Jeune Chercheur

Qualité de données multi-sources et recommandation multi-critère

Laure Berti
GECT, Equipe Systèmes d'Information Multi-Média
Université de Toulon et du Var
B.P. 132, F-83957 La Garde cedex, FRANCE
berti@univ-tln.fr, tél. : (33) 04-94-14-22-20, fax : (33) 04-94-14-20-75

Résumé

Nos travaux s'inscrivent dans un thème de recherche relativement émergent : la qualité des données dans les systèmes d'information. Cet article propose une méthodologie générale de mise en oeuvre de la qualité des données dans un contexte où les sources d'information sont multiples et les valeurs qu'elles proposent, souvent contradictoires. La qualité d'une donnée se décompose en différentes catégories de critères dont les évaluations permettent la détection des erreurs et également, la détection des données de qualité médiocre ou inadaptée à l'utilisateur. Nous introduisons la notion de qualité relative des données multi-sources dans les cas où des données en correspondance, issues de différentes sources, désignent le même signifié par des valeurs contradictoires. Notre approche diffère de celle généralement retenue en intégration de données pour concilier les conflits extensionnels : nous cumulons les informations et méta-informations de qualité et recommandons dynamiquement celles de meilleure qualité, c'est-à-dire, celles les plus appropriées aux besoins et préférences de l'utilisateur. Des algorithmes pour la recommandation des données et de leur source selon leur qualité relative sont introduits et appliqués au domaine spécifique de la veille technologique.

Mots-clés

Qualité de données, démarche qualité, recommandation multi-critère, systèmes d'information spécifiques et application de veille technologique

---

Abstract

Due to its costly impact, data quality is becoming an emerging domain of research. Motivated by its stakes and issues, especially in the application domain of Technological Watch, we propose a generic methodology for modeling and managing data quality in the context of multiple information sources. Data quality has different categories of quality criteria and their evaluation enables the detection of errors and poor data quality. We introduce the notion of relative data quality when several data capture the same reality with contradictory values. Our approach differs from the general approaches for resolving extensional inconsistencies in integration of heterogeneous systems : we cumulatively store data and quality meta-data and we recommend dynamically data with the best quality, or at least, data which are the most appropriate for a specific user. Recommendation algorithms are proposed and applied to the Technological Watch application domain.

Key-words

Data Quality, Data Quality Program, Multicriteria Recommendation, Technological Watch Application and specific Information Systems

1. Introduction

Accentuée par le phénomène Internet, l'incoercibilité de l'information ne cesse de s'accroître avec les développements des technologies de diffusion. De gigantesques gisements de données hétérogènes sont désormais accessibles sous des délais de plus en plus courts. La diversité et le volume sans cesse croissant de ces flux de données rendent économiquement et humainement difficiles les recherches et les traitements plus approfondis de l'information strictement utile. L'utilisateur est instantanément submergé par un flot de données qui, soit ne répondent pas précisément à ses besoins, soit, bien que pertinentes, sont incomplètes, imprécises et/ou contradictoires. De ces différents constats, on s'aperçoit que l'enjeu d'une recherche d'information très ciblée est déplacé du quantitatif vers le qualitatif, du volume de données à leur qualité. De plus, la validation et l'authentification de l'information demeurent délicates. Ce contexte requiert, non seulement, une attitude beaucoup plus active et critique de la part de l'utilisateur, « consommateur de produits d'information », mais aussi, de sérieuses facultés d'interprétation, d'évaluation, de déduction, d'analyse et de synthèse. Si les outils informatiques sont mis à profit pour véhiculer, stocker, gérer et analyser quantitativement une multitude d'informations, rares sont ceux qui assurent ou assistent une analyse critique et qualitative de l'information manipulée.

En amont d'un système d'information ou d'une base de données (lors de la phase d'alimentation en données), cette analyse critique est bien souvent court-circuitée par la facilité d'import et de manipulation des données. En aval, comme si le simple acte de leur saisie faisait foi, les données ne sont a priori que rarement remises en cause, ou qualifiées selon leur fiabilité ou l'intérêt qu'elles pourraient susciter. Il s'avère pourtant essentiel d'évaluer explicitement la qualité les données stockées dans les systèmes d'information de manière à :

1) orienter leur diffusion en fonction du public ciblé,
2) proposer à chaque utilisateur une expertise critique de la qualité du contenu du système,
3) permettre à l'utilisateur de relativiser la confiance qu'il pourrait accorder aux données et, lui permettre ainsi, de mieux en adapter leur usage.

On constate cependant que la qualité des données n'est qu'un thème de recherche relativement émergent dans divers domaines d'application industrielle [Wan96] [Red96], commerciale [SK97], militaire ou scientifique [CP98]. Jusqu'à présent, l'approche généralement adoptée pour mesurer la qualité des données est essentiellement statistique (par échantillonnage sur d'importants volumes de données). Mais, au sein d'une profusion d'informations souvent contradictoires, la qualité d'une seule donnée peut également s'évaluer par comparaison, c'est-à-dire en la comparant à la qualité des données qui sont issues de différentes sources, concernent le même signifié et qui possèdent des valeurs plus ou moins contradictoires. Notre démarche se place donc dans la perpective suivante : plutôt que de résoudre les conflits existants entre les valeurs des données, nous exploitons ces conflits pour évaluer la qualité relative des données les unes par rapport aux autres.

Lors de la consultation, le rôle du système est finalement celui d'un assistant qui propose une sélection adaptative des données « de meilleure qualité ». Il s'agit, en fait, d'une recommandation des données en fonction de leur qualité relative et des besoins de l'utilisateur.

L'objectif de ce travail est le suivant : selon le profil de l'utilisateur ou selon ses exigences explicites (en termes de qualité de données), le système ne présentera, en résultat de requête, que les données de qualité la plus adaptée parmi les données candidates issues des différentes

sources. La recommandation sera justifiée par la trace de l'algorithme de recommandation utilisé (l'utilisateur pouvant choisir différentes stratégies de recommandation).

La suite de l'article s'organise de la manière suivante : la section 2 présente une analyse des travaux sur le thème de la qualité des données, ainsi que notre proposition sur la qualité relative des données dans une perspective multi-source.

