La qualité des données comme condition à la qualité des connaissances : un état de l'art

Laure Berti-Équille

IRISA, Campus Universitaire de Beaulieu
35042 Rennes, France
Laure.Berti-Equille@irisa.fr
http://www.irisa.fr

Résumé. Les travaux actuels sur l'extraction de connaissances à partir des données (ECD) se focalisent sur la recherche de règles intéressantes dont on souhaite pouvoir qualifier l'intérêt ou le caractère exceptionnel, mais dont la validité dépend bien évidemment de celle des données. En amont du processus d'ECD, il semble donc essentiel d'évaluer la qualité des données stockées dans les bases et entrepôts de données afin de : (1) proposer aux utilisateurs une expertise critique de la qualité du contenu d'un système, (2) orienter l'extraction des connaissances en fonction d'un profil ciblé d'utilisateurs et de décideurs, (3) permettre à ceux-ci de relativiser la confiance qu'ils pourraient accorder aux données et aux règles extraites, et leur permettre ainsi de mieux en adapter leur usage, (4) assurer enfin la validité et l'intérêt des connaissances extraites à partir des données. Cet article fait une synthèse de l'état de l'art dans le domaine de la qualité des données en présentant, dans un premier temps, les causes de la non-qualité des données, puis en décrivant un panorama des travaux sur la qualité des données, travaux pertinents dès lors que l'on s'intéresse à modéliser, mesurer et à améliorer la qualité des connaissances "élaborées" à partir des données. Enfin, l'article propose d'exploiter les méta-données décrivant la qualité des données dans le processus d'ECD.
Mots-clés. Qualité des données, méta-données

1. Introduction

Avec la multiplication des sources d'informations disponibles et l'accroissement des volumes de données potentiellement accessibles, l'extraction de connaissances à partir des données a pris une place de premier plan tant au niveau académique qu'au sein des entreprises. En effet, la mise en évidence de liens cachés ou de phénomènes de causalités non-triviales à partir de grandes quantités de données permettra d'aider les décideurs dans leurs choix. Cependant, l'identification, à partir d'une grande collection de données, de motifs valides, nouveaux, potentiellement utiles et compréhensibles dépend de manière très critique de la qualité des données (généralement intégrées) qu'utilisent les algorithmes de fouille de données [CM95] [HG+95] [CDL+97] [Vas00].

Si l'analyse des données peut être réalisée sur des données inexactes, incomplètes, ambiguës et de qualité médiocre, on peut s'interroger sur le sens à donner aux résultats de ces analyses et remettre en cause, à juste titre, la qualité des connaissances ainsi "élaborées".

Les mesures statistiques classiques permettent de détecter des erreurs, des exceptions ou des incohérences d'un jeu de données, mais elles permettent difficilement de discerner et de qualifier les données de qualité douteuse : l'expertise humaine est ici requise pour un jugement qualitatif des données, et c'est également un préalable nécessaire à toute prise de décision. En dépit de son coût, une expertise de la qualité des données semble incontournable dans le contexte de la validation des règles issues d'un processus d'ECD.

L'objet de notre article est de présenter une synthèse de l'état de l'art sur la qualité des données dans la mesure où son impact conditionne la qualité des connaissances extraites. Le cas particulier des données textuelles ne sera pas abordé dans cet état de l'art dont l'objectif est de se situer entre l'approche qualité et l'ECD. Il nous semble donc essentiel d'évaluer préalablement la qualité des données stockées dans les systèmes afin de :

- proposer aux utilisateurs une expertise critique de la qualité du contenu d'un système de stockage des données (base, entrepôt de données, ou système d'information),
- permettre à ceux-ci de relativiser la confiance qu'ils pourraient accorder aux données et aux règles extraites, et leur permettre ainsi, de mieux en adapter leur usage,
- assurer finalement la validité et l'intérêt des connaissances extraites à partir des données.

Les principales motivations de l'article sont de souligner que : 1) la démarche qualité doit s'intégrer dans le processus d'ECD, 2) les techniques d'ECD doivent être employées pour mesurer et consolider la qualité des données, 3) la qualité des données est une étape préalable à l'ECD, fournissant des données additionnelles (méta-données) que doivent prendre en compte les opérateurs d'ECD.

L'article s'organise de la façon suivante : la section 2 répertorie les causes de la non-qualité des données. La section 3 présente un panorama des travaux menés sur la qualité des données. La section 4 propose d'exploiter les méta-données décrivant la qualité des données pour renseigner la qualité des connaissances extraites à partir des données. La section 5 conclut l'article et présente les perspectives de recherche dans le domaine.

2. Les causes et conséquences de la non-qualité des données

La qualité des données suscite, depuis moins d'une dizaine d'années, un vif intérêt, et ce thème émerge en tant que champ de recherche à part entière, comme peuvent l'indiquer l'organisation des premières conférences internationales *Information Quality* (*ICIQ* http://web.mit.edu/tdqm/www/iqc) au *Massachusetts Institute of Technology* [Wan96] [SK97] [CP98], de workshops européens *Data Quality in Cooperative Information Systems* (*DQCIS* http://www.dis.uniroma1.it/~dq/dqcis/), ainsi que les articles et éditions spéciales que lui consacrent les revues *IEEE Transactions on Data and Knowledge Engineering* [WSF95] [BP02] et *Communications of ACM* [TB98].

A l'origine, le constat (souvent cruel) de l'impact de la qualité médiocre des données a tout d'abord été fait en entreprise par des utilisateurs impliqués en pratique et surtout financièrement. La littérature sur le thème de la qualité des données déplore la situation suivante : de nombreuses bases de données ont de graves problèmes de qualité de leurs données. Plusieurs cas décrits dans [Jac92] [Bas95] [Red96], révèlent de nombreux exemples de situations alarmantes concernant la qualité des données dans les bases de données commerciales, médicales, du domaine publique ou de l'industrie [Lau86][CM95][Wan96]. Différents problèmes concernant la qualité des données géographiques sont considérés dans [GJ98]. Le décalage qui existe entre la qualité des données dans la pratique et la qualité des

données en théorie impose une distinction entre les travaux académiques [WSF95] [Nau02] et ceux menés par les « praticiens » en entreprise [Red96].

Les principales causes des problèmes de qualité de données sont d'origines diverses :

- lors de la modélisation conceptuelle des données, lorsque les définitions des attributs n'ont pas été suffisamment bien structurées ou normalisées, que le schéma n'a pas été validé et/ou qu'il manque des contraintes d'intégrité et des procédures pour maintenir la cohérence des données,
- lorsque les interprétations des données sont incohérentes, notamment à cause des différences nationales dans l'usage de certains codes ou symboles,
- lors du développement logiciel : les besoins peuvent ne pas avoir été complètement spécifiés et des erreurs peuvent avoir été introduites lors du processus de développement, notamment entre les phases d'analyse et de conception,
- lorsque la méthode de saisie des données n'a pas été bien conçue et qu'il manque des procédures de vérification systématique : les erreurs humaines sont facilement introduites,
- lorsque la sécurité physique et/ou logique du système laisse à désirer : les données peuvent être corrompues volontairement,
- lorsque les entreprises n'ont pas les moyens de traquer l'âge de leurs données, ni de les mettre à jour ou de les enrichir, ou bien encore d'éliminer les données devenues obsolètes,
- lors de l'intégration ou de la post-intégration de bases de données hétérogènes : les données peuvent être contradictoires ou incohérentes entre les différents systèmes, et les heuristiques d'intégration peuvent s'avérer inadaptées pour chacun des systèmes dont la qualité des données n'est d'ailleurs pas localement homogène,
- lors de la conversion des systèmes ou de leur migration : les programmes de conversion peuvent introduire de nouvelles erreurs ; la rétro-ingénierie peut ne pas considérer ou perdre le contexte de définition, de production ou d'usage de la donnée,

Ces trois derniers points sont particulièrement perçus dans le processus d'ECD.

Ces raisons motivent d'autant plus la prise en compte de la qualité des données comme partie intégrante lors de l'analyse, de la conception, du développement et de la maintenance du système de gestion des données (base ou entrepôt), ainsi que dans toute la chaîne de traitement de l'information [Red96]. Même si de nouvelles fonctionnalités sont ajoutées aux bases de données pour améliorer le processus de prise de décision, ainsi que la qualité des services (sans être trop grandes consommatrices de ressources, ni ajouter une trop grande complexité aux traitements), le besoin de méthodes, de métriques et d'outils pour définir, mesurer, interroger et contrôler la qualité des données dans les bases de données reste très clairement ressenti [Pat93] [CP98] [TB98].

3. Travaux sur la qualité des données

Si la nécessité du contrôle et de la gestion de la non-qualité des données suscite une réelle prise de conscience au sein des différentes communautés de recherche en Bases de Données, Systèmes d'information et Réseaux, la première difficulté réside dans l'absence de consensus sur la notion même de qualité. Tout le monde s'accorde en effet sur le fait que la qualité d'une donnée peut se décomposer en un certain nombre de dimensions, catégories, critères,

facteurs, paramètres ou attributs [WKM93] [FLR94] [Red96] [WSF95] [MR00], mais aucune définition ne fait aujourd'hui l'unanimité. Chaque domaine d'application possède en fait sa propre vision de la qualité de ses données.

Jusqu'à présent, les approches adoptées pour mesurer la qualité des données dans une base de données sont des approches essentiellement statistiques procédant par échantillonnage sur d'importants volumes de données. Ces approches sont centrées sur des méthodes telles que l'inférence sur les données manquantes fondées sur des modèles statistiques, sur la détection automatique et le traitement des exceptions et des données isolées.

