

---

# Renseigner la qualité des connaissances par la fusion d'indicateurs sur la qualité des données

**Laure Berti-Equille**

*IRISA, Campus Universitaire de Beaulieu  
35042 Rennes, France  
Laure.Berti-Equille@irisa.fr*

---

*RÉSUMÉ. L'objet de cet article est de présenter des travaux en cours sur la mesure d'incidence que peut avoir la qualité des données sur la qualité des règles d'association extraites. Dans un premier temps, l'article présente un bref panorama des travaux sur la qualité des données, travaux non négligeables dès lors que l'on s'intéresse à qualifier et à améliorer la qualité des connaissances extraites à partir des données ; une seconde partie présente une méthodologie pour contrôler la qualité des données dans le processus d'extraction de connaissances à partir des données. Nous proposons ensuite de fusionner les indicateurs de la qualité des données afin de renseigner la qualité des règles extraites. Notre approche offre de nombreux avantages dans l'étape de qualification, de validation et de rejet de règles.*

*ABSTRACT. The aim of the article is to present our current works on measuring the impact of data quality on the quality of extracted association rules. The first part of the article reviews previous work on data quality. These quality issues are specially relevant for qualifying and improving the quality of the knowledge discovered into database systems ; the second part presents a methodology for controlling data quality in the process of knowledge discovery in databases. Our approach consists in the fusion of data quality indicators in order to add meta-information on discovered knowledge quality and provides several advantages for the qualification and validation of extracted rules.*

*MOTS-CLÉS : qualité des données, fusion, qualité des connaissances, ECD.*

*KEYWORDS: data quality, fusion, knowledge quality, KDD*

---

## 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, seule la mise en évidence de liens cachés ou de phénomènes de causalités non-triviales à partir de grandes quantités de données aidera les décideurs dans leurs choix stratégiques. 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 [Vas00] [CD+97]. 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 néanmoins de détecter des erreurs, des exceptions ou des incohérences d'un jeu de données, mais elles ne permettent pas de discerner et de qualifier les données de qualité douteuse : l'expertise humaine est ici souvent 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 d'autant plus incontournable dans un contexte de validation des règles issues d'un processus d'ECD.

L'objet de cet article est de présenter nos travaux en cours sur l'incidence que peut avoir la mauvaise qualité des données sur la qualité des connaissances obtenues par extraction de règles d'association. Il nous semble essentiel d'évaluer préalablement la qualité des données stockées dans les systèmes de stockage (base, entrepôt de données, système d'information) afin de : 1) proposer aux utilisateurs une expertise critique de la qualité de leur contenu, 2) orienter l'extraction des connaissances en fonction des profils d'utilisateurs, 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 finalement la validité et l'intérêt des connaissances extraites à partir des données.

L'article s'organise de la façon suivante : la section 2 présente un panorama des travaux menés sur la qualité des données. La section 3 présente une méthodologie générale pour le contrôle de la qualité des données dans le processus d'extraction de connaissances à partir des données. La section 4 décrit la fusion d'indicateurs de la qualité des données pour renseigner la qualité des règles extraites. La section 5 conclut l'article et présente nos perspectives de recherche et de développement.

## 2. Principaux travaux sur la qualité des données

Si la nécessité du contrôle et de la gestion de la 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 définition même de qualité [TB98]. 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 ou attributs [Bro90] [DR92][WSF95][FLR94][Red96], mais aucune définition ne fait aujourd'hui l'unanimité. Quel que soit le domaine d'application (par exemple, dans le domaine des Systèmes d'Information Géographique [GJ98]), les mesures de qualité de données les plus fréquemment mentionnées dans la littérature, sont l'exactitude, la complétude, l'actualité, la cohérence. Pour la suite de l'article et à titre d'exemple, nous utilisons les définitions suivantes pour ces quatre critères de qualité de données (pour plus de détail sur les mesures de qualité des données, nous renvoyons le lecteur à [Red96] [WSF95][FLR94]): (i) l'exactitude d'un jeu de données est définie par le rapport du nombre de valeurs correctes (c'est-à-dire identiques à des données de référence) sur le nombre total de valeurs du jeu de données, (ii) la complétude d'un jeu de données est définie par le rapport du nombre de valeurs nulles sur le nombre total de valeurs du jeu de données, (iii) l'actualité d'un jeu de données est définie par le rapport du nombre de valeurs obsolètes pour une date donnée sur le nombre total de valeurs du jeu de données, (iv) la cohérence d'un jeu de données est définie par le rapport du nombre de valeurs satisfaisant un ensemble de contraintes sur le nombre total de valeurs du jeu de données.