La section 3 propose une méthodologie générale de mise en oeuvre de la qualité des données dans le contexte multi-source. La section 4 définit des algorithmes de recommandation multi-critère. La section 5 illustre un exemple de leur application dans le domaine de la veille technologique. La section 6 conclut sur les perspectives de recherche et de développement issues de ce travail.

2. Qualité des données : la perspective multi-source

La qualité des données est un domaine de recherche qui a suscité depuis longtemps un vif intérêt, mais, qui émerge tout juste comme champ de recherche à part entière, tel que peuvent l'indiquer [WSF95] [Wan96] [SK97] [CP98] [TB98]. Avec le nombre croissant des sources d'information disponibles sur le réseau, le problème de la qualité de leur contenu est devenu crucial. De nombreux exemples révèlent des situations alarmantes concernant la qualité des données dans les bases de données commerciales [Jac92], médicales [SK97], judiciaires [Lau86], du domaine publique [Bas95], ou de l'industrie [Red96]. Les problèmes de non-qualité de données peuvent avoir diverses origines : par exemple,

- lors du développement du système, lorsque les besoins n'ont pas été complètement spécifiés,
- lorsque la méthode de saisie des données n'a pas été suffisamment bien conçue, que les erreurs humaines sont facilement introduites,
- lorsque les entreprises n'ont pas les moyens de mettre à jour ou d'enrichir leurs données,
- lorsque la sécurité physique et/ou logique du système laisse à désirer et que les données peuvent être corrompues volontairement...
- ... au final, lorsque l'utilisateur n'est pas satisfait.

Ces raisons motivent d'autant plus la prise en compte de la qualité des données comme partie intégrante dans la conception, le développement et la maintenance du système, ainsi que dans toute la chaîne de traitement de l'information [Red96], comme les exemples cités ci-dessus l'ont illustré.

2.1. Principaux travaux sur la qualité des données

2.1.1. Modélisation de la qualité des données

Les plus anciens travaux sur la modélisation de la qualité des données ont été menés par [Bro80]. Dans [BP85], les auteurs ont, les premiers, défini la notion de qualité des données et 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] et le consensus sur la définition de la qualité des données n'est toujours pas atteint (voir Tableau 1). Une analyse très complète [WSF95] a recensé les recherches menées sur la qualité des données et en présente un vaste panorama (avec plus de 120 références). Certains travaux intègrent totalement la modélisation et la gestion de la qualité des données dès la conception d'un système d'information. Parmi ces approches, le programme TDQM (*Total Data Quality Management*) [Wan98] [WKM93] propose une méthodologie de modélisation qui guide

pas à pas l'ajout de la dimension qualité sur chaque élément de l'application (entités, attributs, associations). Cette approche, jugée irréalisable en pratique car trop coûteuse, impose l'ajout de labels sur toutes les entités, relations et attributs du modèle.

L'utilisation de méta-données pour l'amélioration de la qualité des données a néanmoins été défendue et préconisée dans [Rot96]. L'auteur incite les producteurs d'informations à assurer la vérification, la validation et la certification de leurs données (VV&C : *Verification, Validation and Certification*) : les méta-données concernant la qualité des données devant être fournies de paire avec les données pour permettre l'estimation et la maintenance de la qualité des données. Dans le domaine des Systèmes d'Information Géographique, la plupart des standards d'échange (ISO 15046-13, CEN prEN 287-008, EDIGéO), incluent d'ailleurs les spécifications de méta-données de qualité [GJ98].

Auteurs et références	***Dichotomie et caractérisation de la qualité des données***	
Brodie [Bro80]	6 concepts	*Intégrité* *Abstraction* *Expressivité sémantique* *Validité* *Maintenance* *Efficacité dans l'utilisation des ressources*
Delen, Rijsenbrij [DR92]	4 dimensions 21 aspects 40 attributs	*Développement et contrôle du SI* *Propriétés statiques de maintenance* *Fonctionnement dynamique* *Importance de l'information : donnée correcte, complète, mise à jour, précise, vérifiable*
Wang et al. [Wan98] [WKM93] Programme TDQM	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 [CPPR98]	3 critères pour la qualité de l'information	*Disponibilité* *Fiabilité* *Adaptabilité*

Tableau 1. Quelques-unes des propositions pour caractériser la qualité des données

2.1.2. Evaluation de la qualité des données

Les premières approches adoptées pour mesurer la qualité des données furent des approches statisticiennes centrées sur des méthodes telles que l'inférence sur les données manquantes à partir des modèles statistiques, la détection automatique et le traitement des exceptions et des données isolées [LU90]. De nombreuses méthodes ont été développées pour mesurer la qualité des données fournies aux utilisateurs conformément à leurs spécifications de qualité, notamment dans le domaine des Systèmes d'Information Géographique [GJ98]. Les principales mesures de qualité de données considérées comme essentielles (quelles que soient les spécificités du domaine d'application) sont au nombre de quatre : l'exactitude, la complétude, la fraîcheur et la cohérence [Red96] [FLR94] [RW95] [SK97] :

- l'exactitude se mesure en détectant le taux de valeurs incorrectes dans la base de données
- la complétude se mesure en détectant le taux de valeurs manquantes dans la base de données
- la fraîcheur se mesure en détectant le taux de valeurs obsolètes dans la base de données
- la cohérence se mesure par rapport à un ensemble de contraintes en détectant les données de la base qui ne les satisfont pas.

Actuellement, les audits [PC93] [WSF95] sont les seuls moyens pratiques de déterminer la qualité des données d'une base (recensement des différents types d'erreurs, élaboration de méthodes pour les détecter, estimation de leur fréquence d'occurrence dans la base par des techniques statistiques appropriées).

Mais, les audits ne résolvent pas les problèmes. La tendance générale est à l'utilisation de méthodes de l'Intelligence Artificielle (validation et correction des données [PC93], identification des exceptions [Sch91], apprentissage, schémas de représentation des connaissances, gestion de l'incertain... [FPSSU96]).

