TECHN 7

Publication technique de la SmalS-MvM n°7 - le 30 janvier 1998

EVALUER ET AMELIORER LA QUALITE DES BASES DE DONNEES

Isabelle Boydens

est consultante à la section des recherches. Elle est responsable du projet "Examen et Amélioration de la qualité du LATG". Elle a entamé ses travaux en tant que chercheur aux universités de Liège et de Bruxelles. Elle s'est spécialisée dans l'analyse critique et la modélisation des bases de données administratives.

Contact : 02/509.59.91

1. Introduction

Nous sommes nombreux à nous souvenir des inconvénients quotidiens qui suivirent la crise pétrolière de 1973 : files devant les pompes à essence, interdiction de circuler le dimanche, ... Aux Etats-Unis, l'embargo pétrolier des Pays Arabes eut un impact particulier. Aux difficultés politiques que connaissait alors le gouvernement de Jimmy Carter s'ajouta une immense vague de suspicion sur la qualité des statistiques en matière d'énergie présentées par les autorités fédérales en vue de justifier la situation économique désastreuse. L'opinion publique américaine reprochait au gouvernement d'avoir été au mieux, naïf et au pire, en collusion avec les seuls producteurs d'information en la matière : les industries pétrolières elles-mêmes. Afin de mettre un terme à la polémique, le *Department of Energy* fut chargé de mettre en place l'un des plus vastes audits de bases de données de tous les temps : pendant près de 5 ans plus de 400 systèmes d'information et près de 2200 bases de données furent inspectés[1]. Des dizaines de sociétés de consultance furent mobilisées à cette fin. Vu l'ampleur de l'information à traiter, il était impossible d'analyser en profondeur chaque base de données et des procédures d'analyse statistique, héritées des méthodes d'analyse de la production industrielle, furent mises en place. Mais progressivement, il est apparu que le problème essentiel ne relevait pas d'une analyse traditionnelle de l'erreur mais d'une question d'interprétation. Le problème le plus significatif résidait en effet dans l'usage croisé de données dont la dénomination était identique mais dont la signification était distincte. Par exemple, certaines bases de données relatives aux *ventes d'énergie* étaient confondues et couplées avec d'autres bases de données relatives à *la consommation en énergie* ou encore, des données collectées dans le contexte des *inventaires* étaient exploitées dans d'autres systèmes en vue de mesurer la *distribution*. Considérées individuellement, les données étaient « correctes » mais les problèmes survenaient lorsque des informations créées dans des contextes différents étaient couplées.

Malgré ces premiers résultats, la question de l'interprétation des bases de données est loin d'être épuisée. Ce n'est que depuis le début des années nonante qu'un nouveau domaine de recherche, appelé « *data quality research* » s'est réellement développé. Il apparaît en effet que dans le monde des industries, des entreprises et des administrations, la qualité de l'information est l'un des enjeux financiers et compétitifs les plus importants. Après une brève définition de ce que l'on entend par « qualité des bases de données », nous présentons et évaluons trois techniques récentes destinées à analyser et à améliorer la qualité des données

[1] A. S. LOEBL, *Accuracy and Relevance and the Quality of Data* dans *Data Quality Control. Theory and Pragmatics*, G. E. LIEPINS et V. R. R. UPPULURI éds, (Série "*Statistics : Textbooks and Monographs*"), New York, Marcel Dekker, Inc, vol. 112, 1990, pp. 103-141.

informatisées. Cette étude s'inscrit par ailleurs dans le cadre plus large des méthodes d'amélioration du "*software process*"[2] dont elle envisage spécifiquement les questions liées à la qualité de l'information.

2. La qualité des bases de données : définitions

2.1. Base de données

Une base de données est une collection finie et structurée de données codifiées destinées à représenter certains aspects du réel observable, appelés « domaine d'application ». La structure d'une base de données (i.e., son « schéma ») inclut le domaine de définition des données. Leurs propriétés sont définies en termes de contraintes d'intégrité, assertions logiques exprimant les valeurs admises de chaque donnée et de leurs interrelations. Quelle qu'en soit la complexité apparente, la conception d'un schéma de base de données repose toujours sur un processus de transformation et de simplification du réel observable, processus qui en fonde le caractère opérationnel.