De nombreuses propositions ont été développées pour mesurer la qualité des données fournies aux utilisateurs conformément à des spécifications préalables de qualité, notamment dans le domaine des Systèmes d'Information Géographique [GM95] [GJ98]. Actuellement, les techniques de mesure les plus utilisées sont les techniques d'échantillonnage et les audits [PC93] qui demeurent d'après [WSF95] les seuls moyens pratiques de déterminer la qualité des données dans une base (recensement des différents types d'erreurs, élaboration de méthodes pour les détecter, estimation de leur fréquence d'occurrences dans la base, etc.).

Parallèlement aux approches "statisticiennes" et "fréquentistes" pour la mesure de la qualité des données, quelques approches "subjectivistes", plus orientées vers l'assistance aux utilisateurs ont émergé [Ber99][Ber02]. Pour résumer, les différents travaux sur la qualité des données peuvent être classés selon leur objectif qui est :

- de définir rigoureusement chaque dimension de la qualité des données, ainsi que le mode opératoire (ou protocole) de mesure selon une méthode scientifique particulière,
- de créer un standard universel présentant l'ensemble des dimensions opérationnelles de la qualité des données pour l'application considérée,
- de proposer une assistance à l'utilisateur pour qu'il définisse et évalue lui-même la qualité des données qu'il acquiert et manipule.

Parmi ces trois courants de recherche sur la qualité des données, différents travaux se sont plus particulièrement intéressés à l'une des trois activités suivantes qui seront décrites ci-après :

- la modélisation de la qualité des données,
- la mesure de la qualité des données,
- la correction des erreurs et l'amélioration de la qualité des données.

3.1 Modélisation de la qualité des données

Les plus anciens travaux sur la modélisation de la qualité des données ont été menés par [Bro80] qui propose six concepts pour caractériser la qualité des données gérées par un système d'information (voir TAB. 1). Dans [BP85], les auteurs ont, les premiers, défini la notion de qualité des données ainsi que ses dimensions. La terminologie et les concepts fondamentaux sur ce thème ont également été précisés dans [DR92][FLR94]. De nombreuses propositions ont depuis été faites [Wan96] [SK97] [CP98], mais le consensus sur la définition de la qualité des données n'est toujours pas atteint. Le tableau 1 synthétise quelques-unes de ces propositions.

L'analyse très complète menée dans [WSF95] a recensé les différentes approches concernant la qualité des données sur les plans managérial, organisationnel, financier, légal

Auteurs et références	*Dimensions de la qualité des données*	
Brodie [Bro80]	6 concepts	*Intégrité* *Niveau d'abstraction du modèle conceptuel* *Expressivité sémantique* *Validité par rapport à des données de référence* *Maintenance des données* *Efficacité dans l'utilisation des ressources*
Delen, Rijsenbrij [DR92]	4 dimensions 21 aspects 40 attributs	*Développement et contrôle du S.I.* *Propriétés statiques de maintenance* *Fonctionnement dynamique* *Importance de l'information : donnée correcte, complète, mise à jour, précise, vérifiable*
Wang et al. *Programme TDQM* [Wan98] [WKM93] [WSF95]	4 catégories 179 attributs	*Qualité intrinsèque* *Qualité d'accessibilité* *Qualité contextuelle* *Qualité de la représentation*
Redman [Red96]	- 4 dimensions pour les valeurs - 8 dimensions pour le format de représentation	- *Précision, complétude, actualité, cohérence* - *Donnée appropriée, interprétable, portable, précision du format, flexibilité du format, possibilité de représenter les valeurs nulles, utilisation efficace, cohérence*
Calabretto, Pinon, Poullet, Richez [CP+98]	3 critères pour la qualité de l'information	*Disponibilité, Fiabilité, Adaptabilité*
Aebi, Perrochon [AP93]	3 composantes	*Donnée correcte, complète, minimale*
DISA [Wan96, pages 155-171]	6 caractéristiques	*Précision, Complétude, Consistance, Actualité* *Unicité, Validité*
TIPS-QCT [TIPS99]	8 descripteurs pour la qualité des documents	***Qualité scientifique**: exactitude, précision, originalité, complétude, actualité, qualité de démonstration, qualité de la liste des références, qualité de la méthodologie; **Lisibilité**: qualité du style d'écriture, qualité de la structure logique, adéquation des illustrations, absence des répétitions, clarté de l'expression des idées; **Public visé**: niveau technique; **Fraîcheur**: date de publication, **Autorité**: réputation de l'auteur, réputation du journal ou de la conférence; **Disponibilité**: longévité, imprimabilité; **Popularité**: nombre des lecteurs, des citations; **Qualité d'identification**: citabilité*

TAB. 1 - *Quelques propositions caractérisant la notion de qualité de données*

et technique dans le cadre de la gestion d'un projet dédié à la qualité des données. L'article fournit une liste remarquable de références bibliographiques sur le sujet (plus de

120). Chacun des auteurs cités dans le tableau 1 ont proposé et défini plusieurs dimensions de la qualité des données aux niveaux conceptuel et logique, ainsi que des mesures objectives pour les évaluer. La section 4.1 en décrira les plus pertinentes au regard du processus d'ECD. Pour ce qui concerne les autres dimensions non détaillées ici, nous invitons le lecteur à consulter les articles cités en référence.

Certains travaux intègrent totalement la modélisation et la gestion de la qualité des données dès la conception du système d'information. Parmi ces approches de modélisation, le programme *TDQM* (*Total Data Quality Management*) mené par Wang et al. au *Massachusetts Institute of Technology* [Wan98] [WKM93] propose une méthodologie de modélisation basée sur le modèle Entité/Association qui guide pas à pas l'ajout de la dimension qualité sur chaque élément du modèle (entités, attributs, associations). Elle se décompose en quatre étapes décrites ci-après.

Phase 1) Modélisation du domaine d'application

Classiquement, la première étape consiste à modéliser le domaine d'application, sous la forme d'un modèle sémantique étendu de type Entité/Association. La Figure 1, adaptée de [WKM93] présente un exemple de l'achat d'un produit par une société.

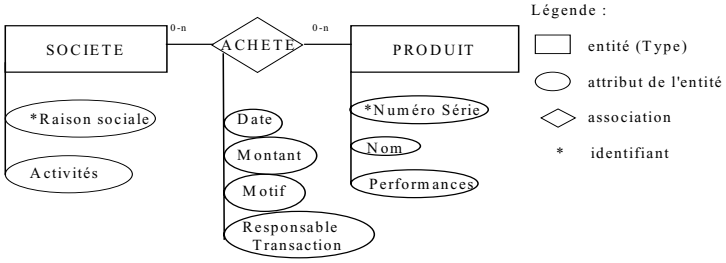


FIG. 1 - *Modélisation du domaine d'application*

Phase 2) Ajout de paramètres subjectifs sur la qualité des données

La seconde étape, présentée dans la Figure 2, vise à rendre explicites des paramètres de qualité subjectifs sur chaque type d'entité, attribut ou association. Ces paramètres subjectifs tels que la cohérence, la criticité, le niveau de détail de la donnée... sont ajoutés en tant que qualificatifs aux attributs de premier ordre définis lors de l'étape de modélisation précédente.

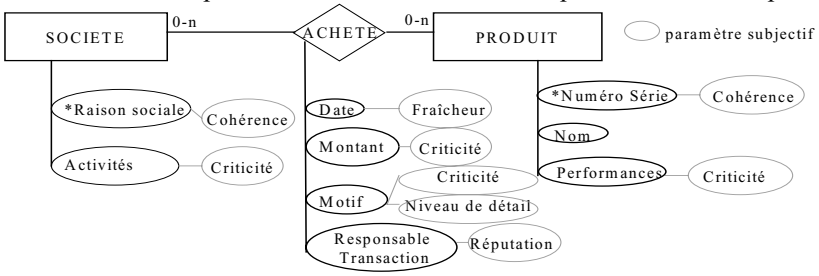


FIG. 2 - *Ajout des paramètres subjectifs concernant la qualité des données*

Phase 3) Ajout d'indicateurs objectifs sur la qualité des données

Des indicateurs de qualité objectifs sont, par la suite, ajoutés (Figure 3). Ils correspondent uniquement aux paramètres mesurables. En effet, les indicateurs objectifs sont des résultats de calculs automatiques (par des méthodes ou des contraintes). Ils sont ajoutés en tant que qualificatifs à l'attribut de l'entité ou de l'association déjà existants pour lequel le paramètre subjectif associé est mesurable.

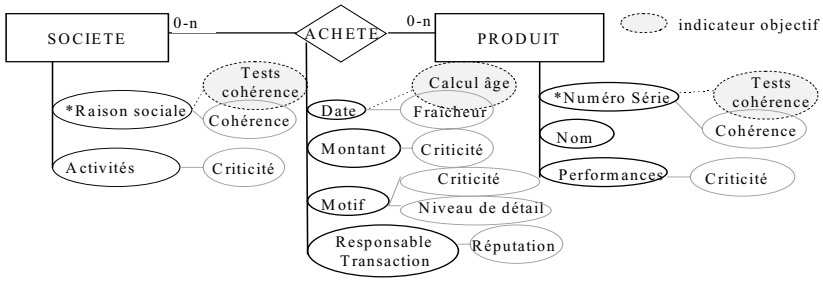


FIG. 3 - *Ajout des indicateurs objectifs concernant la qualité des données*

Phase 4) Intégration des vues-qualité au Modèle Conceptuel de Données

La dernière étape de la modélisation consiste à intégrer, au schéma conceptuel global, les différentes vues-qualité qui ont pu être définies dans les trois phases précédentes [WKM93].

