastian-Coleman. Measuring data quality for ongoing improvement : a data quality assessment framework. Morgan Kaufmann, 2013. 11

[Shafer 76] G. Shafer. A mathematical theory of evidence. Princeton University Press, 1976. 151

[Shankaranarayanan 06] G. Shankaranarayanan, A. Even & S. Watts. *The role of process metadata and data quality perceptions in decision-making*. Journal of Information Technology Management, vol. XVII, no. 1, pages 50–67, 2006. 41

[Shannon 93] C. Shannon. The collected papers of claude e. shannon. In Sloane, N.J.A. and Wyner, A.D., editeurs, IEEE Press, New York, NY, USA, 1993. 13

[Sheridan 05] T. Sheridan & R. Parasuraman. *Human-automation interaction*. Reviews of human factors and ergonomics, vol. 1, no. 1, pages 89–129, 2005. 41

[Simon 62] H. Simon. *The architecture of complexity*. Proceedings of the American Philosophical Society, vol. 106, pages 467–482, 1962. 7

[Singh 13] Rajeev Pratap Singh & K.K. Pattanaik. *An Approach to Composite QoS Parameter based Web Service Selection*. Procedia Computer Science, vol. 19, pages 470–477, 2013. 53

[Singpurwalla 02] N. Singpurwalla, J. Booker & T. Bement. Fuzzy logic and probability applications : Bridging the gap, chapitre Probability theory, pages 55–71. ASA-SIAM, 2002. 149, 150

[Singpurwalla 04] D. Singpurwalla & J. Booker. *Membership functions and probability measures of fuzzy sets (with comments)*. Journal of the american statistical association, vol. 99, no. 467, pages 867–889, 2004. 150, 152

[Smarandache 04] F. Smarandache & J. Dezert. Advances and applications of dsmt for information fusion, (vol. 1). American Research Press, Rehoboth, 2004. 171, 172, 174

[Smets 94a] P. Smets. *What is Dempster-Shafer model*. In R. Yager, M. Fedrizzi & J. Kacprzyk, editeurs, Advances in the Dempster-Shafer theory of evidence, pages 225–254. John Wiley & Sons, 1994. 152

[Smets 94b] P. Smets & R. Kennes. *The transferable belief model.* Artificial Intelligence, vol. 66, no. 2, pages 193–234, 1994. 172

[Smets 96] P. Smets. *Imperfect Information : Imprecision and Uncertainty.* In A. Motro & P. Smets, editeurs, Uncertainty management in information systems : From needs to solutions, pages 225–254. Kluwer Academic Publishers, 1996. 56

[Smithson 89] M. Smithson. Ignorance and uncertainty : Emerging paradigms. Springer Verlag, New York, 1989. 55

[Solano 12] M.A. Solano & G. Jernigan. *Enterprise data architecture principles for High-Level Multi-Int fusion : A pragmatic guide for implementing a heterogeneous data exploitation framework.* In Proceedings of the 15th International Conference on Information Fusion (FUSION'12), pages 867–874, July 2012. v, 21

[Sølvberg 93] A. Sølvberg & D.C. Kung. Information systems engineering : An introduction. Springer-Verlag, 1993. 10, 14

[Srivastava 83] R. Srivastava. *Reliability modeling of information systems with human elements : A new perspective*, 1983. 88

[Srivastava 85] R. Srivastava. *A note on Internal Control Systems with Control Components in Series.* The Accounting Review, vol. LX, no. 3, pages 504–507, 1985. 88

[Stefanidis 11] K. Stefanidis, E. Pitoura & P. Vassiliadis. *Managing contextual preferences.* Information Systems, vol. 36, pages 1158–1180, 2011. 9

[Strong 97] D. Strong, Y. Lee & R. Wang. *Data quality in context.* Commununications of the ACM, vol. 40, no. 5, pages 103–110, May 1997. 40

[Stvilia 08] B. Stvilia & L. Gasser. *An activity theoretical model for information quality change.* First Monday, vol. 13, no. 4, 2008. 68

[Tait 05] P. Tait. Introduction to target recognition. The Institution of Electrical Engineers, Stevenage, Herts, U.K., 2005. 109, 111, 112

[Todoran 11] I.G. Todoran. *Fusion de données hétérogènes.* Rapport technique, Télécom Bretagne, mars 2011. 133

[Todoran 13] I.G. Todoran, L. Lecornu, A. Khenchaf & J.M. Le Caillec. *Information quality evaluation in fusion systems.* In Proceedings of the 16th International Conference on Information Fusion (FUSION'13), pages 906–913, July 2013. 75, 87, 96