Actuellement, les techniques de mesure les plus utilisées sont les techniques d'échantillonnage [PC93] et les audits 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, édition des données, suivi de données, détection automatique des erreurs et traitement des exceptions et données isolées). 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).

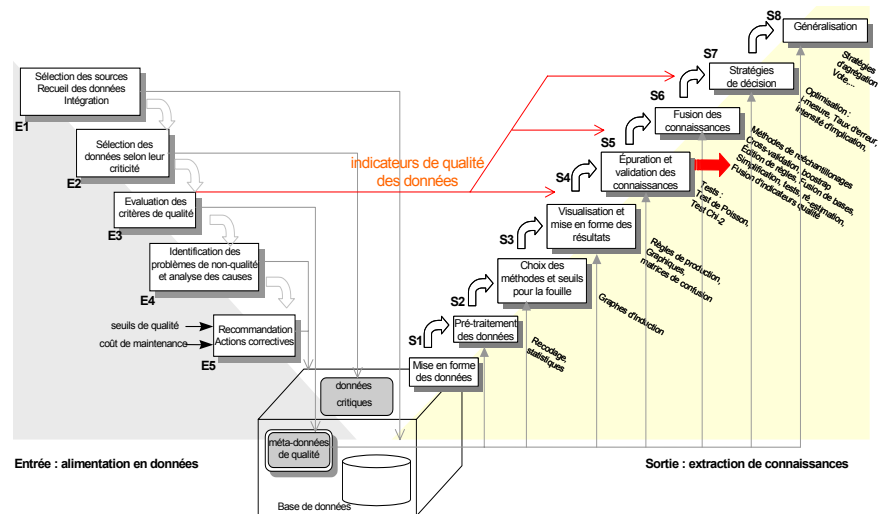
## 3. Méthodologie pour contrôler la qualité des données dans un processus d'ECD

Parce que, selon nous, la compréhension du processus d'alimentation et de qualification des données dans une base est indissociable de l'interprétation des résultats obtenus par ECD, nous proposons une méthodologie pour resituer la fusion

des indicateurs de qualité des données en amont du processus d'extraction de connaissances à partir des données (Figure 1).

### 3.1. 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.



**Figure 1.** Démarche générale pour contrôler la qualité des données pour l'ECD

L'étape **E3** consiste à calculer les scores de qualité des données 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. 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.

### 3.2. 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 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. Fusionner la qualité des données pour renseigner la qualité des règles

Notre hypothèse initiale est que la qualité d'une règle d'association dépend de la qualité des données sur lesquelles elle se base. L'objet de cette section est donc de proposer 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  $j$  ( $j=1, \dots, k$ ) 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 :  $Q(D_i)^T = [c_{i1}, \dots, c_{ik}]$

**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  $j=1, \dots, k$ , le poids affecté au score de chaque critère de qualité  $c_{ij}$  avec les deux hypothèses suivantes :

$$w_j \geq 0 \text{ et } \sum_{j=1}^k w_j = 1$$

Les poids  $w_1, \dots, w_k$  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  $j = 1, \dots, k$

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 [CCR78].

**Définition 3. Qualité d'une règle :** Soit une règle d'association  $R$  de la forme  $X \rightarrow Y$  dans laquelle la prémisse  $X$ , et la conclusion  $Y$  sont des conjonctions de variables telles que : l'extension de  $X$  est  $g(X) = x_1 \wedge x_2 \wedge \dots \wedge x_n$  et l'extension de  $Y$  est  $g(Y) = y_1 \wedge y_2 \wedge \dots \wedge y_n$ . 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 :

$$\begin{aligned} Quality(R) &= Quality(X) \circ Quality(Y) \\ &= Quality(x_1) \circ Quality(x_2) \circ \dots \circ Quality(x_n) \circ Quality(y_1) \circ Quality(y_2) \\ &\quad \circ \dots \circ Quality(y_n) \end{aligned}$$

**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  $\circ$  est une fonction commutative et associative de  $T \times T \rightarrow T$ . 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 1 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 4 critères de qualité précédemment définis.