Cette approche a cependant été jugée irréalisable dans la pratique [Red96], car l'ajout de labels sur toutes les entités, relations et leurs attributs semble être une solution trop coûteuse. L'utilisation de méta-données (données sur les données) pour l'évaluation et l'amélioration de la qualité des données a néanmoins été préconisée dans [Rot96]. L'auteur incite les producteurs d'informations à assurer la vérification, validation et certification de leur données (*VV&C : Verification, Validation and Certification*), les méta-données concernant la qualité des données doivent être fournies avec les données pour permettre le processus d'estimation et de maintenance de la qualité des données. Elles sont structurées à différents niveaux d'abstraction : au niveau de la base de données, au niveau des éléments structurels des données (relation, attribut) et au niveau de la valeur de la donnée.

Peu de projets se sont focalisés sur la qualité de données dans un environnement distribué [NLF99][Ber03][CD+97]. Dans le domaine de la modélisation, citons *DWQ* (*Data Warehouse Quality*) [CD+97] [Vas00] qui est un projet centré sur la modélisation formelle de la qualité de l'information pour optimiser la conception d'un entrepôt de données intégrant la gestion de méta-données de qualité.

Plus récemment et dans le contexte des documents et des données semi-structurées, la tendance est bien à l'ajout et l'utilisation de méta-données (labels) pour l'annotation qualitative [CP+98][TIPS99][HBP00][Ber02]. Le projet *DESIRE* [HBP00] a produit une liste détaillée de standards de qualité à utiliser pour choisir des ressources du web selon diverses catégories de critères de qualité : 1) les critères liés à la politique de la diffusion de la ressource, 2) des critères liés au contenu, 3) des critères liés à la forme, 4) des critères liés à la gestion de la qualité de la documentation. Dans le cadre du projet européen *TIPS* [TIPS99], plusieurs services ont été développés pour réutiliser des évaluations faites par des relecteurs humains sur les publications scientifiques : *QCT* (*Quality Control Tools*) vise à

rassembler les évaluations des documents afin d'enrichir leur indexation traditionnelle avec des informations qualité (voir TAB. 1 pour les descripteurs de qualité utilisés dans *TIPS-QCT*).

Le plus souvent, les propositions de définition des dimensions de la qualité des données sont menées dans le cadre d'une démarche plus générale et de nombreuses méthodologies ont été proposées pour mettre en oeuvre la gestion et l'amélioration de la qualité des données dans un système d'information (TAB. 2).

Auteurs	***Nom de la méthodologie***	***Descriptif***
Wang et al. [Wan98][WKM93]	TDQM *Total Data Quality Management*	1) Définir la qualité de l'information : les spécifications de la qualité du produit d'information font partie intégrante du processus de conception 2) Mesurer la qualité de l'information : des métriques de qualité de l'information sont définies pour détecter les erreurs, vérifier la fraîcheur, la complétude, la cohérence des données 3) Analyser la qualité de l'information par des méthodes d'investigation qui recherchent les causes des problèmes de qualité d'information (SPC, analyse sur diagrammes de Pareto) 4) Améliorer la qualité de l'information
Ballou, Tayi [BT89]	Méthodologie d'allocation de ressources pour l'amélioration de la qualité des données	Utiliser efficacement des ressources pour la maintenance de l'intégrité des données par des heuristiques permettant l'équilibre des coûts entre paramètres d'estimation et maintenance de la qualité des données
Paradice, Fuerst [PF91]	Méthodologie pour assurer la qualité des données dans un système d'information de gestion *(MIS Management Information System)*	Formuler le taux d'erreurs sur les données stockées

TAB. 2 - *Quelques méthodologies pour la mise en œuvre de la qualité des données*

3.2. Mesure de la qualité des données

Les premières approches adoptées pour mesurer la qualité des données dans une base de données furent des approches basées sur les méthodes statistiques [Spe85] [LU90] telles que l'inférence sur les données manquantes, la prédiction sur la précision des estimations à partir des données disponibles, l'édition des données, le suivi de données, la détection automatique des erreurs et le traitement des exceptions et données isolées [PF91] [Red96]. Il est toutefois recommandé que la qualité des données soit mesurée sur les aspects de l'information qui sont considérés comme critiques [FLR94] [Bon96]. Les bases de données modélisent une portion du monde réel en constante évolution. Or, dans le même temps, la qualité des données peut devenir obsolète. Le processus d'estimation et de contrôle doit être périodiquement reconduit en fonction de la dynamique du monde réel modélisé ainsi que des changements de centres d'intérêt (des données jugées critiques à un instant donné ne le restent pas nécessairement).

On peut signaler que dans de nombreux cas (pour les systèmes d'information géographique mais aussi en extraction de connaissances à partir de textes), les mesures de qualité sont faites par comparaison avec des données de référence et bien sûr, l'utilisation d'un sous-ensemble certifié de données coûte cher.

Quel que soit le domaine d'application, les mesures de qualité de données les plus fréquemment mentionnées dans la littérature, sont l'exactitude, la complétude, l'actualité, la cohérence [FLR94] [BP95] [RW95] [Red96] dont les définitions générales sont les suivantes (la section 4.1 définit les métriques les plus pertinentes pour le processus d'ECD) :

- l'exactitude se mesure en détectant le taux de valeurs correctes dans la base de données,
- la complétude se mesure en détectant le taux de valeurs manquantes dans la base de données,

- l'actualité se mesure en détectant le taux de valeurs obsolètes dans la base de données par rapport à une date prédéfinie (la fraîcheur se mesure, quant à elle, par une comparaison de la date de saisie à la date courante),
- la cohérence se mesure par rapport à un ensemble de contraintes en détectant les données de la base qui ne les satisfont pas.

Sur le plan formel, [RW95] proposent en particulier une algèbre relationnelle étendue par des estimations de l'exactitude des données basées sur des hypothèses de distributions uniformes des valeurs incorrectes sur l'ensemble des tuples et des attributs de la base de données.

Selon une approche logique, [Cho94] propose une définition axiomatique pour caractériser une base de données résultant d'une fusion de sources d'information contradictoires (bases de connaissances ou bases de données), en les classant selon un ordre total de fiabilité relative aux thèmes d'information qu'elles contiennent. La fusion des différentes sources d'information est menée selon deux attitudes :

- l'attitude suspicieuse qui consiste à suspecter toutes les informations fournies par une source qui contredit une source plus fiable. Elle se traduit par une perte de confiance totale en la source la moins fiable,
- l'attitude confiante qui consiste à suspecter seulement l'ensemble minimal des informations fournies par une source qui contredit une source plus fiable qu'elle.

Dans la plupart des travaux, on suppose initialement que l'évaluation d'un critère de qualité est homogène quelle que soit la donnée (ou la source) considérée. Cette hypothèse de travail est simplificatrice car, en pratique, l'évaluation de certains critères de qualité devrait nécessiter une partition uniforme de l'élément dont la qualité est à évaluer. Par exemple, la crédibilité d'une source de données devrait être ″évaluée par parties″, c'est-à-dire que la source peut être considérée crédible selon les thèmes ou les entités qu'elle renseigne, par analogie avec les travaux de L. Cholvy [Cho94] qui classe les sources selon un ordre total de fiabilité relative à un thème particulier. Certains critères comme la fiabilité d'une source, la réputation des auteurs... imposent des évaluations globales qui ne reflètent pas la réalité d'une appréciation (qui n'est généralement pas homogène sur l'ensemble des informations proposées). Il serait donc plus réaliste de partitionner la couverture descriptive de la source avant d'en évaluer la qualité.

3.3. Amélioration de la qualité des données

3.3.1. Détection des erreurs

L'utilisation des techniques d'apprentissage pour la validation et la correction des données sont abordées dans [PC93]. Ce livre montre en particulier comment les règles inférées à partir des instances de la base de données par les méthodes d'apprentissage peuvent être utilisées pour identifier les exceptions et faciliter le processus de validation des données. D'autres approches similaires peuvent être trouvées dans [Sch91].

La tendance générale est à l'utilisation grandissante des méthodes de l'Intelligence Artificielle pour la validation des données [FP+96] (apprentissage, schémas de représentation des connaissances, gestion de l'incertain...). Ainsi, le prototype décrit dans [SWK93] vérifie

si les données existantes sont correctes : les auteurs parlent de validation et de nettoyage des données au moyen d'un langage de données logique (*LDL++*). Le système emploie des contraintes pour la validation des données et des règles de nettoyage.

L'utilisation de techniques statistiques pour améliorer la correction des bases de données et l'introduction d'un nouveau type de contraintes d'intégrité ont par ailleurs été proposées dans [HZ95]. Les contraintes sont dérivées à partir d'une instance de la base de données en utilisant les méthodes statistiques traditionnelles (échantillonnage, régression...). Chaque mise à jour de la base de données est validée si elle satisfait ces contraintes statistiques.

3.3.2. Nettoyage des données

L'intérêt des métriques sur la qualité des données est de pouvoir les exploiter à des fins d'amélioration de la qualité de la base. Le nettoyage de données fait partie des stratégies d'amélioration automatique de la qualité des données et se décompose en trois étapes : i) auditer les données afin de détecter les incohérences, ii) choisir les transformations pour résoudre les problèmes de non-qualité, iii) appliquer les transformations choisies au jeu de données.

Pour la mise en œuvre, sont utilisés des outils commerciaux d'audit tels que *ACR/Data* d'Unitech Systems ou *Migration Architect* d'Evoke et des outils de transformation (ETL – Extraction/Transformation /Loading) tels que *Data Junction* ou *DataStage* d'Ascential Software.