[Todoran 14a] I.G. Todoran, L. Lecornu, A. Khenchaf & J.M. Le Caillec. *Assessing information quality in information fusion systems.* In NATO Symposium on Analysis Support to Decision Making in Cyber Defence & Security, 2014. 82, 98, 104

[Todoran 14b] I.G. Todoran, L. Lecornu, A. Khenchaf & J.M. Le Caillec. *A methodology to evaluate important dimensions of information quality in systems.* Journal of Data and Information Quality (JDIQ) - (sous révision), 2014. 115

[Todoran 14c] I.G. Todoran, L. Lecornu, A. Khenchaf & J.M. Le Caillec. *Toward the quality evaluation in complex information systems.* In I. Kadar, editeur, Signal Processing, Sensor/Information Fusion and Target Recognition XXIII. Proceedings of SPIE Defense, Security, and Sensing Symposium, Vol. 9091, 90910N, 2014. 96, 98, 104

RÉFÉRENCES BIBLIOGRAPHIQUES

[Toumi 07] A. Toumi. *Intégration des bases de connaissances dans les systèmes d'aide à la décision : Application à l'aide à la reconnaissance de cibles radar non-coopératives*. PhD thesis, Université de Bretagne Occidentale, 2007. vi, 112, 113

[Turban 05] E. Turban, J. Aronson & T.P. Liang. Decision support systems and intelligent systems. Pearson Education, 7 edition, 2005. 3, 4, 6

[Van Trees 68] H. Van Trees. Detection, estimation and modulation theory, part i : Detection, estimation and linear modulation theory. John Wiley & Sons, 1968. 90

[Verykios 02] V. Verykios & A. Elmagarmid. *The Purdue University Data Quality Project*. In Data Quality, eds. Wang, R. and Ziad, M. and Lee, Y., pages 119–137. Kluwer Academic Publishers, 2002. 143

[Waltz 90] E. Waltz & J. Llinas. Multisensor data fusion, chapitre 11 : System modeling and performance evaluation, pages 389–423. Artech House, 1990. 54, 62, 64, 112

[Wand 96] Y. Wand & R. Wang. *Anchoring Data Quality Dimensions in Ontological Foundations*. Commununications of the ACM, vol. 39, no. 11, page 86–95, 1996. 40

[Wang 95] R. Wang, M.P. Reddy & H. Kon. *Toward quality data : An attribute based approach*. Decision Support Systems, vol. 13, pages 349–372, 1995. 33, 34

[Wang 96] R. Wang & D. Strong. *Beyond accuracy : what data quality means to data consumers*. Journal of Management Information Systems, vol. 12, no. 4, pages 5–33, March 1996. vii, 40, 44, 46, 66, 140

[Wang 02] R. Wang, M. Ziad & Y. Lee. Data quailty. Kluwer Academic Publishers, 2002. 33, 34

[Warren 99] L. Warren. *Strategic information synthesis by globular knowledge fusion*. Proceedings of Information, Decision and Control, pages 407 – 412, 1999. 103

[Weaver 49] W. Weaver. *The mathematics of communication*. Scientific American, vol. 181, no. 1, pages 11–15, 1949. 13

[Weise 06] E. Weise. *Study : Medication errors harm 1.5M a year*. USA Today, 21 juillet 2006. `http://usatoday30.usatoday.com/money/industries/health/2006-07-20-drug-errors_x.htm`, consulté le 31 mai 2014. 39

[Wilkinson 98] A. Wilkinson, T. Redman, E. Snape & M. Marchington. Managing with total quality management. MacMillan Business, 1998. 67

[Xiao 09] J. Xiao. *Gestion des incertitudes dans le processus de développement de systèmes complexes*. PhD thesis, Institut National Polytechnique de Toulouse, 2009. 4

[Xu 92] L. Xu, A. Krzyzak & C.Y. Suen. *Methods of Combining Multiple Classifiers and Their Application to Handwriting Recognition*. IEEE Transactions on Systems,Man and Cybernetics, vol. 22, no. 3, 1992. 166

[Yager 87a] R. Yager. *Toward a theory of conjunctive variables*. International Journal of General Systems, vol. 13, pages 203–227, 1987. 153

[Yager 87b] R. Yager & R. Kennes. *On the Dempster-Shafer framework and new combination rules*. Information Sciences, vol. 41, pages 93–138, 1987. 172

[Yager 91] R. Yager. *Connectives and quantifiers in fuzzy sets*. Fuzzy Sets and Systems, vol. 40, pages 39 – 75, 1991. 100

[Yager 05] R. Yager. *Uncertainty Management for Intelligence Analysis.* In Data Fusion for Situation Monitoring, Incident Detection, Alert and Response Management, pages 35–60. IOS Press, 2005. 156, 159