2.2. Qualité d'une base de données

Au sens envisagé ici, le terme « qualité » s'inscrit dans une échelle appréciative de valeurs pratiques, la qualité étant définie en termes de critères positifs et rejoignant le concept « d'excellence ». Une base de données est dite de « qualité » si elle permet effectivement de représenter ce pour quoi elle a été initialement conçue, c'est-à-dire, si elle satisfait les besoins des utilisateurs. Quatre caractéristiques découlent de ce qui précède :

- La notion de qualité est multidimensionnelle : elle peut par exemple concerner la précision, la cohérence, la consistance, la crédibilité ou encore l'actualité de l'information;
- Certaines dimensions sont mesurables (par exemple, la consistance logique d'une base de données) et d'autres ne le sont pas (par exemple, la crédibilité de l'information diffusée par une base de données);
- Certaines dimensions sont concurrentes : une information actuelle et récente ne sera pas nécessairement cohérente car les processus de test et de correction n'auront pas été menés à leur terme. Inversement, une information logiquement cohérente et consistante ne sera pas nécessairement actuelle et récente;
- La notion de qualité varie en fonction des usages : alors que l'exploitation administrative d'une base de données exige une précision maximale en vue de traiter équitablement chaque enregistrement, l'exploitation statistique de la base de données tolère une certaine marge d'erreur.

Ces quatre caractéristiques invalident le label « *qualité totale* » que l'on retrouve pourtant dans de nombreux travaux en matière de qualité informatique. La qualité de l'information, par définition, ne peut être « totale » : elle relève inévitablement d'un compromis entre plusieurs critères.

---

[2] M. DE DECKER, *Software Process Improvement* dans *Techno, Publication technique de la SmalS-MvM*, Bruxelles, n°6, novembre 1997.

2.3. Les coûts de la "non-qualité"

L'estimation des "coûts de la non-qualité" n'est pas aisée. Ajoutons que s'il est relativement aisé d'évaluer combien coûte la mise en oeuvre d'une procédure d'amélioration, les bénéfices escomptés sont plus difficiles à chiffrer en raison des aspects non mesurables, mais néanmoins cruciaux, qui accompagnent l'amélioration de la qualité d'un système informatique, tels que la crédibilité ou la fiabilité de l'information. A titre indicatif, plusieurs études menées aux Etats-Unis dans des secteurs divers (banques, assurances, agences de voyage...) font état d'un taux d'erreur de 5% à 30% dans les bases de données (ce taux étant, par exemple, évalué sur base du rapport entre le nombre d'enregistrements contenant au moins une erreur logique et le nombre total d'enregistrements d'une base de données). En termes financiers, les coûts de la "non-qualité" sont évalués à une perte d'environ 5 à 10 % du revenu des entreprises examinées (citons par exemple les coûts en contrôles, correction et maintenance de données de qualité douteuse, les coûts liés au traitement des plaintes des clients non satisfaits ou encore à la réparation des préjudices)[3].

3. Trois méthodes d'évaluation et d'amélioration de la qualité des bases de données

Les résultats concrets des travaux en matière de qualité des bases de données couvrent de nombreux aspects du cycle de vie d'une base de données. Nous avons sélectionné ici trois méthodes particulièrement représentatives et complémentaires. Chacune de ces méthodes mériterait de longs développements : nous en présentons une synthèse ainsi qu'une évaluation critique, tant sur le plan de leur efficacité que du rapport « coût/bénéfice » correspondant :

1. Le "Data Tracking" : assurer le suivi des données;
2. Le "Data Tagging" : étiqueter les données;
3. Le "Data Cleansing" : "nettoyer" les données.

[3] T. C. REDMAN, *Data Quality for the Information Age*, Boston-London, Artech House Publishers, 1996.