Parmi les outils académiques d'extraction et de transformation, citons Potter's Wheel [RH01] et AJAX [GF+00] qui permettent l'extraction de structures et expressions régulières, la translation de valeurs de données (par application de fonctions de formatage), la transformation de l'ensemble des valeurs de tuples (lignes) et d'attributs (colonnes). Les opérations de transformation possibles sont récapitulées dans le tableau 3 avec un exemple extrait de [RH01].

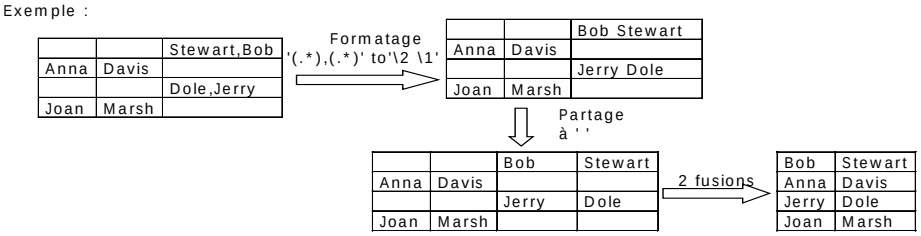
D'autres opérations, telles que le clustering et l'appariement sont employées pour la détection et l'élimination de doublons : *AJAX* [GF+00] est une extension du langage déclaratif SQL permettant de spécifier chaque transformation de données (*matching, merging, mapping, clustering*) nécessaire au processus de nettoyage des données. Ces transformations normalisent si possible les formats de données et détectent les paires d'enregistrements qui se rapportent le plus probablement au même objet. Cette étape d'élimination des doublons est appliquée si des données approximativement redondantes sont trouvées et un appariement multi-table calcule des jointures par similarité entre des données distinctes ce qui permet leur consolidation.

Le nettoyage des données est une étape préalable essentielle au processus d'ECD mais elle demeure coûteuse et sa mise en œuvre avec les outils actuels n'est ni incrémentale ni interactive. Les scripts combinant les différentes opérations de nettoyage décrites précédemment peuvent également introduire de nouvelles erreurs.

TRANSFORMATION	DÉFINITION FORMELLE
Formatage	$\phi(R, i, f) = \{(a_1,\dots, a_{i-1}, a_{i+1},\dots, a_n, f(a_i))\mid (a_1,\dots, a_n) \in R\}$
Ajout	$\alpha(R, x) = \{(a_1,\dots, a_n, x)\mid (a_1,\dots, a_n) \in R\}$
Suppression	$\pi(R, i) = \{(a_1,\dots, a_{i-1}, a_{i+1},\dots, a_n)\mid (a_1,\dots, a_n) \in R\}$
Copie	$\kappa((a_1,\dots, a_n), i) = \{(a_1,\dots, a_n, a_i)\mid (a_1,\dots, a_n) \in R\}$
Fusion	$\mu((a_1,\dots, a_n,), i, j, glue) = \{(a_1,\dots, a_{i-1}, a_{i+1}, \dots, a_{j-1}, a_{j+1},\dots, a_n, a_i \oplus \text{glue} \oplus a_j)\mid (a_1,\dots, a_n) \in R\}$
Division	$\delta((a_1,\dots, a_n), i, \text{pred}) = \{(a_1,\dots, a_{i-1}, a_{i+1}, \dots, a_n, a_i, \text{null})\mid (a_1,\dots, a_n) \in R \wedge \text{pred}(a_i)\} \cup \{(a_1,\dots, a_{i-1}, a_{i+1}, \dots, a_n, \text{null}, a_i)\mid (a_1,\dots, a_n) \in R \wedge \neg\text{pred}(a_i)\}$
Partage	$\omega((a_1,\dots, a_n,), i, splitter) = \{(a_1,\dots, a_{i-1}, a_{i+1},\dots, a_n, \text{left}(a_i, splitter), \text{right}(a_i, \text{splitter}))\mid (a_1,\dots, a_n) \in R\}$
Repliement	$\lambda(R, i_1, i_2,\dots, i_k) = \{(a_1,\dots, a_{i1-1}, a_{i1+1},\dots, a_{i2-1}, a_{i2+1},\dots, a_{ik-1}, a_{ik+1},\dots, a_n, a_{il})\mid (a_1,\dots, a_n) \in R \wedge 1 <= l <= k\}$
Sélection	$\sigma(R, \text{pred}) = \{(a_1,\dots, a_n)\mid (a_1,\dots, a_n) \in R \wedge \text{pred}((a_1,\dots, a_n))\}$

Notation : R est une relation avec n attributs, i et j sont les indices des attributs ; a_i représente la valeur d'un attribut pour un tuple donné; x et *glue* sont des valeurs, f est une fonction de transformation d'une valeur en une autre ; $x \oplus y$ concatène x et y ; splitter est une position dans une chaîne de caractères ou une expression régulière ; left(x, splitter) est la partie gauche de x après avoir partagé la chaîne de caractères à la position indiquée par la variable splitter ; pred est une fonction retournant un booléen.

Exemple :



TAB. 3 - *Opérations de transformations utilisées pour le nettoyage des données*

3.4. La qualité des données dans un domaine d'application spécifique : les données géographiques

La communauté des Systèmes d'Information Géographique (S.I.G.) a adopté, dans la plupart de ses standards, des méta-données qui incluent des spécifications sur les éléments de qualité des données géographiques à différent niveaux, selon la granularité des ensembles de données considérés [GM95] [GJ98]. L'appréciation d'un jeu de données géographiques repose sur l'estimation de plusieurs éléments de qualité distincts pouvant être classés de la façon suivante :

- *les éléments quantitatifs* : résultats de mesure qui sont une estimation directe de la qualité,
- *les éléments qualitatifs* : informations descriptives qui permettent une appréciation indirecte de la qualité d'un jeu de données par un expert,
- *les éléments de qualité spécifiques à l'utilisateur* : informations pouvant compléter la description de la qualité du jeu de données.

Pour implémenter les informations de qualité selon le standard d'échange retenu, deux cas se présentent :

- soit pour chaque élément de qualité, les sous-éléments de qualité associés aux données sont identifiés et leur définition précisément fixée. Cette démarche est suivie dans la pré-norme EDIGéO où les informations de qualité, appelées des *descripteurs de qualité* sont assimilables à des méta-données.
- soit le standard définit un modèle général de description des éléments de qualité qui spécifie leur organisation.

Les éléments de qualité sont décrits dans le diagramme de la Figure 4 selon le formalisme UML.

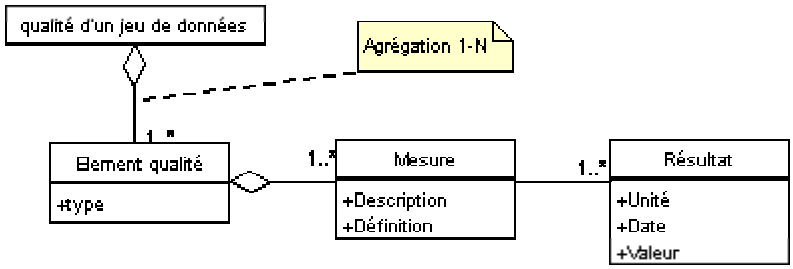


FIG. 4. Diagramme de classes pour la qualité d'un jeu de données

Le tableau 4 d'après [Ber99] propose un récapitulatif des éléments de qualité retenus dans les normes EDIGéO, CEN et ISO.

Dans le domaine des systèmes d'information géographique, seuls les éléments quantitatifs font l'objet d'une estimation directe de la qualité par comparaison avec le terrain nominal (c'est-à-dire le jeu de données de référence). Dans ce contexte, l'ISO identifie deux façons d'évaluer la qualité d'un jeu de données géographiques:

- *l'inspection automatique* : un contrôle systématique de toutes les informations contenues dans le jeu de données est effectué automatiquement. Ce type de contrôle est particulièrement adapté au contrôle de la cohérence logique (contrôle d'intégrité des données). Il est notamment possible de vérifier que l'échange des données est conforme à un format de données, ou que les attributs de chaque entité ont été renseignés conformément aux valeurs prévues dans les spécifications (*tests of valid values*), ou bien que les règles topologiques sont respectées (*specific topological test*),
- *l'évaluation d'après un échantillon* : l'estimation de la qualité des données ne peut se faire sur la totalité des données contenues dans le jeu de données. Le contrôleur extrait alors des sous-ensembles représentatifs du jeu de données et les compare aux données de référence. Pour mesurer et analyser les erreurs, certaines approches (qui s'apparentent à la consolidation de données) s'appuient sur la comparaison du jeu de données avec une source de qualité supérieure ; par exemple, en identifiant plusieurs méthodes pour l'estimation de la précision planimétrique dont (1) l'estimation déductive basée sur la connaissance des erreurs introduites à chaque étape de la production des données (2) la cohérence interne (*internal evidence*) (3) la comparaison avec la source à l'origine des données et (4) la comparaison avec une source de précision supérieure.