En synthèse, les différents travaux sur la qualité des données peuvent être classés en trois grands courants selon leur objectif qui est : 1) soit de définir rigoureusement chaque dimension de la qualité des données ainsi que leur mode de mesure sur la base d'une méthode scientifique, 2) soit de créer un ensemble standard universel de dimensions opérationnelles pour la qualité des données, 3) soit de proposer une assistance à l'utilisateur pour qu'il définisse et évalue lui-même la qualité des données qu'il manipule.

2.3. Qualité de données relative

Parallèlement aux approches statisticiennes et fréquencistes pour mesurer la qualité des données, quelques approches subjectivistes, plus orientées vers l'assistance aux utilisateurs émergent. Nous nous positionnons dans ce courant de recherche et étendons la proposition de [SLW97] [Wan98] (voir partie grisée dans le Tableau 1) en décomposant la qualité d'une donnée en 4 catégories de critères [Ber98] (voir Tableau 2). La première catégorie regroupe les critères indiquant la qualité de la gestion de la donnée par le système (accessibilité, confidentialité, sécurité,...). La seconde catégorie regroupe les critères indiquant la qualité de la représentation de la donnée par le système (ou qualité de la modélisation). La troisième catégorie regroupe les critères intrinsèques de qualité de la donnée (absence d'erreur, exactitude, qualité de la source...).

Qualité d'une donnée		
Catégorie	***Sous-Catégorie***	***Critère***
Qualité de la gestion de la donnée par le système		Accessibilité Confidentialité Facilité d'échange Facilité de maintenance Facilité de manipulation Facilité de recherche Sécurité...
Qualité de la représentation de la donnée par le système		Compréhension Concision Interprétation ...
Qualité intrinsèque de la donnée		Absence d'erreur Cohérence Exactitude Qualité de la source...
Qualité relative de la donnée	**Qualité contextuelle**	***Référentiel temporel*** ***discret*** Actualité Fraîcheur... ***continu*** Traçabilité Variabilité Volatilité... ***Référentiel applicatif*** Criticité Pertinence Utilité...
	Qualité relative à l'utilisateur	***Référentiel cognitif*** Fiabilité Originalité Rareté Vraisemblance... ***Référentiel affectif*** Intérêt Préférences...
	Qualité relative aux données homologues	Comparaison Contradictions Redondance Similarité...

Tableau 2. Les catégories de critères de qualité d'une donnée et leur liste non exhaustive

Enfin, nous proposons la catégorie de la ***qualité relative des données*** qui comprend trois sous-catégories de critères : l'une relative au contexte de l'application (basée sur le référentiel temporel ou purement applicatif), la seconde sous-catégorie relative à l'utilisateur (basée sur le référentiel cognitif ou affectif), et enfin, la dernière sous-catégorie de critères évalués par une comparaison avec des *données homologues* (ou données en correspondance), c'est-à-dire des données issues de différentes sources et représentant la même réalité avec des valeurs plus ou moins contradictoires. Pour proposer une recommandation des données homologues basée sur leur qualité relative, nous nous intéressons aux critères de cette dernière sous-catégorie (grisée dans le Tableau 2) en introduisant la perspective multi-source dans l'évaluation de la qualité des données. Les travaux menés en intégration de systèmes hétérogènes, en particulier, sur les problèmes d'incohérences extensionnelles (problèmes d'identification des tuples, conflits de valeurs, contradictions) rejoignent notre problématique.

Si ces problèmes sont unanimement reconnus [KS92] [PAGM96], relativement peu de solutions ont été proposées pour résoudre les conflits entre les instances de données [Sad91] [LSS94]. Les problèmes d'identification d'entité et de valeurs de données incohérentes sont évoqués dans la littérature selon trois approches utilisant : soit des informations sémantiques du domaine [SM91] [UW91], soit la théorie des probabilités [CP87] [DeM89] [TCY93], soit la théorie de Dempster-Shafer [Lee92].

Les approches retenues pour concilier les hétérogénéités entre les valeurs de données sont habituellement de :

- préférer les valeurs des sources les plus fiables [KS92] [PAGM96],
- mentionner, pour chaque valeur, l'identifiant de sa source d'origine et se baser sur la crédibilité de cette source [WM90],
- stocker avec la donnée, des méta-données [SM91], notamment les mesures de fiabilité de sa source [Sad91].

Les principaux inconvénients de ces approches sont les suivants :
- il ne semble pas évident de déterminer quelle est la source la plus fiable, ni de définir la méthode pour mesurer la fiabilité d'une source,
- même si la fiabilité des sources est fournie, il faut alors supposer que la fiabilité reste homogène pour toutes les données que propose la source, ce qui est rarement le cas en pratique,
- le stockage de la donnée, de sa fiabilité ou de celle de sa source augmente de façon prohibitive les besoins de stockage et nécessite des modifications importantes dans le traitement classique des requêtes.

Dans la suite de l'article, nous retiendrons les mêmes approches que celles adoptées pour concilier les conflits entre valeurs, en proposant des solutions à certains de leurs inconvénients. Plutôt que de concilier les conflits, nous les exploitons pour évaluer la qualité relative des données et proposons une méthodologie pour sa mise en oeuvre.

3. Méthodologie pour la mise en oeuvre de la qualité des données

Dans un contexte multi-source, l'objectif de notre système d'information est de recommander les données les plus appropriées aux besoins d'un utilisateur. Ainsi, pour la même requête, différentes données seront proposées aux différents utilisateurs requérants, selon leurs exigences en termes de qualité de données. Par exemple, un comptable privilégiera la précision des données qu'il manipule pour ses conversions en euros, alors que le chef d'entreprise n'aura pas besoin d'une telle précision pour son chiffre d'affaire. En conséquence, il s'agit principalement de :
- cumuler toutes les données homologues (la dernière donnée collectée n'étant pas toujours la meilleure),
- stocker toutes les données critiques dans une base où, pour une même instance, les attributs peuvent avoir des valeurs multiples,
- exploiter les conflits entre les valeurs des différentes sources,
- évaluer la qualité relative des données.