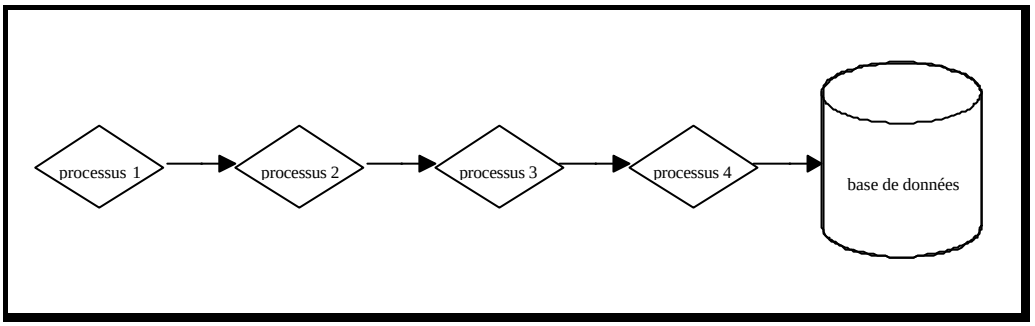
3.1. Le « data tracking *» : assurer le suivi des données*

Le « *data tracking* » (littéralement « suivi des données ») est une technique développée par les ingénieurs d'un des laboratoires de recherche du géant américain des télécommunications, *AT&T Laboratories*[4]. La méthode est destinée à évaluer les processus qui précèdent l'intégration de l'information dans une base de données. La technique repose sur le principe selon lequel de nombreuses erreurs naissent en amont de la base de données et comporte quatre phases dont les trois premières sont ici présentées sur base d'un exemple fictif.

3.1.1. Identification des processus

La première étape consiste à identifier les processus et programmes d'intégration de l'information (figure 1). Par exemple, ces processus peuvent permettre de transformer progressivement l'information entrante dans un format commun et homogène ou encore, de générer de nouvelles valeurs issues des programmes de test et de correction de l'information.

Figure 1. Identification des processus



3.1.2. Identification et évaluation des transformations opérées sur les enregistrements

L'étape suivante consiste à prélever une collection d'enregistrements destinés à être intégrés dans le système d'information (input) et à évaluer les transformations opérées d'un fichier intermédiaire à l'autre. Il s'agit ensuite, avec l'aide des gestionnaires de la base de données, de distinguer :

- d'une part, les transformations qui correspondent à des phases de normalisation prévues par les spécifications initiales (par exemple, insertion ou suppression d'espaces et de délimiteurs ou modification de formats, comme l'illustre la table 1 : la modification de l'attribut 1 par le processus 4),

[4] Voir par exemple : Y. U. HUH, F. R. KELLER, T. C. REDMAN et A. R. WATKINS, *Data Quality* dans *Information and Software Technology*, 32-8, 1990, pp. 559-565. T. C. REDMAN, *Data Quality for Telecommunications* dans *IEEE Journal on Selected Areas in Communications*, 12-2, 1994, pp. 306-312.

- d'autre part, les transformations indésirables qui génèrent des déformations de valeurs et témoignent d'une inadéquation entre spécifications attendues et programmation effective. Ces déformations n'émanent pas nécessairement d'une « erreur humaine » lors de la programmation, elles peuvent également être dues à la survenance, dans la collection de données examinées, de nouveaux cas de figure non prévus dans les spécifications initiales. Pour cette raison, leur analyse nécessite l'expertise des spécialistes du domaine d'application.

Table 1. Identification des transformations opérées sur un enregistrement

	Processus 1	**Processus 2**	**Processus 3**	**Processus 4**	**Base de données**
Attribut 1	XYZ1	XYZ1	XYZ1	**XYZ1-001**	**XYZ1-001**
Attribut 2	Oui	Oui	**Non**	Non	Non
Attribut 3			K	K	K
Attribut 4		1500	**5100**	5100	5100
Attribut 5		Z	Z	Z	**1**
Attribut 6					**OK**
Date	01/03/89	02/03/89	20/03/89	25/03/89	01/04/89

3.1.3. Quantification des changements, identification des erreurs et amélioration des processus

Il s'agit ensuite d'évaluer sur la collection d'enregistrements initialement sélectionnés le pourcentage d'erreurs parmi les changements décelés par attribut et par processus (table 2). Cette évaluation permet d'identifier les processus et programmes dont la structure doit faire l'objet d'une révision ultérieure.