		EDIGéO (AFNOR 1992)	CEN (CEN 1996)	ISO (ISO 1998)
Utilisation de la norme		Echange de données	Description des informations de qualité	Principe de description des informations de qualité
Eléments qualité	Précision planimétrique	Précision planimétrique *Précision planimétrique absolue* Précision altimétrique *Précision altimétrique absolue* Précision métrique *Précision métrique absolue*	Eléments qualité considérés comme paramètres de qualité primaires s'appliquant à tous les jeux de données	*Précision absolue* *Précision relative* *Précision de position du pixel* *Précision de forme* *Stabilité de position* *Stabilité de position relative*
	Précision sémantique	*Nombre d'éléments correctement codifiés*		*Précision d'un attribut définie par une valeur* *Précision de classification* *Classification incorrecte*
	Consistance logique	*Nombre d'éléments conformes à la règle*		*Cohérence de domaine* *Cohérence de format* *Cohérence topologique*
	Actualité	*Date d'observation* *Type de mise à jour* *Pérennité de la mise à jour* *Date de mise à jour* *Date de fin de validité* *Historique*		*Précision* *Cohérence temporelle* *Validité temporelle*
	Exhaustivité			*Elément excédentaire* *Elément manquant*
	Généalogie		Production *Producteur* *Cause de production*	
	Usage		*Référence à l'institution utilisant les données* *Type d'usage*	
Ensemble de données ciblé		Jeu de données Catégorie d'objet (entité ou relation) Donnée individuelle	Jeu de données Sous-ensemble du jeu de données	Série de jeu de données Jeu de données Sous-ensemble d'un jeu de données avec des caractéristiques communes

Les sous-éléments de qualité sont représentés en italique.

TAB. 4 - *Éléments de qualité selon les normes EDIGéO, CEN et ISO*

Les éléments de qualité qualitatifs ne font pas l'objet de tests spécifiques et sont estimés par un expert. De façon générale, l'évaluation de la qualité pour un jeu de données géographiques repose sur une estimation des informations de qualité dont les résultats sont présentés de façon synthétique dans les méta-données qui renseignent chaque jeu de données. Des comptes-rendus de qualité peuvent également décrire l'ensemble des mesures et résultats relatifs à la qualité du jeu de données.

Il est intéressant de constater que la communauté des S.I.G. a été pionnière pour proposer des standards incluant des méta-données de qualité et des procédures de contrôle des données, notamment afin d'assurer la recette des données et la certification de leurs systèmes. Cependant, la mise en oeuvre pratique et effective reste problématique à cause de son coût. Les producteurs de données sont chargés de contrôler la qualité des jeux de données fournis à leurs clients ou, tout au moins, de leur fournir des moyens de contrôle des données (par exemple, la source de données de référence).

4. Exploitation des méta-données sur la qualité dans le processus d'ECD

4.1. Définition des principales métriques de qualité des données

Comme nous l'avons déjà évoqué, les métriques permettant d'évaluer les différentes facettes de la qualité des données sont définies le plus souvent *ad hoc* selon les besoins et les contraintes d'une application et/ou d'un usage. Les métriques de qualité des données sont appréhendées notamment dans la littérature sur les bases de données et l'ECD théorique comme des indices de plausibilité et des fonctions générales. Mais leur définition rentre dans les tâches d'ECD. Dans la mesure où aucune étude comparative n'a encore été menée sur l'ensemble des métriques existantes proposées pour l'évaluation de la qualité d'une base ou d'entrepôt de données, la présente section sera volontairement limitée aux principales métriques de qualité des données dans un cadre relationnel. Elle sera articulée selon les niveaux conceptuel et logique.

4.1.1. Métriques pour évaluer la qualité d'une modélisation de type entité-association

Il est intéressant d'analyser la qualité et en particulier la complexité du modèle conceptuel à l'origine du schéma d'une base de données, et ce, principalement pour en évaluer la maintenabilité. Ces mesures sont des supports quantitatifs afin de comparer des alternatives de conception et permettre l'identification des problèmes de conception qui ont nécessairement un impact plus ou moins direct sur la qualité des données stockées. Les métriques proposées dans la littérature et présentées dans le tableau 5 sont des mesures objectives et subjectives de la complexité et des facteurs de qualité d'un modèle conceptuel selon le formalisme Entité-Association. Les métriques relatives aux modèles conceptuels objet ne sont pas présentées ici, car les données exploitées dans le processus de découverte de connaissances sont majoritairement stockées dans des bases et entrepôts relationnels ayant le plus souvent des modèles conceptuels sous-jacents de type entité-association. La validation de ces métriques reste un problème de recherche ouvert.

Du point de vue de la qualité des connaissances extraites à partir des données, les métriques relatives à la qualité et à la complexité du modèle conceptuel originel peuvent être exploitées en tant que méta-données et permettre de nuancer la portée des règles extraites.

AUTEURS	MÉTRIQUES	VALIDATION	
		THÉORIQUE	EMPIRIQUE
Gray et al. [GC+91]	Complexité structurel du modèle : $E=\sum_{i=1}^{n} Ei$ avec n entités Ei Complexité pour l'entité i : $D_i=R_i*(a*FDA_i+b*NFDA_i)$ avec 0 < a <= b, Ri = nombre d'associations, $FDAi$ = nombre d'attributs en dépendances fonctionnelles, $NFDAi$ = nombre d'attributs sans dépendance fonctionnelle	Non	Non
Moody et al. [MS+98]]	**Complétude** : nombre d'items du modèle ne correspondant pas aux spécifications des utilisateurs, nombre de spécifications non représentées dans le modèle, nombre d'items du modèle qui correspondent aux spécifications mais qui sont mal définis, nombre d'incohérences dans la modélisation **Intégrité** : nombre de contraintes non satisfaites par les données, nombre de contraintes incluses dans le modèle qui ne correspondent pas exactement à la réalité modélisée **Flexibilité :** nombre d'éléments dans le modèle sujets à modification, coût estimé des modifications, importance stratégique des modifications **Compréhensibilité** : estimation par l'utilisateur du caractère compréhensible et interprétable du modèle **Correction** :nombre de violations des conventions du modèle de données, nombre de violations des formes normales, nombre d'instances redondantes dans le modèle **Simplicité** : nombre d'entités et associations ; somme pondérée ($aN^E + bN^R + cN^A$), où N^E est le nombre d'entités, N^R le nombre d'associations et N^A le nombre d'attributs **Intégration** : nombre de conflits avec le modèle de données commun, nombre de conflits avec les systèmes existants **Implémentabilité** : estimation du risque technique, estimation du risque de planification, , estimation du coût de développement, nombre d'éléments du niveau physique inclus dans le modèle	Non	Non
Genero et al. [GP+00]	**NE** : nombre total d'entités dans le modèle ; **NA** : nombre total d'attributs d'entités et d'associations (simples ou composés); **DA** : nombre d'attributs dérivés (i.e. attributs dont la valeur peut être déduite ou calculée) ; **CA** : nombre total d'attributs composés ; **MVA** : nombre total d'attributs multi-valués ; **NR** : nombre total d'associations dans le modèle ; **M:NR** : nombre total d'associations M:N ; **1:NR** : nombre total d'associations 1:N et 1:1 ; **N-AryR** : nombre total d'associations N-aires ; **BinaryR** : nombre total d'associations binaires ; **NIS-AR** : nombre total d'associations IS_A (généralisation/ spécialisation) ; **RefR** : nombre total d'associations cycliques ; **RR** : nombre total d'association redondantes	Oui	Partiellement
Piattini et al. [PG+00]	**RvsE:** rapport entre le nombre d'associations NR et le nombre d'entités NE du modèle $RvsE=\left(\frac{NR}{NR+NE}\right)^2$ avec $NR + NE > 0$. **DA:** rapport entre le nombre d'attributs dérivés NDA et le nombre maximal d'attributs dérivés possibles NA : $DA=\frac{NDA}{NA-1}$ avec $NA > 1$. **CA:** rapport entre le nombre d'attributs composés NCA et le nombre d'attributs total NA : $CA=\frac{NCA}{NA}$ avec $NA > 0$. **RR:** rapport entre le nombre d'associations redondantes NRR et le nombre d'associations NR : $RR=\frac{NRR}{NR-1}$ avec $NR > 1$. **M:NRel:** rapport entre NM:NR, le nombre d'associations M:N sur le nombre d'associations NR : $M:NRel=\frac{NM:NR}{NR}$ avec $NR > 1$. **FLeaf :** mesure de la complexité des hiérarchies (IS_A) : $FLeaf=\frac{NLeaf}{NE}$ avec $NLeaf$: nombre d'entités filles dans une hiérarchie généralisation/spécialisation et NE nombre d'entités dans chaque hiérarchie, $NE > 0$. **IS_Arel** : nombre moyen de supertypes directs et indirects par entité non racine $ALLSup$: $IS\_ARel=FLeaf-\frac{RLeaf}{ALLSup}$	Non	Partiellement

TAB. 5 - *Métriques utilisées pour évaluer la complexité et la qualité d'un modèle conceptuel de données*

Les métriques multi-dimensionnelles telles que la complétude, l'intégrité, la flexibilité définies par Moody et al. [MS+98] et rappelées dans le tableau 5 peuvent être des indicateurs à apporter aux connaissances extraites par ECD puisque les données exploitées risquent de ne pas correspondre aux spécifications initiales (complétude), ne pas respecter des contraintes du domaine modélisé (intégrité) ou encore peuvent s'avérer obsolètes ou inutiles (flexibilité). D'autre part, les métriques indiquant le nombre d'associations redondantes ou d'attributs composés peuvent corroborer certaines règles triviales.