Comme toute autre démarche de qualité [Red96] [Wan98] [SK97] [CP98] [Bon96], nous définissons les étapes essentielles pour mettre en oeuvre la qualité relative des données dans un environnement multi-source. La démarche repose sur les étapes suivantes (Figure 1) :

1- spécifications préalables des besoins
2- sélection des critères de qualité pour les sources et leurs données
3- sélection des sources d'information, recueil des données et mise en correspondance
4- sélection des données critiques
5- évaluation des critères de qualité
6- identification des problèmes de non-qualité et analyse de leurs causes
7- recommandation des données et définition de nouveaux objectifs de qualité.

Etape 1. Spécifications préalables des besoins

Cette étape est une phase préalable à l'étude et à la mise en oeuvre de la qualité des données. Les besoins doivent être clairement exprimés, en particulier, sur la qualité des données attendue par

les différents utilisateurs (opérateurs de saisie, clients consultants, Direction et décideur final). Nécessitant une réelle implication de ces acteurs, elle a pour but de préciser le vocabulaire du domaine d'application, ainsi que le vocabulaire sur les dimensions (critères) de la qualité des données, et ce, afin de s'assurer que chacun utilise les mêmes concepts avec la même signification. Elle permettra de construire une ontologie du domaine des informations et des méta-informations de qualité. Elle fixe les limites de l'étude et ses priorités (par exemple, évaluer la qualité des données manipulées par un traitement particulier, ou bien, améliorer la qualité des données numériques pour telle classe d'objets...).

Etape 2. Sélection des critères de qualité pour les sources et leurs données

Directement dépendante de la première étape, cette étape a pour but de choisir les critères sur lesquels s'évalue la qualité des sources et celle de leurs données. Les critères de qualité d'une source peuvent être de deux types :

- des critères objectifs (comptage du nombre de renseignements fournis par la source indiquant sa largeur descriptive, comptage des caractéristiques renseignées indiquant sa profondeur descriptive,...)
- et des critères subjectifs (réputation des auteurs, crédibilité,...).

Les critères de qualité des données doivent répondre au mieux aux priorités définies à l'étape précédente. Ils peuvent être choisis selon les différentes catégories du Tableau 2, parmi la liste des critères proposée par [WKM93] ou bien encore selon l’une des propositions du Tableau 1.

Etape 3. Sélection des sources, recueil des données et mise en correspondance

Le but de cette étape est de sélectionner les sources d'information qui vont alimenter en données la base, puis de recueillir les données proposées par ces différentes sources et enfin, de mettre en correspondance les données qui désignent la même réalité. Les sources sont sélectionnées selon leur qualité évaluée sur les critères choisis à l'étape précédente.

Pour renseigner les attributs de la base multi-source, le recueil des données peut être fait de façon automatique (par import) ou manuelle (par un opérateur de saisie). Le flux des données doit être précisément décrit : les acteurs et leurs interventions (création, mise à jour, utilisation, destruction) qui peuvent modifier la qualité des données doivent être identifiés. Cette étape met en évidence des données non renseignées ou, bien renseignées mais non utilisées. Elle permet une remise en cause de leur existence et la formalisation des règles de décision, ou bien encore la distinction de différents niveaux de qualité pour les attributs selon leur utilisation particulière (Figure 1). La mise en correspondance de telles données sera supervisée et faite manuellement par un opérateur, expert du domaine. Les données issues de différentes sources peuvent avoir des valeurs contradictoires ou recouvrantes. Des attributs à valeurs multiples ayant la mention de l'identifiant de leur source d'origine en seront le résultat de stockage au sein de la ***base de données multi-sources*** (voir Figure 2).