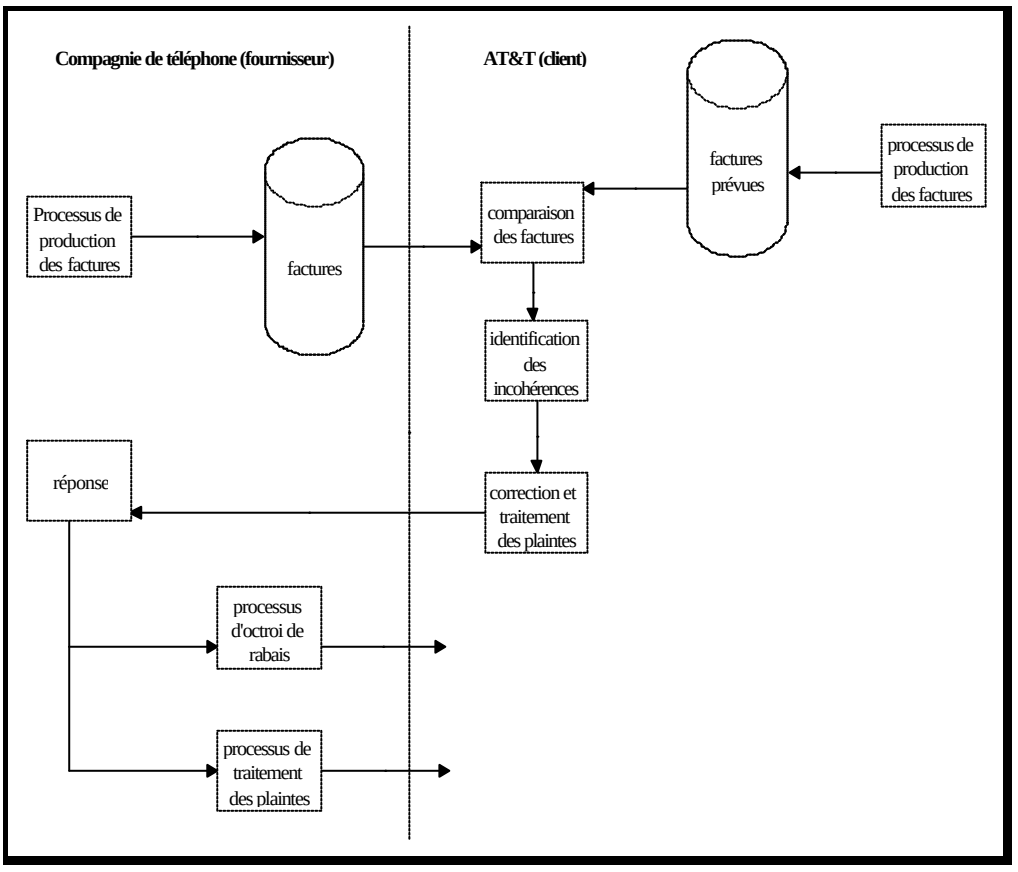
Table 2. Quantification des changements opérés sur une collection d'enregistrements

Attribut	**Processus**	**Changements**	**Enregistrements Lus**	**Pourcentage**
Attribut 2	Processus 2	3	100	3%
Attribut 2	Processus 3	21	94	22%
Attribut 2	Processus 4	2	91	2%
Attribut 3	Processus 4	1	91	1%
Attribut 4	Processus 3	4	94	4%
Attribut 4	Processus 4	7	91	8%

3.1.4. Rationalisation des processus et *"reengineering"*

Dans un cadre plus large, la technique du « *data tracking* » peut donner lieu à une révision complète et à une rationalisation des processus de traitement de l'information. L'objectif consiste à réduire le nombre de procédures précédant l'intégration de l'information dans la base de données. Le processus de facturation reliant *AT&T Laboratories* et ses fournisseurs fut revu dans cet esprit[5]. Initialement, la procédure faisait l'objet d'un traitement et d'une double vérification chez les fournisseurs d'une part et chez *AT&T Laboratories*, d'autre part (figure 2).