| CRITÈRE DE QUALITÉ | FONCTION DE FUSION DE SCORES "o" | QUALITÉ DE LA RÈGLE $x \rightarrow y$                                                                                          |
|--------------------|----------------------------------|--------------------------------------------------------------------------------------------------------------------------------|
| Actualité          | $\min[c_x, c_y]$                 | L'actualité de la règle $x \rightarrow y$ est jugée de façon pessimiste égale au plus faible score d'actualité, donc à $c_y$ . |
| Exactitude         | $c_x \cdot c_y$                  | L'exactitude de la règle $x \rightarrow y$ est définie comme la probabilité d'exactitude des jeux de données $x$ et $y$ .      |
| Complétude         | $c_x + c_y - c_x \cdot c_y$      | La complétude de la règle $x \rightarrow y$ est définie comme la probabilité que l'un des deux jeux de données soit complet.   |
| Cohérence          | $\max[c_x, c_y]$                 | La cohérence de la règle $x \rightarrow y$ est jugée de façon optimiste égale au score de cohérence le plus élevé.             |

**Tableau 1.** Exemple de fonctions pour la fusion des scores de qualité des données

## 5. Conclusion

La plupart des bases de données produites actuellement sont entachées d'incertitude et contiennent des erreurs. Face à cette réalité et à la multiplication des échanges de

données, tous les acteurs impliqués dans l'utilisation et l'analyse des données prennent conscience de la nécessité de contrôler les données. Le contrôle de la qualité des données doit en premier lieu se dérouler lors des étapes de production ou d'acquisition des données et être pris en compte au final lors de l'interprétation et la validation des règles extraites par un processus d'ECD. L'article resitue dans une méthodologie générale le contrôle de la qualité des données sur la qualité des connaissances. Des méta-données relatives à la qualité des données sont utilisées et, en particulier, les scores des critères de qualité sont fusionnés afin de renseigner la qualité des règles qui les mettent en œuvre. Les mesures de qualité des données et les fonctions de fusion des scores pour évaluer la qualité des règles sont en cours d'intégration dans WEKA<sup>†</sup> (*Waikato Environment for Knowledge Analysis*).

## 5. Bibliographie

- [Bro90] Brodie M. L., Data quality in information systems, *Information and Management*, vol. 3, p. 245-258, 1980.
- [CCR78] Charnes, A., W. Cooper, and Rhodes E., Measuring the efficiency of decision making units. *European Journal of Operational Research*, vol. 2, p. 429-444, 1978.
- [CD+97] Clavanes D., DeGiacomo G., Lenzerini M., Nardi D., Rosati R., Data integration in datawarehousing. Technical Report DWQ-UNIROMA-001, DWQ Consortium, 1997.
- [DR92] Delen G., Rijsenbrij D., The specification, engineering and measurement of information systems quality, *Journal of Software Systems*, no.17, p. 205-217, 1992.
- [GJ98] Goodchild M., Jeansoulin R. (Ed.), *Data quality in geographic information : from error to uncertainty*, Hermès, 1998.
- [FLR94] Fox C., Levitin A., Redman T., The notion of data and its quality dimensions, *Information Processing and Management*, vol. 30, no. 1, 1994.
- [HY81] Hwang, C.-L. and Yoon K., *Multiple Attribute Decision Making*. no. 186, *Lecture Notes in Economics and Mathematical Systems*, Springer, 1981.
- [PC93] Parsaye K., Chignell M. Data quality control with smart databases, *AI Expert*, 8(5):23 – 27, 1993.
- [Red96] Redman T., *Data quality for the information age*, Artech House Publishers, 1996.
- [Saa80] Saaty, T. L., *The Analytic Hierarchy Process*. New York: McGraw-Hill, Inc, 1980.
- [TB98] Tayi G. K., Ballou D. P., Examining Data Quality, *Communications of the ACM*, vol. 41, no. 2, p. 54-57, 1998.
- [Vas00] Vassiliadis P., *Data Warehouse Modeling and Quality Issues*, PhD thesis, National Technical University of Athens, 2000.
- [WSF95] Wang R., Storey V., Firth C., A framework for analysis of data quality research, *IEEE Transactions on Knowledge and Data Engineering*, vol. 7, no. 4, p. 670-677, 1995.

---

<sup>†</sup> WEKA : <http://www.cs.waikato.ac.nz/ml/weka>