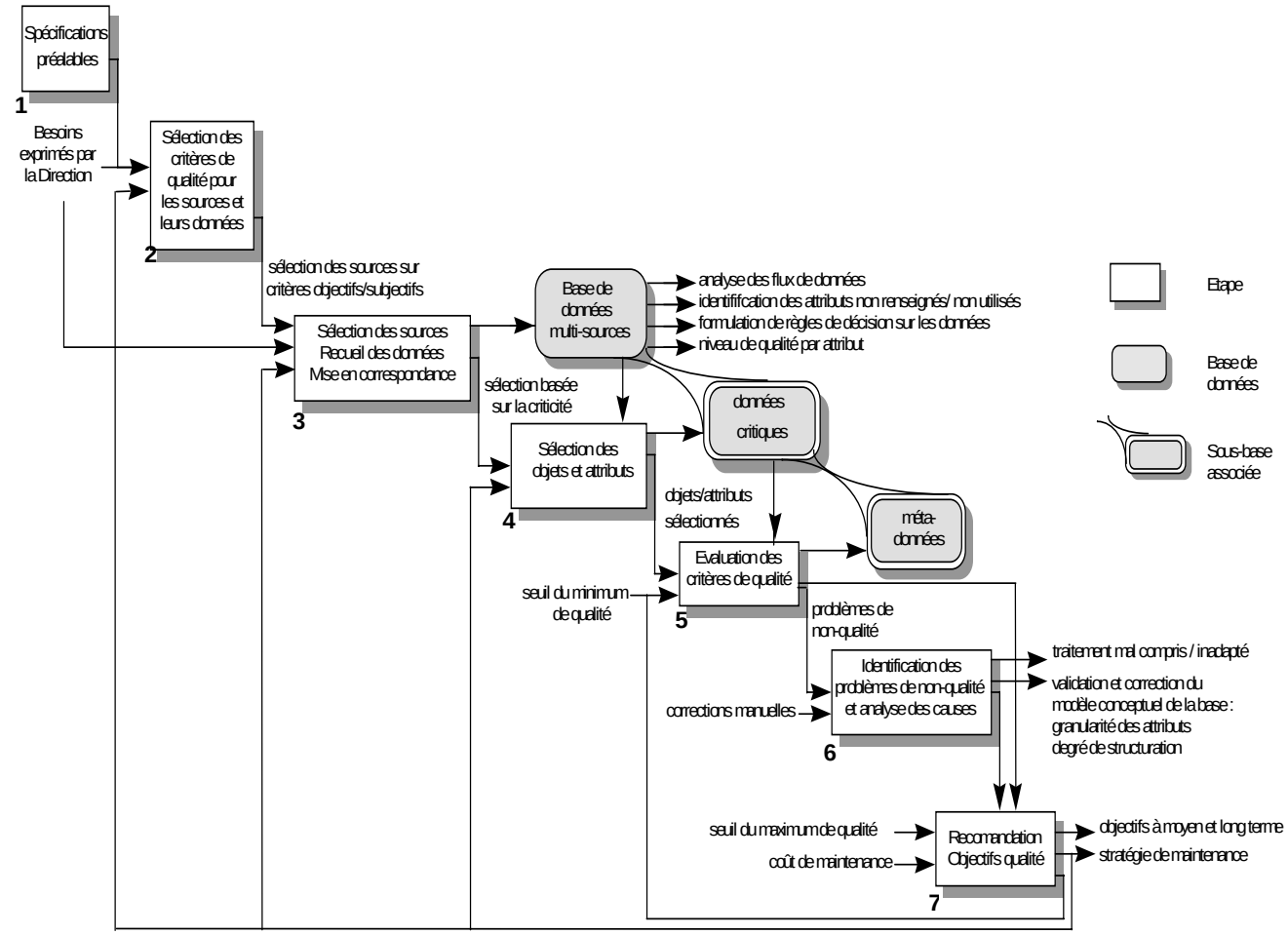

Figure 1. Démarche générale pour la mise en oeuvre de la qualité des données multi-sources

Etape 4. Sélection des données critiques

Le but de cette étape est de sélectionner les classes, les objets et les attributs *critiques* dans la base de données multi-sources. Toutes les données n'ont pas la même importance, elles ne sont pas équivalentes d'un point de vue objectif pour l'entreprise : elles 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.

Les objets et attributs *critiques* seront classés puis sélectionnés selon leur importance vis-à-vis du traitement/objectif envisagé. Ils constitueront une ***sous-base de données critiques*** (voir Figure 1). Par exemple, pour une démarche de veille concurrentielle, la volonté est de connaître les produits des principaux concurrents de l'entreprise, leurs caractéristiques et performances : les informations sur ces entités et sur leurs propriétés réelles auront un degré d'importance plus élevé et seront représentées par des objets et attributs critiques.

Etape 5. Evaluation des critères de qualité

Le but de cette étape est d'évaluer, pour chaque objet et attribut critiques, les critères de qualité choisis à l'étape 2. Chaque critère de qualité sera évalué par :

- des mesures directes,

- des mesures indirectes pouvant être:
  . des indicateurs objectifs de qualité (date de création, mise à jour, méthode de collecte,...),
  . des indicateurs subjectifs de qualité nécessitant une évaluation "par expérience" de l'opérateur (expertise de la donnée),
  . des cartes de mesures pour suivre les variations des données [Red96],
  . des approximations.

Les procédures de mesure objective et les protocoles d'évaluation subjective doivent être explicitement décrits. En cas de conflits entre des valeurs issues de différentes sources et dans l'incertain, la qualité relative repose principalement sur l'expérience de l'opérateur qui supervise la saisie, évalue les données et leur affecte des indicateurs subjectifs de qualité. Un seuil de tolérance à la non-qualité peut être exprimé selon la criticité de la donnée (Figure 1 : seuil du minimum de qualité).

Aux données critiques seront associées des méta-données de qualité, c'est-à-dire des couples de type (critère de qualité, évaluation). Ces méta-données seront stockées au sein d'une sous-base de méta-données associée aux données critiques. Par exemple, un des critères de qualité essentiels dans la démarche de veille concurrentielle (ayant été choisis à l'étape 2) pourra être la vraisemblance de la donnée. Pour chaque donnée, ce critère et son évaluation constituera une méta-donnée qui lui sera associée.

Etape 6. Identification des problèmes de non-qualité et analyse de leurs causes

Le but de cette étape est de faire apparaître les problèmes de non-qualité des données (erreurs ou données de qualité médiocre) et d'en analyser les causes. Après avoir identifié les acteurs et les traitements qui créent, maintiennent, collectent, visualisent ou utilisent chaque entité/attribut (voir Figure 1 : analyse des flux de données), il peut apparaître qu'un traitement soit mal utilisé ou inadapté et/ou ne satisfasse pas les utilisateurs. Il s'agit de détecter ces anomalies de traitement. Les efforts se concentrent sur l'identification des problèmes dans les activités manuelles ou informatisées et sur les opportunités d'amélioration liées à la qualité des données.

Cette analyse de l'existant doit notamment identifier : qui réalise la sélection des sources d'information, qui réalise le recueil des données multi-sources, leur mise en correspondance, ainsi que les activités de saisie, de sélection des données, l'évaluation des critères de qualité, quelles sont les activités qui sont informatisées mais qui restent manuelles, quelles sont les corrections manuelles, quelle est la satisfaction des utilisateurs... L'identification des corrections manuelles sert de base à la validation du modèle conceptuel de la base de données multi-sources. En particulier, il s'agit d'insister sur les corrections répétitives effectués par un utilisateur et de s'interroger sur l'utilité d'un attribut non utilisé pour formaliser et automatiser des règles de décision. Cette étape a pour but de rechercher les causes de non-qualité des données en classant les événements perturbant la chaîne de l'information en événements externes et internes, prévus, prévisibles ou non, répétitifs (cycliques) ou ponctuels (aléatoires). Il s'agit de savoir/prévoir quelles données sont/seront affectées par ces événements. Si un événement affecte les données d'un seul objet, il faut être attentif aux conséquences indirectes de tels changements sur d'autres objets.

Etape 7. Recommandation des données et définition de nouveaux objectifs de qualité

L'évaluation et l'analyse de la (non-)qualité des données dans la base permet une recommandation des données de meilleure qualité ou, tout au moins, celles qui, parmi les

données homologues paraissent les plus appropriées au besoin de l'utilisateur requérant. La recommandation doit être adaptative, multi-critère et recalculée dynamiquement en fonction :
- de chaque nouvel arrivage de données,
- des préférences de l'utilisateur consultant qui fixe un seuil de qualité de données requise (Figure 1 : seuil du maximum de qualité).