Figure 2. La vérification des factures à *AT&T Laboratories* avant restructuration

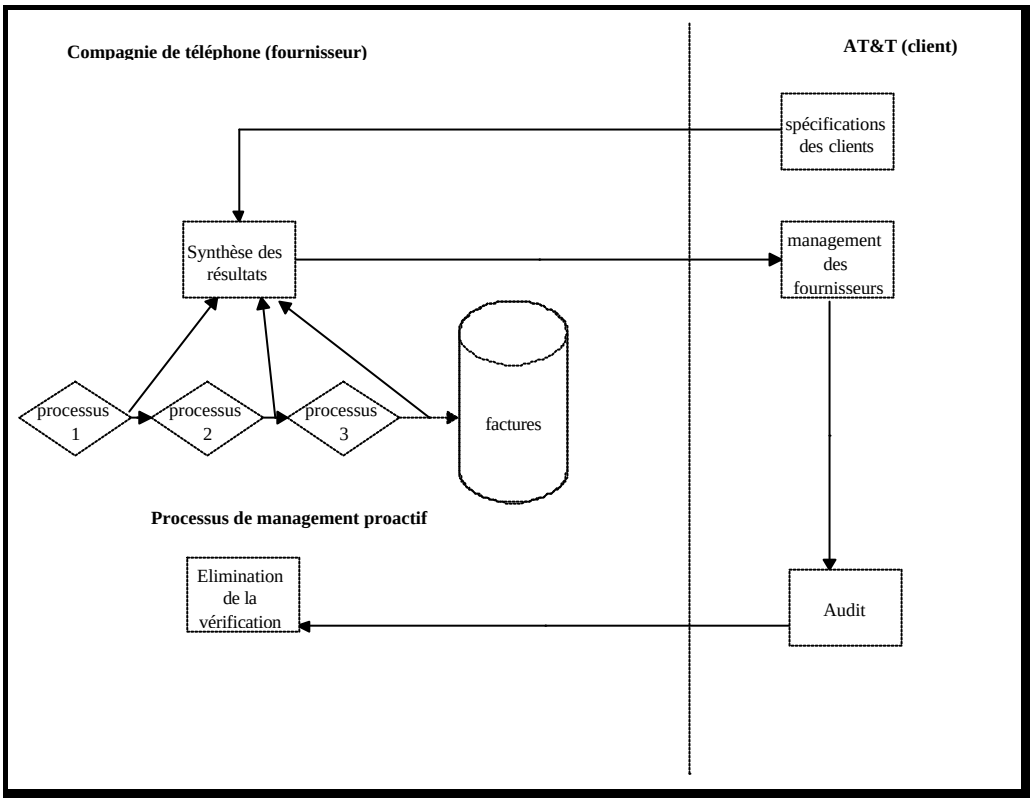


Ce processus donnait lieu à une duplication des bases de données (l'une concernant les montants facturés et l'autre, les factures «prévues ») et des traitements mais aussi à de multiples procédures judiciaires en cas de litige sur les montants à payer. En particulier, la duplication des données et des programmes multipliait d'une part, le risque d'erreur et d'incohérence et d'autre part, les coûts en traitements informatiques pour chaque partenaire. Pour cette raison, la structure de vérification fut entièrement revue dans le cadre d'un partenariat (figure 3) : le processus de vérification et de traitement étant centralisé chez les

[5] T. C. REDMAN, *Improve Data Quality for Competitive Advantage* dans *Sloan Management Review*, hiver 1995, pp. 99-106.

fournisseurs sous le contrôle *d'AT&T Laboratories*, via des procédures régulières d'inspection. Cette rationalisation a permis une amélioration de la qualité des données ainsi qu'une simplification des procédures informatiques : les gains financiers de l'opération furent évalués aux deux tiers du montant antérieurement consacré au traitement de l'information.

Figure 3. La vérification des factures à *AT&T Laboratories* après restructuration



3.1.5. Evaluation de la méthode

Les points forts du «*data tracking* » résident dans le fait que la base de données est envisagée dans son contexte, ce qui permet de localiser l'origine et la cause des éventuelles erreurs. L'approche est structurelle et permet une meilleure connaissance du système d'information considéré. Pour cette raison, les résultats obtenus par la révision finale des processus seront durables et le rapport en termes de « coûts/bénéfices » est positif. Cependant, seules les erreurs logiquement et formellement décelables sont analysées. Selon nous, la technique doit être élargie par une prise en considération d'une part, de la nature des processus et d'autre part, des problèmes d'interprétation non décelables formellement. C'est ce que nous envisageons au point suivant avec la technique du « *data tagging* ».