4.1.2. Métriques pour évaluer la qualité d'une base relationnelle

Parmi les nombreuses métriques proposées dans la littérature pour évaluer différentes dimensions de la qualité des données, citons les travaux de Naumann et al. [NL99][Nau02] qui définissent plusieurs critères de qualité objectifs et subjectifs permettant d'établir des stratégies d'exécution de requêtes dans une architecture de type médiateur/adaptateurs (incluant la création, la sélection et l'ordonnancement de plans de requêtes selon la qualité des données et des sources). Dans ce contexte, plusieurs sources de données hétérogènes peuvent répondre à tout ou partie d'une requête. Ainsi, la qualité du résultat d'une requête est fonction de la qualité des sources et des données qui le compose. Le tableau 6 en présente les définitions. Les travaux de Motro et Rakov [MR97] sont intéressants dans la mesure où ils proposent et définissent la notion d'homogénéité de la qualité. En effet, la qualité d'une base de données, d'une vue ou d'un résultat de requête n'est bien souvent pas uniforme sur l'ensemble des données considérées. [MR97] propose une métrique pour évaluer l'homogénéité de la précision des données. Ces travaux méritent d'être adaptés et étendus aux différentes dimensions de la qualité des données.

AUTEURS	DEFINITION DES METRIQUES	
Naumann, Leser [NLF99]	Spécifiques à la source de données	**Facilité de compréhension** : jugement de l'utilisateur de 1 à 10 basé sur la présentation des données **Réputation** : jugement de l'utilisateur de 1 à 10 basé sur des préférences personnelles et son expérience **Fiabilité** : score de 1 à 10 basé sur la précision des méthodes expérimentales de recueil et/ou de production des données **Actualité** : fréquence de mise à jour mesurée en nombre de jours
	Spécifiques aux requêtes	**Disponibilité** : pourcentage de temps pendant lequel la source de données est accessible **Prix** : prix d'une requête en US dollars **Temps de réponse** : temps d'attente moyen en secondes par requête **Précision** : pourcentage de données sans erreur **Pertinence** : pourcentage d'objets du monde réel représentés par la source de données
	Spécifiques aux attributs	**Complétude** : pourcentage de valeurs non nulles
Motro, Rakov [MR97]	**Complétude** : proportion des informations correctement stockées D par rapport au monde réel modélisé W: $C=\frac{\lvert D\cap W\rvert}{\lvert W\rvert}$ **Précision** (soundness) : proportion des informations correctement stockées : $S=\frac{\lvert D\cap W\rvert}{\lvert D\rvert}$ **Homogénéité de la qualité d'une vue** : différentiel moyen de précision entre une vue v et toutes ses sous-vues possibles v_i $Hs(v)=\frac{1}{N}\sum_{i=1}^{N}\lvert S(v)-S(v_i)\rvert$	
Labrinidis, Roussopoulos [LR01]	**Fraîcheur** : probabilité d'accéder une version à jour d'une vue v de la base de données sur une période donnée T = [ti,tj] : $Fresh(v)=\frac{1}{T}\times\int_{ti}^{tj} f(d_i)^t$ avec $f(d_i)^t=\begin{cases}0, si la donnée di est obsolète au temps t\\1, si la donnée di n'est pas obsolète au temps t\end{cases}$	

TAB. 6 - *Métriques utilisées pour évaluer la qualité des données*

La qualité des données n'est jamais définitivement acquise et elle évolue à chaque instant. En effet, à chaque action sur les données, peuvent être introduites des erreurs. Ne rien faire et laisser "vieillir" les données engendre également des problèmes de non-qualité. En cela, de récents travaux tels que ceux de Labrinidis et Roussopoulos [LR01] sur la fraîcheur des données proposent une mesure de probabilité de la fraîcheur des données (TAB. 6).

Ces métriques ainsi que toutes celles proposées de façon plus informelle dans [FLR94] [BP95] [Red96] et [WKM93] (cf. TAB. 1) peuvent constituer un ensemble important de méta-données permettant de renseigner les connaissances extraites à partir des données. Mais, en ajoutant aux résultats du processus d'ECD un tel tableau de bord de la qualité des données, le risque est d'en affecter la lisibilité et l'interprétabilité. D'un autre côté, l'alternative à l'évaluation de la qualité des données au sein du processus d'ECD pourrait consister à évaluer la qualité des régularités trouvées (par un calcul de l'espérance, la moyenne ou une agrégation de la qualité des données couvertes)... or cette possibilité conduit à calibrer différemment la recherche (en l'occurrence par des méta-requêtes telles qu'il est inutile de rechercher des régularités dont la fréquence est de même ordre que le bruit). Ici, connaître la qualité locale des données est pertinent car cela permet d'ajuster les seuils de fréquence raisonnables en fonction de la région explorée.

4.2. Description d'une méthode pour intégrer la qualité des données au processus d'ECD

Dans cette section, nous proposons une méthodologie pour intégrer et exploiter les méta-données décrivant la qualité des données grâce à la fusion de ses indicateurs en amont du processus d'extraction de connaissances à partir des données (Figure 5).

4.2.1. Contrôler la qualité des données en amont du processus d'ECD

L'étape E1 consiste à sélectionner les sources de données qui vont alimenter la base et à mettre en place des procédures de recueil (saisie ou import massif) et d'intégration des données. Dans l'étape E2, il s'agit de sélectionner les données *critiques* de la base de données (vue matérialisée). Toutes les données n'ont pas la même importance, elles ne sont pas équivalentes d'un point de vue "stratégique" pour l'entreprise et ne doivent donc pas être considérées de façon uniforme. La notion de *criticité* est utilisée pour comparer l'importance des données les unes par rapport aux autres vis-à-vis du traitement ou de l'objectif envisagé. Elle est déterminée par l'expert ou de façon semi-automatique. Les critères de qualité sont également choisis et spécifiés à cette étape.

L'étape E3 consiste à calculer les mesures de qualité des données (scores) pour les critères retenus à l'étape précédente (par exemple, la complétude, la fraîcheur ou l'exactitude des données) et à les stocker sous forme de méta-données. Un *score de qualité* est calculé à une date donnée par la combinaison pondérée des résultats de mesures directes et indirectes, quantitatives et qualitatives des critères de qualité. L'ensemble des méta-données de qualité sont associées aux données. Le but de l'étape E4 est (1) de détecter les *problèmes de qualité des données* (les erreurs ou les données de qualité médiocre) tout au long du traitement de l'information et (2) d'en analyser les causes.

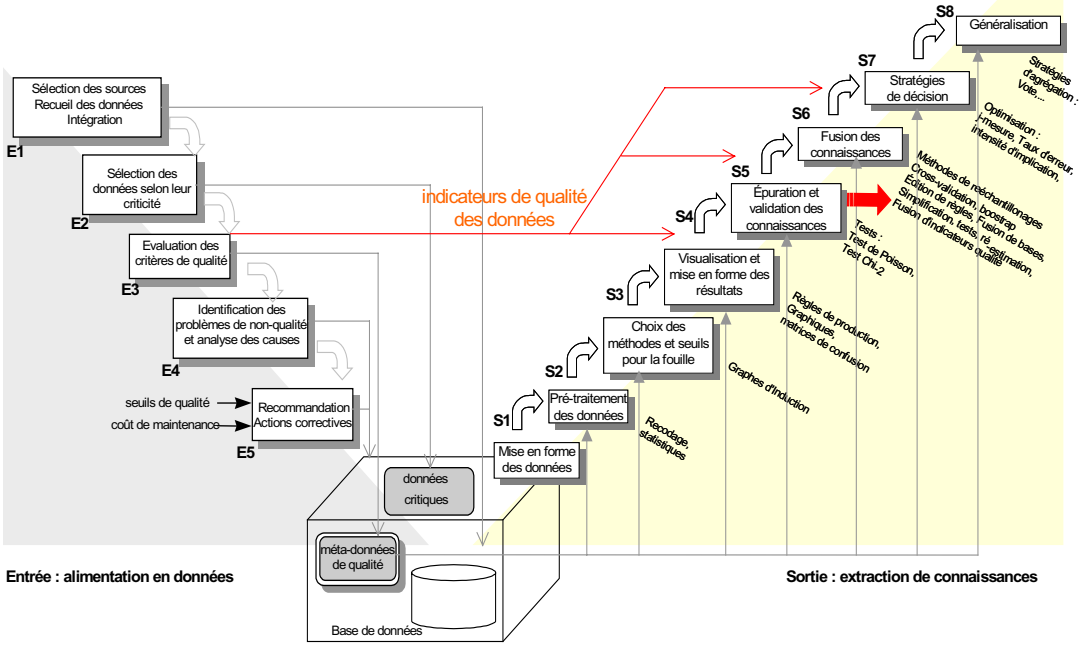


FIG. 5. Démarche générale pour contrôler la qualité des données pour l'ECD

L'étape E5 a pour but de proposer une recommandation des données jugée de "meilleure qualité" compte-tenu des coûts d'acquisition et d'expertise ou, tout au moins, des données qui paraissent les plus appropriées aux besoins de l'utilisateur requérant ou au processus d'ECD qui sera mis en œuvre par la suite.

4.2.2. Préparer les données et exploiter les méta-données

L'étape de préparation des données (regroupant les étapes S1 et S2) consiste en une succession de tâches telles que le choix des objets, des variables, leur pondération, le nettoyage, le recodage des données, le traitement des données manquantes, la détection des données anormales, l'homogénéisation des jeux de données, la discrétisation des attributs continus et la réduction de la dimension. Les données atypiques posent souvent problème lors de l'extraction de connaissances. Soit on les considère comme des erreurs de saisie (étape E4), auquel cas on procède aux mêmes traitements que pour les données manquantes, soit on les considère comme une observation hors champ d'étude. S'il est difficile d'expliquer la particularité d'une observation, le risque est de ne pas considérer un individu porteur d'informations qui peuvent être essentielles pour la compréhension des mécanismes que l'on cherche justement à mettre à jour. L'intérêt des méta-données relatives à la qualité des données (issues de l'étape E3) réside, à ce stade, dans leur exploitation afin de renseigner et présenter objectivement la qualité des règles extraites (voire conforter le caractère exceptionnel de certaines règles - étapes S5, S6 et S7).