L'entreprise doit également se fixer des objectifs de progrès sur la qualité de ses données à moyen et à long terme, définir des ressources et des stratégies à mettre en oeuvre pour les atteindre et pour assurer le soutien et l'amélioration de la qualité de ses données. La stratégie de maintenance de la base doit prendre en compte l'évolution des besoins des utilisateurs en termes de qualité de données.

4. Recommandation multi-critère des données selon leur qualité relative

La démarche précédente présente une vision globale et concrète de la mise en œuvre de la qualité des données dans un environnement multi-source. Dans cette section, nous nous intéressons plus précisément à la recommandation des données selon leur qualité relative, recommandation pour laquelle nous proposons des algorithmes. Le système qui les met en œuvre peut s'apparenter aux systèmes de recommandation basée sur le contenu pour des données semi-structurées [RV97].

4.1. Evaluation subjective et représentation des critères de qualité relative

Chaque donnée est évaluée sur un ensemble de critères de qualité. L'évaluation de la qualité relative nécessite une expertise de la donnée par un opérateur, spécialiste du domaine. L'évaluation peut être relativement subjective. L'échelle adoptée est présentée dans le Tableau 3.

Terme	Graduation
très bon	1.00
bon	0.75
moyen	0.50
mauvais	0.25
très mauvais	0.00

Tableau 3. Echelle linguistique et ordinale pour l'évaluation d'un critère de qualité

La base de données multi-sources est une base de données non déterministe [VN95] dont les attributs possèdent, pour la même instance, plusieurs valeurs plus ou moins contradictoires et de qualité inégale. Chaque objet de la base de données multi-sources possédera la structure présentée dans la Figure 2 :

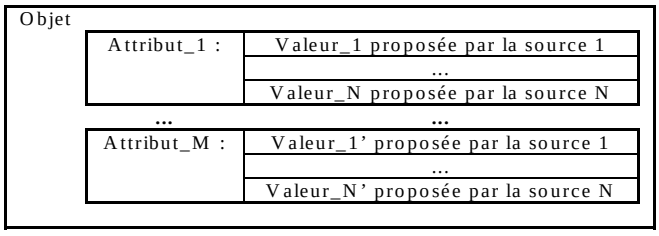

Figure 2. Structure d'un objet dans la base de données multi-sources

Chaque valeur est évaluée sur chacun des critères de qualité (choisis lors de l'étape 2 de la démarche). Un opérateur, expert du domaine, sera chargé de l'expertise et de l'évaluation subjective de chaque donnée. Typiquement, la structure de chaque valeur est telle que le montre la Figure 3 :

Valeur	Identifiant de la source	Qualité évaluée et datée
		Critère C_1: Evaluation E_1
		...
		Critère C_n : Evaluation E_n

Figure 3. Structure d'une valeur dans la base de données multi-sources

4.2. Calcul des scores et ordonnancement de qualité

L'objectif de la recommandation multi-critère des données selon leur qualité relative est de proposer la donnée *correcte la plus appropriée*. Toutes les données sont a priori jugée *correcte* si elles sont conformes à leur source d'origine. La donnée *la plus appropriée* est celle qui répond le mieux à la qualité requise par l'utilisateur (celui-ci peut par exemple privilégier la fraîcheur des données à la fiabilité ou à la précision).

Il est souvent nécessaire de prendre en compte l'importance relative des critères de qualité les uns par rapport aux autres et, s'il y a lieu, leur affecter une pondération qui sera utilisée dans la recommandation. Dans ce cas, les critères ne sont pas interchangeables et l'utilisateur requérant peut attribuer un poids à chacun des critères de qualité selon l'importance qu'il lui accorde.

Nous avons choisi d'appliquer aux données et à leur source, les méthodes existantes pour la sélection et l'évaluation de produits logiciels [FC94]. La recommandation consiste à calculer pour chaque donnée un score ou bien un ordonnancement de sa qualité.

Parmi les nombreux modèles de calcul de scores et d'ordonnancements existants, nous avons implanté les modèles de calcul décrits dans [FC94] [And90]. Pour la suite de l'article, nous avons choisis d'appliquer, à un exemple, quatre de ces modèles car ils prennent en compte l'aspect relatif de la qualité :

- *l'affectation linéaire du poids (AL)* : cette méthode permet le calcul d'un score de qualité pour chaque donnée tel que :

$$Qscore(D_i) = \sum_{k=1}^{l} W_k E_{ik}$$ où W_k est le poids du $k^{ième}$ critère de qualité et E_{ik}, l'évaluation du $k^{ième}$ critère pour la $i^{ième}$ donnée D_i

- *l'élimination par aspects (EA) :* cette méthode classe les critères de qualité par importance, puis élimine les données ayant le plus mauvais score de qualité pour le critère le plus important. L'opération est répétée jusqu'à ce qu'il ne reste plus qu'une seule donnée pour la recommandation ou bien que tous les critères aient été considérés.

- *la méthode d'Anderson (AND)* [And90] : cette méthode utilise trois matrices indiquant chacune la situation relative des données les unes par rapport aux autres :
  1) la matrice contenant la concordance des évaluations, c'est-à-dire qui considère la pondération des critères pour l'évaluation desquels "une donnée i est meilleure qu'une donnée j ",
  2) la matrice de discordance qui considère la pondération des critères pour l'évaluation desquels "une donnée i est plus mauvaise qu'une donnée j ",

3) la matrice de magnitude des évaluations qui considère "de combien une donnée i est meilleure qu'une donnée j ".

- *la méthode de Subramanian et Gershon (SG) :* sur le même principe que la méthode d'Anderson, la pondération des critères de qualité se fait lors de la construction des matrices de concordance et de discordance.

Nous avons appliqué ces modèles de calcul aussi bien pour la recommandation des données structurées (valeurs d'attribut pour un objet multi-source), que pour la recommandation des données semi-structurées (recommandation des sources de données).