3.2 Le « data tagging » : étiqueter les données

3.2.1. Présentation de la méthode

La technique du « *data tracking* » permet de déceler les violations de contraintes d'intégrité logiquement identifiables. Mais il est apparu que l'usage exclusif des contraintes logiques ne suffisait pas en vue de garantir la qualité de l'information. Prenons un exemple simplifié sur base de la table 3. Par exemple, pour le trimestre T, le montant de cotisation à payer par les employeurs relevant de la catégorie d'activité A est égal à la somme des rémunérations déclarées R multipliée par le taux de cotisation C. Mais les tests formels et logiques ne permettent pas de savoir si le montant des rémunérations déclarées (donnée pour le moins cruciale...) est correct.

Table 3. Table des cotisations sociales à payer

Employeur - id	Rémunérations déclarées	Taux de cotisation	Cotisations sociales à payer
XO1.236	20.000	12.36%	2472
KO1.658	256.658	14.25%	36573,75
RD1.258	1.258.236	14.69%	184834,86

Le « *data tagging* » (littéralement « étiquetage des données ») est une méthode développée aux Etats-Unis par le *Massachusetts Institute of Technology*[6]. La méthode consiste à enrichir le schéma d'une base de données en y ajoutant des informations qui permettent aux utilisateurs d'en évaluer la qualité. L'approche comporte trois étapes :

- Identification des dimensions « subjectives » de la qualité jugées cruciales par les utilisateurs: par exemple, actualité ou encore, fiabilité de l'information.
- Identification des indicateurs « objectifs » de la qualité permettant de mesurer certains aspects des dimensions prédéfinies : par exemple, la fiabilité de la donnée « rémunérations déclarées » peut être mieux connue si, pour chaque occurrence de la relation, on connaît les informations suivantes :

- date d'intégration (et donc, de test) et de correction de l'information : on saura de la sorte quand l'information a été intégrée et en cas d'erreur, corrigée.

- catégorie d'activité de l'employeur : la qualité présumée d'une donnée peut être évaluée sur base du secteur d'activité dont émane l'information. En effet, dans certains secteurs, au sein desquels la main d'oeuvre est très fluctuante, les données envoyées à l'administration sont moins fiables et souvent sujettes à des rectifications et à des corrections.

- devise du montant : avec l'intégration probable de l'Euro, ce champ supplémentaire sera provisoirement indispensable en vue d'interpréter toute donnée monétaire.

- Intégration des indicateurs dans le schéma de la base de données, comme l'illustre la table 4, sur base de notre exemple simplifié :

---

6 Voir par exemple : R. Y. WANG, H. B. KON, and M. P. REDDY, *Toward Data Quality : An Attribute-Based Approach* dans *Decision Support Systems***,** Elsevier Science, 13, 1995, pp. 349-372. R. Y. WANG and M. P. REDDY, *Quality Data objects* dans *Total Data Quality Management (TDQM) Research Program Sloan School of Management, Massachusetts Institute of Technology*, Décembre 1992 TDQM-92-06.

Table 4. Indicateurs de la qualité relatifs à la donnée « rémunérations déclarées »

Employeur-id	Rémunérations déclarées	Date d'intégration	Date de correction	Devise	Secteur d'activité
XO1.236	20.000	01/05/1997	05/10/1997	Euro	Construction
KO1.658	256.658	02/10/1997	non corrigé	BEF	Intérim
RD1.258	1.258.236	01/01/1997	pas d'erreur	BEF	Assurances

3.2.2. Evaluation de la méthode

La méthode du « *data tagging* » rejoint le domaine des systèmes de méta-information[7]. Elle représente une aide à l'interprétation de l'information au-delà des tests de contrôle logique. Cependant, l'ajout de nouvelles données dans des bases de données déjà très volumineuses risque dans certains cas de poser des problèmes de performance. Par ailleurs, comment tester la qualité des indicateurs de la qualité ? Afin d'éviter ces écueils, il est vivement recommandé de privilégier comme indicateurs de la qualité:

- des informations déjà présentes dans la base de données : souvent les données s'éclairent mutuellement et leur mise en relation constitue un moyen économique d'en connaître la qualité,
- des données directement générées à partir du système comme, dans notre exemple, les dates d'intégration et de correction de l'information.