4.3. Proposition de prise en compte des indicateurs de la qualité des données dans le processus d'ECD

Selon l'hypothèse que la qualité d'une règle d'association dépend de la qualité des données sur lesquelles elle se base, notre démarche est de proposer une méthodologie qui repose sur la donnée de critères de qualité et sur un mécanisme d'agrégation (fusion). L'objet de cette section est d'intégrer des métriques de qualité de données au calcul de la qualité d'une règle d'association. Nous donnons ci-après les définitions qui formalisent notre proposition.

Définition 1. ***Vecteur qualité d'un jeu de données :*** Soient le jeu de données D_i et c_{ij} un score de qualité du jeu de données D_i pour le critère j ($j=1,\ldots,k$) défini sur [0,1] à l'issue de l'étape E3. Le vecteur qualité du jeu de données $Q(D_i)$ est défini comme étant le vecteur colonne $k \times 1$ des scores de qualité sur les k critères, dont la transposée est :
$Q(D_i)^T = [c_{i1},\ldots, c_{ik}]$.

Définition 2. ***Qualité d'un jeu de données*** **:** Soient le jeu de données D_i et le vecteur qualité de ce jeu de données $Q(D_i)$. On note w_j avec $j = 1,\ldots,k$, le poids affecté au score de chaque critère de qualité c_{ij} avec les deux hypothèses suivantes : $w_j \geq 0$ et $\sum_{j=1}^{k} w_j = 1$

Les poids $w_1,\ldots,w_k$ sont des constantes positives qui reflètent l'importance des critères de qualité les uns par rapport aux autres. La qualité du jeu de données D_i est définie et calculée en fonction d'une méthode de pondération et d'ordonnancement sur le vecteur qualité de $Q(D_i)$ telle que : $Quality(D_i) = ranking(Q(D_i),W)$ où $ranking \in [0,1]$ et W, le vecteur $k \times 1$ des poids w_j pour $j = 1,\ldots,k$. La fonction *ranking* peut être choisie parmi les méthodes d'aide multicritère à la décision telles que SAW, TOPSIS [HY81], AHP [Saa80], ELECTRE ou DEA [Nau02].

Définition 3. Qualité d'une règle : Soit une règle d'association R de la forme $X \quad Y$ dans laquelle la prémisse X, et la conclusion Y sont des conjonctions de variables telles que : l'extension de X est $g(X)= x_1 \wedge x_2 \wedge \ldots \wedge x_n$ et l'extension de Y est $g(Y)= y_1 \wedge y_2 \wedge \ldots \wedge y_{n'}$. Nous définissons la qualité de la règle d'association R par une fonction de fusion "**o**" des scores de qualité des jeux de données présents en prémisse et en conclusion de la règle telle que :
$Quality(R) = Quality(X)$ o $Quality(Y)$
$= Quality(x_1)$ o $Quality(x_2)$ o … o $Quality(x_n)$ o $Quality(y_1)$ o $Quality(y_2)$ o … o $Quality(y_{n'})$

Définition 4. Fusion des scores de qualité : Soit T le domaine de valeurs d'un score de qualité c_{ij} du jeu de données D_i, la fonction de fusion des scores, notée o est une fonction commutative et associative de $T \times T \quad T$. La fonction de fusion peut avoir différentes interprétations selon le critère de qualité considéré, afin de refléter les propriétés de la mesure de chaque critère. Le tableau 7 présente quelques exemples de fonctions de fusion permettant d'agréger les scores de qualité pour deux jeux de données x et y sur chacun des 3 critères de qualité précédemment définis.

CRITÈRE DE QUALITÉ	FONCTION DE FUSION DE SCORES "o"	QUALITÉ DE LA RÈGLE x -> y
Fraîcheur	$\min[c_X, c_Y]$	La fraîcheur de la règle *x* -> *y* est jugée de façon pessimiste égale au plus faible score de fraîcheur, donc à c_Y.
Précision	$c_X \cdot c_Y$	La précision de la règle *x* -> *y* est définie comme la probabilité de précision des jeux de données *x* et *y*.
Complétude	$c_X + c_Y - c_X \cdot c_Y$	La complétude de la règle *x* -> *y* est définie comme la probabilité que l'un des deux jeux de données soit complet.

TAB. 7 - *Exemple de fonctions pour la fusion des scores de qualité des données*

Les limites de ce travail se situent à trois niveaux :

- pour les données qui ont permis d'établir la prémisse et la conclusion d'une règle d'association, le problème de l'uniformité (ou d'homogénéité) des mesures des critères de leur qualité risque de biaiser l'interprétation de la qualité de la règle ;
- la portée de ces indicateurs dont le but n'est que d'informer sur la qualité des données à l'origine des règles peut être discutable : il y a dans notre définition de la qualité d'une règle une volonté de prise en compte de la qualité des données initiales mais qui s'accompagne d'une « perte du contexte », en particulier sur la manière de fusionner les critères, aussi, est-il préférable d'effectuer l'agrégation des indicateurs le plus tard possible pour éviter le lissage ;
- le consensus sur la définition des critères de qualité des données n'étant atteint, leur fusion pour qualifier la qualité d'une règle d'association peut engendrer des problèmes d'interprétation et réduire la lisibilité des règles déjà difficile, de plus la définition des fonctions de fusion selon chaque critère reste un problème difficile.

4.4. Discussion

L'éclairage que nous avons proposé pour ce panorama sur les travaux liés à la qualité des données a permis de souligner la place prédominante qu'ont les méthodes statistiques dans la mesure de la qualité.

Les approches que nous avons présentées précédemment laissent toutefois des problèmes ouverts et bon nombre de perspectives de recherche dans les domaines de : (1) la modélisation (2) la mesure et (3) la mise en oeuvre de la qualité des données intégrées ou non au processus d'extraction de connaissances :

- en l'absence de consensus sur la définition de la qualité, il est souvent préférable d'opter pour la flexibilité : d'une part, en proposant à l'utilisateur un ensemble minimal de critères de qualité de données qui devraient être évalués de façon systématique et en lui laissant le choix des critères à considérer, par la suite, lors de la validation des règles extraites, et d'autre part, en lui fournissant les moyens de définir lui-même un ensemble cohérent de critères de qualité pour ses données. Ici encore, la détermination d'un ensemble minimal et cohérent de critères de qualité est problématique ;
- l'approche souvent préconisée pour la mesure de la qualité est mixte puisqu'il est question d'associer la composante quantitative (calculs statistiques) et la composante

qualitative (expertise plus subjective) pour la mesure de la qualité des données ; cette "mixité" devrait légitimement transparaître dans la mesure de la qualité des règles pouvant être extraites à partir des données ;

- à partir des nombreux travaux sur la qualité des données [Red96] [Wan98][Vas00][CP98] [Bon96] [SK97][Wan96], il semble intéressant de proposer un support méthodologique pour la mise en oeuvre de la qualité des données dans le processus d'extraction de connaissances.
- dans bien des situations, des données peu précises et/ou peu fiables peuvent être acquises relativement vite et facilement, alors que l'acquisition de données précises et plus exactes nécessite plus de temps, de moyens financiers et humains. Il s'agit d'un problème d'optimisation de ressources [BP95] et de modèle de coût [BP02]. Le but est de réaliser un bon équilibre entre les deux, selon le projet à mener. Les problèmes d'optimisation de ressources qui surviennent dans le contexte de la qualité des données ont été soulevés dans [BT89] [BP02]. Dans un programme d'amélioration de la qualité des données ou bien encore de mesure de la qualité des règles extraites, il est nécessaire de considérer le problème de l'allocation optimale de ressources limitées, en utilisant différents paramètres tels que le coût des erreurs, le taux d'erreurs et l'efficacité des procédures de correction. Il est clair que le coût de la qualité des données/règles est important, que ce soit (1) à l'acquisition des données (2) à l'expertise de leur qualité et (3) à la maintenance des méta-données associées (4) ou à l'extraction et l'élagage des règles.

5. Conclusion

La plupart des bases de données produites actuellement sont entachées d'incertitude et contiennent des erreurs ou des données de qualité "médiocre". Face à cette réalité, aux montants financiers en jeu et à la multiplication des échanges de données, tous les acteurs impliqués dans l'utilisation et l'analyse des données ont pris conscience de la nécessité de mesurer et contrôler la qualité des données. Le contrôle de la qualité des données doit en premier lieu se dérouler lors de la production ou lors de l'acquisition des données et être considéré au final dans l'interprétation et la validation des règles extraites par le processus d'ECD.

L'objet de cet article a été dans un premier temps, de dresser les causes de la non-qualité des données, et de présenter un panorama des travaux sur la qualité des données, dans la mesure où ces travaux sont à prendre en compte pour la qualification et l'amélioration de la qualité des connaissances extraites à partir des données. Après avoir défini formellement les métriques de qualité des données exploitables, nous proposons une méthode qui permet l'intégration des méta-données de qualité des données au processus d'ECD par fusion des indicateurs renseignant la qualité des connaissances extraites.

Références

[AP93] Aebi D., Perrochon L., Estimating data accuracy in a federated database environment. Proc. of the 7th Intl. Conf. on Information Systems and Management of Data (CISMOD'93), 1993.