4.3. Stratégies de recommandation

Le choix du modèle de calcul des scores de qualité dépend de la stratégie de recommandation que l'utilisateur souhaite. Le modèle de calcul sera choisi selon ses propriétés telles que :

- la compensation (ou non) des scores : équilibre entre les évaluations des différents critères de qualité,
- l'utilisation des informations : certains modèles utilisent les évaluations de tous les critères de qualité de toutes les données simultanément, alors que d'autres modèles comparent les données sur la base d'un seul critère,
- les préférences sur les critères : certaines permettent d'affecter un poids numérique ou un ordonnancement des critères de qualité,
- la complexité des calculs,
- un seuil du minimum de qualité : une donnée dont la qualité sera inférieure à ce seuil sera rejetée pour la recommandation,
- un ordonnancement unique des données.

Nous avons vérifié les propositions de [And90] et [FC94] telles que chaque modèle de calcul possède respectivement les propriétés suivantes (Tableau 4) :

Modèle	Compensation	Utilisation Information	Préférences	Complexité	Seuil	Unicité	Type de résultat
AL	Oui	Bonne	Oui	Elevée	Non	Oui	Ordonnancement
EA	Non	Faible	Oui	Faible	Oui	Non	Solution unique
AND	Oui	Très Bonne	Oui	Elevée	Non	Oui	Classement
SG	Oui	Très Bonne	Oui	Elevée	Non	Non	Classement

Tableau 4. Propriétés des modèles de calcul

4.4. Algorithmes de recommandation multi-critère

A partir de ces modèles de calcul et pour déterminer quelle est la source de meilleure qualité, nous définissons un accumulateur de score pour chaque source de données et proposons l'algorithme de recommandation (1) suivant :

Algorithme Recommandation_Sources (1)

Déclaration
Soit Sacc[s] l'accumulateur de score pour la source s

Début
Initialisation de l'accumulateur à 0
Pour chaque source
Retrouver l'ensemble des critères de qualité évalués {c_1, c_2,..., c_l} pour la source s
/* l : nombre de critères décrivant la qualité de la source s */
Pour chaque critère de qualité c_k évalué (avec k = 1,..., l)
Faire
Début
Retrouver l'ensemble des sources {$s_{ck,1}$, $s_{ck,2}$,... ,$s_{ck,mk}$}
pour lesquelles le critère c_k a été évalué
/* mk : nombre de sources pour lesquelles le critère de qualité c_k a été évalué */
Pour chaque source $s_{ck,j}$ (avec j = 1,..., mk)
Faire
Sacc[$s_{ck,j}$] = Sacc[$s_{ck,j}$] + Sscore($s_{ck,j}$, c_k)
/* Sscore($s_{ck,j}$, c_k) : fonction donnant le score de la source $s_{ck,j}$
pour le critère de qualité c_k */
Fin
Tri des accumulateurs de score par ordre décroissant
Affichage des valeurs par ordre selon les scores des accumulateurs
Fin.

La fonction Sscore($s_{ck,j}$, c_k) est calculée à partir d'un des modèles de calculs précédents dont le choix dépend de la stratégie de recommandation retenue.

Dans le cas où plusieurs valeurs de données sont en conflit pour un même attribut d'objet (Figure 2), nous proposons l'algorithme de recommandation (2).

Algorithme Recommandation_Valeurs (2)

Déclaration
soit Vacc[v] l'accumulateur de score pour la valeur v

Début
Initialisation de l'accumulateur à 0
Pour chaque objet critique o de la base de données multi-sources
Retrouver l'ensemble des attributs critiques {a_1, a_2,..., a_l} de l'objet o
/* l : nombre d'attributs critiques décrivant l'objet o */
Pour chaque attribut critique a_i (avec i = 1,..., l)
Retrouver l'ensemble des valeurs {$v_{ai,1}$, $v_{ai,2}$,..., $v_{ai,mi}$} de l'attribut a_i
/* mi : nombre de valeurs de l'attribut a_i proposées par les différentes sources
qui décrivent l'attribut a_i */
Pour chaque valeur $v_{ai,j}$ (avec j = 1,..., mi)
Faire
Début
Pour chaque critère de qualité c_k évalué pour la valeur $v_{ai,j}$ (avec k = 1,...,nj)
/* nj : nombre de critères de qualité pour la valeur $v_{ai,j}$ */
Retrouver l'ensemble des valeurs {$v_{ck,1}$, $v_{ck,2}$,... ,$v_{ck,ml}$}
pour lesquelles le critère c_k a été évalué
/* ml : nombre de valeurs pour lesquelles le critère de qualité ck

```
                    a été évalué (ml ≤ mi) */
            Pour chaque valeur v_ck,p (avec p = 1,..., ml)
                Faire
                  Début
                    Retrouver la source s de la valeur v_ck,p
                    Vacc[v_ck,p] = Vacc[v_ck,p] + Vscore(v_ck,p, c_k, s)
                    /* Vscore(v_ck,p, c_k, s) : fonction donnant le score
                                          de la valeur v_ck,p pour le critère
                    de qualité c_k et en considérant la qualité de sa source s */
                  Fin
        Fin
Tri des accumulateurs de score par ordre décroissant
Affichage des valeurs par ordre selon les scores des accumulateurs
Fin.
```

La fonction $Vscore(v_{ck,p}, c_k, s)$ est calculée à partir d'un des modèles de calculs précédents, selon la stratégie de recommandation choisie. La recommandation des données est basée sur un constat simple : la qualité d'une donnée dépend de la qualité de sa source et inversement. La fonction $Vscore(v_{ck,p}, c_k, s)$ prend donc en compte le score de qualité de la source de la valeur.

5. Application au domaine de la veille technologique

5.1. Exemple

La démarche de mise en oeuvre de la qualité des données dans un contexte multi-source s'applique tout particulièrement à la veille technologique [BG98]. En effet, celle-ci repose sur les trois principales étapes suivantes (Figure 4) :

A) la collecte des descriptions d'un objet réel faites par différentes sources et le stockage des données dans une base,

B) l'évaluation de différents critères de qualité des données,

C) la recommandation des données lors d'une consultation.

A), B) et C) correspondent respectivement aux étapes 3, 5 et 7 de la démarche présentée précédemment.