Par conséquent, la sélection et l'identification des indicateurs de qualité relèvent d'une analyse approfondie du système d'information étudié et requièrent à la fois une profonde connaissance du domaine d'application et de la structure des bases de données concernées.

3.3. Le « data cleansing » : « nettoyer » les données

3.3.1. Présentation de la méthode

Le « *data cleansing* » (littéralement « nettoyage » des données) est une méthode directement issue des techniques du *Datawarehousing*[8]. Ces dernières procèdent en effet à l'intégration de bases de données hétérogènes en vue d'obtenir une information agrégée et un support homogène à la prise de décision. Dans ce cadre, une opération préalable de mise en cohérence de l'information est indispensable. Plusieurs logiciels sont apparus sur le marché en vue d'automatiser cette opération de nettoyage[9] qui consiste par exemple :

- à remplacer systématiquement l'intitulé d'un champ par un autre (par exemple le champ « genre » par le champ « sexe ») en vue d'assurer l'homogénéité de l'information,
- à vérifier la cohérence logique des données (sur base de spécifications prédéfinies) et à corriger automatiquement les incohérences décelées.

---

[7] I. BOYDENS, *Les systèmes de méta-information* dans *Techno, Publication technique de la SmalS-MvM*, Bruxelles, n°1, avril 1997.

[8] A. DE KONING, *Le Datawarehousing* dans *Techno, Publication technique de la SmalS-MvM*, Bruxelles, n°2, mai 1997.

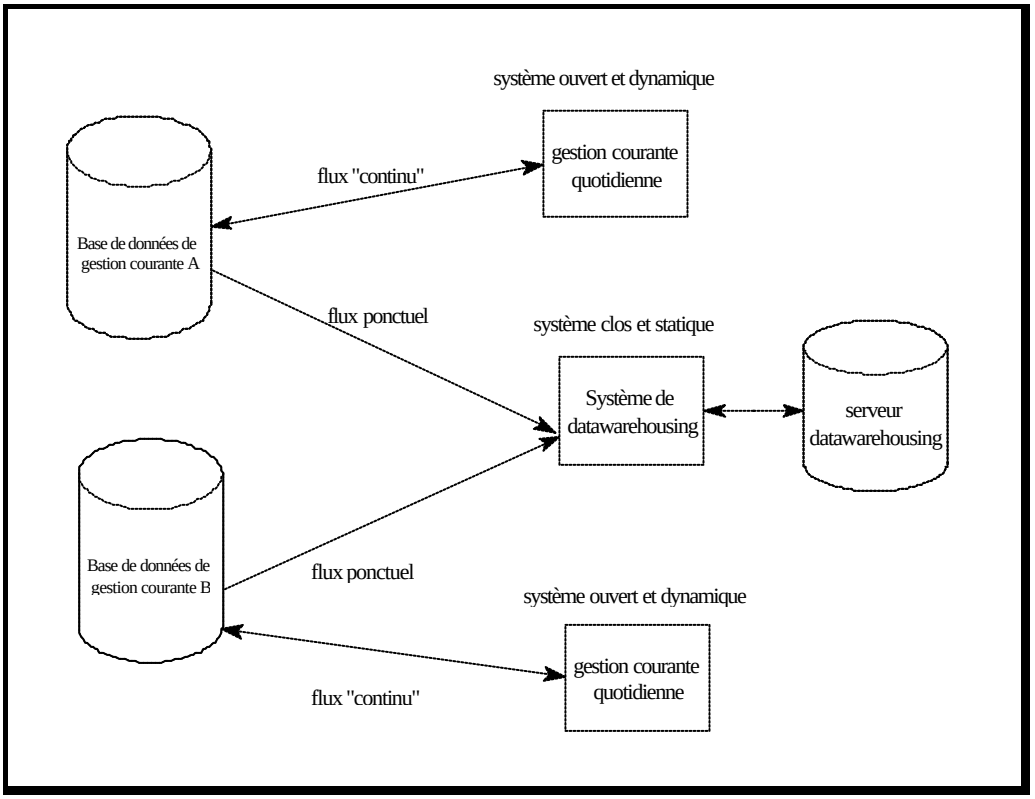
[9] Voir par exemple : M. HURWICZ, *Take your Data to the Cleaners* dans *Byte*, janvier 1997, pp.97-102.