[Bas95] Bash R. (Ed.), Electronic informartion delivery : ensuring quality and value, 1995.
[Ber99] Berti L., Qualité des données et leur recommandation : modèle conceptuel, formalisation et application à la veille technologique, Thèse de l'Université de Toulon et du Var, 1999.
[Ber02] Berti L., Annotation et recommandation collaboratives de documents selon leur qualité. Revue ISI-NIS, Numéro Spécial Recherche et Filtrage d'Information, 7(1-2/2002):125-156, 2002.
[Ber03] Berti L., Quality-extended query processing for distributed sources. Proc. of the Intl. Workshop on Data Quality in Cooperative Information Systems, DQCIS'2003, Siena, Italy, Janvier 2003.
[Bon96] Bonjour E., La qualité et la mise à jour des Bases de données Techniques utilisées en GPAO, Thèse de l'Université de Franche-Comté, 1996.
[BP85] Ballou D., Pazer H., Modeling data and process quality multi-input, multi-output information systems, Management Science, 31(2):150-162, 1985.
[BP95] Ballou D.P., Pazer H.. Designing information systems to optimize the accuracy-timeliness tradeoff. Information Systems Research, 6(1), 1995.
[BP02] Ballou D.P., Pazer H., Modeling completeness versus consistency tradeoffs in information decision contexts. IEEE TKDE, 15(1):240-243, 2002.
[Bro80] Brodie M. L., Data quality in information systems, Information and Management, vol. 3, p. 245-258, 1980.
[BT89] Ballou D.P., Tayi G.K., Methodology for allocating ressources for data quality enhancement. Com. of the ACM, 32(3):320 - 329, 1989.
[CDL+97] Clavanese D., DeGiacomo G., Lenzerini M., Nardi D., Rosati R.. Data integration in datawarehousing. Technical Report DWQ-UNIROMA-001, DWQ Consortium, 1997.
[Cho94] Cholvy L., A logical approach to multi-source reasoning. Lecture Notes in Artificial Intelligence, volume 808, 1994.
[CM95] Celko J., McDonald J. Don't warehouse dirty data. Datamation, 41(18), 1995.
[CP98] Chengalur-Smith I., Pipino L. (Ed.), Proc. of the 3rd Intl. Conf. on Information Quality, MIT, Cambridge, 1998.
[CP+98] Calabretto S., Pinon J. M., Poullet L., Richez M.-A., De la qualité de l'information à la qualité de la documentation, Document Numérique, 12(1):37-52, 1998.
[DR92] Delen G., Rijsenbrij D., The specification, engineering and measurement of information systems quality, Journal of Software Systems, no.17, p. 205-217, 1992.
[FLR94] Fox C., Levitin A., Redman T., The notion of data and its quality dimensions, Information Processing and Management, vol. 30, no. 1, 1994.
[FP+96] Fayyad U., Piatetsky-Shapiro G., Smyth P., Uthurusamy R. (Ed.), Advances in Knowledge Discovery and Data Mining, AAAI Press, MIT Press, 1996.
[GC+91] Gray R., Carey B., McGlynn N., Pengelly A., Design metrics for database systems, *BT Technology J.,* 9(4), 1991, 69-79.
[GF+00] Galhardas H., Florescu D., Shasha D., and Simon E., An Extensible Framework for Data Cleaning, Proc. of the 16th Intl. Conf. on Data Engineering (ICDE 2000), 2000.
[GJ98] Goodchild M., Jeansoulin R. (Ed.), Data quality in geographic information : from error to uncertainty, Hermès, 1998.
[GM95] Guptill S.C., Morrison J.L., ed. Elements of spatial data quality. Elsevier Applied Science, 1995.
[GP+00] Genero M., Piattini M., Calero C., Serrano M., Measures to get better quality databases, *ICEIS 2000,* Stafford, July 2000, 49-55.

[HG+95] Hammer J., Garcia-Molina H., Widom J., Labio W., Zhuge Y. The Standford data warehousing project. Data Engineering Bulletin, 18(2):41 - 48, 1995.
[HBP00] Hiom D., Belcher M., Place E., People Power and the Web: Building Quality Controlled Portals, TERENA Networking Conference, 2000. http://www.desire.org/
[HY81] Hwang, C.-L. and K. Yoon, Multiple Attribute Decision Making, Lecture Notes in Economics and Mathematical Systems, 186, Springer, 1981.
[HZ95] Hou W.C., Zhang Z., Enhancing database correctness : a statistical approach. Proc. of the ACM SIGMOD Intl. Conference on Management of Data, 1995.
[Jac92] Jacso P., CD-ROM software, dataware and hardware : evaluation, selection and installation, 1992.
[Lau86] Laudon K. C., Data quality and due process in large interorganizational record systems, Com. of the ACM, p. 4-11, 1986.
[LR01] Labrinidis A., Roussopoulos N., Update propagation strategies for improving the quality of data on the web, Proc. of the 27th VLDB Conf., 2001.
[LU90] Liepins G., Uppuluri V., Data quality control : theory and pragmatics, M. Dekker, New-York, 1990.
[MR97] Motro A., Rakov I., Not all answers are equally good : estimating the quality of database answers. Flexible Answering Systems, Kluwer Academic Publishers, p.1-21, 1997.
[MR00] Mihaila G. A., Raschid L., and Vidal M.-E., Using quality of data metadata for source selection and ranking. Proc. of the WebDB Workshop, pages 93-98, 2000.
[MS+98] Moody L., Shanks G., Darke P., Improving the Quality of Entity Relationship Models - Experience in Research and Practice, *Proc. of the 7th Intl. Conf. on Conceptual Modelling (ER '*98), p. 255-276, 1998.
[Nau02] Naumann F., Quality-Driven Query Answering for Integrated Information Systems, Springer 2002.
[NLF99] Naumann F., Leser U., and Freytag J., Quality-driven integration of heterogeneous information systems. Proc. of the 25th VLDB Conf., 1999.
[Pat93] Patterson B., The need for data quality. Proc. of the 19th VLDB Conf., 1993.
[PC93] Parsaye K., Chignell M. Data quality control with smart databases, AI Expert, 8(5):23 - 27. 1993.
[PF91] Paradice D.B., Fuerst W.L., An MIS data quality management strategy based on an optimal methodology. Journal of Information Systems, 5(1):48 - 66, 1991.
[PG+00] Piattini M., Genero M., Calero C., Polo C., Ruiz F., Database Quality, In Chapter 14: Advanced Database Technology and Design, Eds. Mario Piattini and Oscar Díaz, Artech House, 2000, 485-509.
[Red96] Redman T., Data quality for the information age, Artech House Publishers, 1996.
[RH01] Raman V., Hellerstein J. M., Potter'swheel : an interactive data cleaning system, Proc. of the 27th VLDB Conf., 2001.
[Rot96] Rothenberg J., Metadata to support data quality and longevity, Proc. of the 1st IEEE Metadata Conf., 1996.
[RW95] Reddy M. P., Wang R., Estimating data accuracy in a federated database environment, Proc. of the 9th Intl. Conf. CISMOD, p. 115-134, 1995.
[Saa80] Saaty, T. L., The Analytic Hierarchy Process. New York: McGraw-Hill, Inc., 1980.
[Sch91] Schlimmer J., Learning determinations and checking databases, Proc. of the AAAI-91 Workshop on Knowledge Discovery in Databases, 1991.

[SK97] Strong D., Kahn B. (Ed.), Proc. of the 2nd Intl. Conf. on Information Quality, MIT, Cambridge, 1997.
[SLW97] Strong D., Lee Y., Wang R., Data quality in context, Com. of the ACM, 40(5), p. 103-110, 1997.
[Spe85] Spencer B.D. Optimal data quality. American Statistical Association, 80(391):564 - 573, 1985.
[SWK93] Sheth A., Wood C., Kashyap V., Q-data : Using deductive database technology to improve data quality, pp. 23 - 56. Kluwer Academic Press, 1993.
[TB98] Tayi G. K., Ballou D. P., Examining Data Quality, Com. of the ACM, 41(2):54-57, 1998.
[TIPS99] TIPS Documentation, Quality Control Tools User Requirements, V-Framework Programme IST-1999-10419, 1999. http://tips.sissa.it
[Vas00] Vassiliadis P., Data Warehouse Modeling and Quality Issues, PhD thesis, Department of Electrical and Computer Engineering, National Technical University of Athens (Greece), 2000.
[Wan96] Wang R. (Ed.), Proc. of the 1st Intl. Conf. on Information Quality, MIT, Cambridge, 1996.
[Wan98] Wang R., A product perspective on Total Data Quality Management, Com. of the ACM, 41(2): 58-65, 1998.
[WKM93] Wang R., Kon H. B., Madnick S. E., Data quality requirements analysis and modeling, Proc. of the 9th Intl. Conf. on Data Engineering, p. 670-677, 1993.
[WSF95] Wang R., Storey V., Firth C., A framework for analysis of data quality research, IEEE Transactions on Knowledge and Data Engineering, 7(4):670-677, 1995.

Summary

Current works on Knowledge Discovery in Databases (KDD) are focused on retrieving relevant patterns which one wishes to be able to qualify the interest or the exceptional character, but whose validity obviously depends on data qulity. Upstream the KDD process, it seems essential to evaluate the quality of the data stored in the databases and datawarehouses in order to: (1) to propose to the users a critical expertise on the quality of the system content, (2) to direct the knowledge extraction according to a targeted profile of users and decision makers, (3) to allow them to relativize confidence which they could grant to the data and the extracted rules, and thus, to allow them to better adapt their data usage, (4) to finally ensure the validity and the interest of the knowledge extracted from the data. This article gives a synthesis of the state of art in the field of data quality while presenting, initially, the causes of data non-quality, then by describing a panorama of works on data quality, relevant works since one is interested to model, measure and to improve quality of the knowledge generated from data. Lastly, the article proposes to exploit the metadata describing data quality in the KDD process.
Key words. Data quality, metadata.