[Ye 08] J. Ye, S. McKeever, L. Coyle, S. Neely & S. Dobson. *Resolving uncertainty in context integration and abstraction : context integration and abstraction.* In Proceedings of the 5th international conference on Pervasive services, ICPS '08, pages 131–140, New York, NY, USA, 2008. ACM. 9

[Zeleny 87] M. Zeleny. *Management Support Systems : Towards integrated knowledge management.* Human Systems Management, vol. 7, no. 1, pages 59–70, 1987. 11

[Zeng 04] L. Zeng, B. Benatallah, A. Ngu, M. Dumas, J. Kalagnanam & H. Chang. *QoS-aware middleware for Web services composition.* IEEE Transactions on Software Engineering, vol. 30, no. 5, pages 311–327, May 2004. 50

[Zhu 03] Y. Zhu. Multisensor decision and estimation fusion. Kluwer Academic Publishers, Boston, 2003. 167

[Zimmermann 80] H.-J. Zimmermann & P. Zysno. *Latent connectives in human decision making.* Fuzzy Sets and Systems, vol. 4, pages 37 – 51, 1980. 100

[Zins 07] C. Zins. *Conceptual approaches for defining data, information, and knowledge.* Journal of the American Society for Information Science and Technology, vol. 54, no. 4, pages 479–493, 2007. 10

Technopôle Brest-Iroise - CS 83818
29238 Brest Cedex 3
France
Tél : + 33 (0)2 29 00 11 11
www.telecom-bretagne.eu

Résumé

Le développement des réseaux sociaux et le déploiement d'un nombre de plus en plus élevé de capteurs font que la quantité des données disponibles à traiter ne cesse d'augmenter. Mais avoir à sa disposition une quantité impressionnante de données n'est pas suffisant, car un autre facteur joue un rôle aussi important : la qualité des données et des informations extraites. Il existe dans la littérature plus de 20 méthodologies d'évaluation de la qualité de l'information (QI). Malheureusement, toutes ces méthodologies ont une vision boîte noire du système d'information (SI) et essaient d'évaluer la QI en utilisant des formulaires. Cette procédure n'est pas intuitive, est très subjective et demande beaucoup de temps. Nous proposons une nouvelle méthodologie, en 3 étapes, d'évaluation de la QI d'un SI complexe. Dans la première étape le SI est décomposé en modules élémentaires. Grâce à cela, il est possible d'étudier la qualité en entrée et en sortie de chaque module : qualité locale. Dans la deuxième étape chaque module est caractérisé par une fonction de transfert de qualité réalisant le lien entre la qualité en entrée et celle en sortie. Dans la troisième étape, la qualité du SI global est évaluée en utilisant la qualité locale et les fonctions de transfert de qualité des modules. Celle-ci est appelée qualité globale et est évaluée par la propagation de la qualité à travers le SI. La validation est réalisée en considérant 2 applications : un système de reconnaissance automatique de cibles radar et un système d'aide au codage médical. Cette méthodologie permet non seulement une évaluation instantanée de la qualité mais aussi une explication de la qualité à plusieurs niveaux.

Mots-clés : Qualité des données, Qualité de l'information, Mesures de qualité, Fonction de transfert de qualité, Système d'information complexe, Modélisation de l'incertitude, Système radar, Codage médical

Abstract

The recent development of social networks and the deployment of different types of sensors made possible to record and store in databases large amounts of heterogeneous data. The quantity of available data is not the only problem to be handled and a new concept has arisen: the quality of data and its influence over information system performance. Assessing the quality of the information proposed by an information system has become one of the major research topics in the last two decades. A quick literature survey shows that a significant number of information quality frameworks are proposed in different domains of application. Unfortunately, they do not provide a feasible methodology that is both simple and intuitive to be implemented in practice. In order to address this need, we propose a new information quality methodology. Our methodology makes use of existing frameworks and proposes a three-step process capable of tracking the quality changes through the system. In the first step we propose decomposing the information system into its elementary modules. Having access to each module allows us to locally define the information quality. Then, in the second step we model each processing module by a quality transfer function, capturing the module's influence over the information quality. In the third step we make use of the previous two steps in order to estimate the quality of the entire information system. Thus, our methodology allows informing the end user on both output quality and local quality. The proof of concept of our methodology has been carried out considering two applications: an automatic target recognition system and a diagnosis coding support system.

Keywords : Data quality, Information quality, Quality measures, Quality transfer function, Complex information system, Uncertainty theory, Radar automatic target recognition system, Medical encoding

TELECOM Bretagne

n° d'ordre : 2014telb0343

Télécom Bretagne

Technopôle Brest-Iroise - CS 83818 - 29238 Brest Cedex 3

Tél : + 33(0) 29 00 11 11 - Fax : + 33(0) 29 00 10 00