3.3.2. Evaluation de la méthode

Le « *data cleansing* » se justifie entièrement dans le cadre d'une opération de « *Datawarehousing* » qui consiste à extraire ponctuellement et à agréger des informations issues de bases de données hétérogènes à des fins pratiques (voir figure 4). Mais il va de soi que les avantages d'une telle démarche résident moins dans la précision individuelle de chaque enregistrement « nettoyé » que dans la rationalisation massive de vastes collections de données. Dès lors, dans le cadre de l'analyse et de l'amélioration de la qualité des bases de données de gestion quotidienne, l'usage exclusif des techniques du « *data cleansing* » ne nous semble ni opportun, ni rentable pour les raisons suivantes :

- la technique agit exclusivement sur les données et ne considère ni le schéma de la base de données ni les processus qui l'encadrent : elle ne permet donc pas de connaître la cause et l'origine des problèmes de qualité;
- dans le cadre d'une base de données de gestion, traitant un flux d'information continu, la mise en oeuvre ponctuelle d'un processus de correction automatisé n'est ni efficace ni rentable : à peine corrigées, les données sont vite remplacées par d'autres données dont la qualité reste douteuse;
- lorsqu'ils assurent une correction automatique de l'information, les logiciels de «*data cleansing* » introduisent une cohérence parfois artificielle dans une collection de données, induisant de la sorte une perte d'information. En effet, dans le cadre d'une base de données de gestion, il est souvent utile de garder une trace du processus de correction de l'information erronée, a fortiori dans le domaine administratif où la force probante de l'information doit être respectée.

Figure 4. Bases de données de gestion et Datawarehousing



4. Conclusions : synthèse et perspectives

Après avoir brièvement défini en quoi consistait la qualité d'une base de données et quels en étaient les enjeux, nous avons présenté et évalué trois méthodes actuelles d'amélioration de la qualité des données. L'évaluation repose à la fois sur l'efficacité des techniques et sur leurs rapports en termes de « coûts/bénéfices » :

- le « *data tracking* » repose sur une analyse des processus et permet de déceler l'origine et les causes structurelles des erreurs logiques affectant l'information : la méthode assure dès lors une amélioration de la qualité rentable à long terme;
- le « *data tagging* » permet l'interprétation des dimensions de la qualité dont l'évaluation échappe aux contraintes de la logique formelle. En ce sens, la méthode pourrait apporter une réponse partielle aux problèmes d'interprétation évoqués dans l'introduction à propos des bases de données pétrolières. La mise en oeuvre efficace de cette approche repose sur le choix minutieux et économique des données qui feront office d'indicateurs de la qualité;
- enfin, le « *data cleansing* » se justifie essentiellement dans le cadre de systèmes d'information agrégés destinés à la prise de décision au sein d'une entreprise ou d'une administration. Appliqué aux bases de données de gestion courante, l'usage exclusif des logiciels de « *data cleansing* » n'est toutefois ni rentable ni efficace.

A cela, il convient d'ajouter que :

- toute méthode d'analyse et d'amélioration des bases de données constitue nécessairement une approche pluridisciplinaire : elle requiert à la fois une connaissance approfondie du domaine d'application concerné et une expertise des méthodes de conception des systèmes d'information,
- l'approche implique une étroite collaboration entre gestionnaires et utilisateurs des bases de données et requiert une solide infrastructure organisationnelle.

Enfin, il est fondamental de distinguer d'une part, les aspects méthodologiques généralisables à toutes les bases de données et d'autre part, les caractéristiques propres à certains domaines d'application. Dans le cadre du projet « examen et amélioration de la qualité du LATG », nous travaillons à la mise au point d'une méthode spécifiquement adaptée aux bases de données administratives.