A. Sélection des sources et recueil des données et méta-données de qualité

Pour la cellule Veille Technologique d'une entreprise, un produit concurrent ***P*** est une entité réelle qui n'est jamais entièrement et/ou précisément connue. Pour connaître ses caractéristiques et faire, entre autre, des études plus approfondies, les veilleurs qui n'ont pas nécessairement un accès direct au matériel. Ils doivent donc se fier à un ensemble de sources d'informations textuelles (rapports techniques, plaquettes du constructeur, des brochures commerciales...). L'environnement multi-source est constitué d'un ensemble de sources, notées S1, S2 et S3 sur la Figure 4. Chaque source propose des caractéristiques pour décrire le produit ***P***. Les valeurs proposées se recoupent ou sont contradictoires. Dans l'exemple, « longueur de 12m » pour S1 et « long : 15,023 » pour S3.

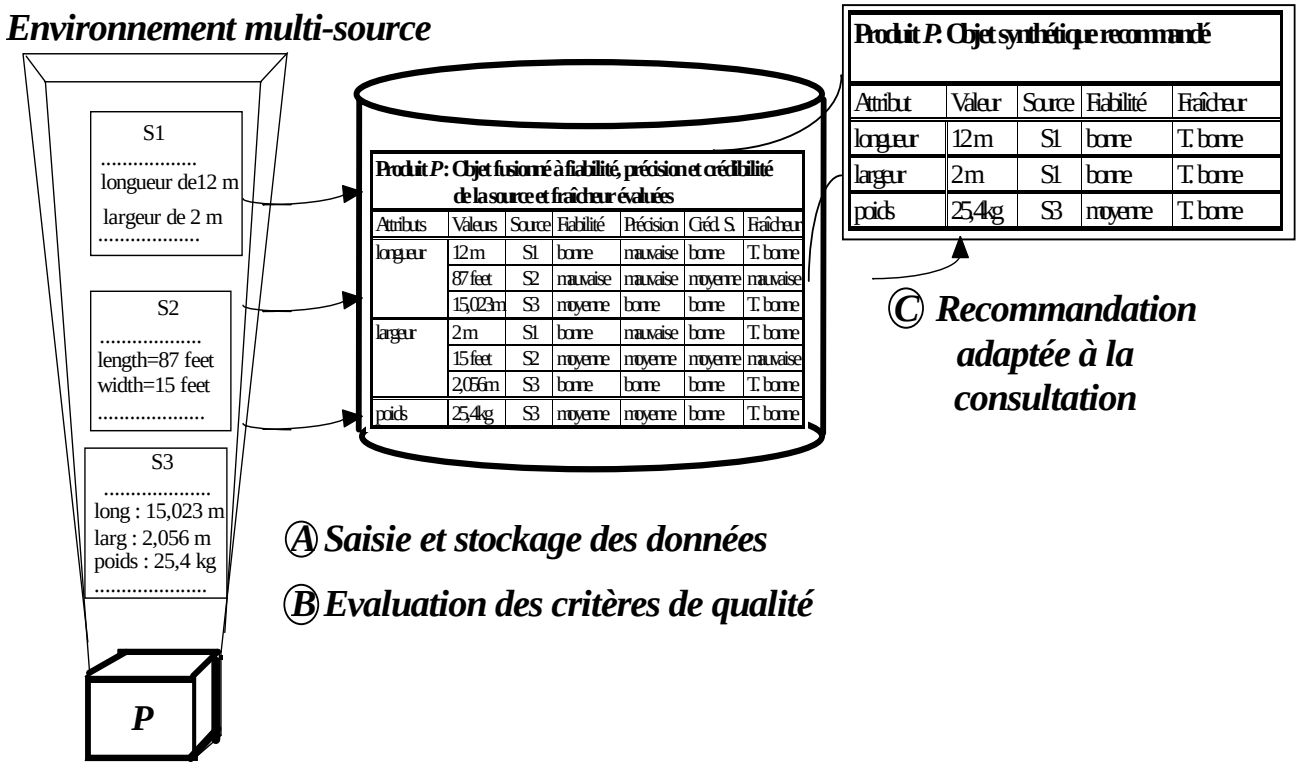

Figure 4 : Les étapes impliquant le veilleur et le système d'information

Dans un premier temps, les veilleurs devront rassembler toutes les descriptions sur le produit ***P*** faites par les sources disponibles. Puis, ils doivent structurer et stocker ces descriptions dans la base de données multi-sources sous la forme d'un *objet fusionné* dont les attributs ont des valeurs multiples avec la mention de leur source d'origine. Dans l'exemple, le produit ***P*** a pour longueur le Tableau 5 :

longueur	12 m	S1
	87 feet	S2
	15,023 m	S3

Tableau 5. « longueur » : un attribut à valeurs multi-sources

B. Evaluation des critères de qualité pour chaque valeur

Dans l'exemple, on suppose que les quatre critères de qualité retenus lors des étapes de spécifications et de sélection des critères de qualité (étapes 1 et 2 de la démarche) sont : la fiabilité de la donnée, sa précision, la crédibilité de sa source et la fraîcheur de la donnée.

Ces critères sont classés par degré d'importance par l'opérateur (par exemple, Fiabilité : 60%, Précision : 20%, Crédibilité : 15% et Fraîcheur : 5%). La qualité de chaque valeur est évaluée selon ces critères par le veilleur, expert du domaine et chargé de la saisie des méta-données. Dans l'exemple, la fiabilité de la valeur « 12 m » proposée par la source S1 est évaluée « bonne ».

C. Recommandation adaptative des valeurs

L'utilisateur qui consulte le système peut avoir des exigences en termes de qualité de données et ordonner lui-même les critères de qualité qu'il juge prépondérants. L'utilisateur spécifiera lui-même ses préférences sur les critères de qualité pour que lui soit recommandé un jeu d'informations adapté à son besoin. Ainsi, lors de la consultation, le système proposera une recommandation adaptée des valeurs au sein de chaque attribut de l'objet fusionné. Un objet synthétique (ou objet de synthèse) sera construit à la demande à partir des valeurs existantes de l'objet fusionné et selon les meilleures évaluations pour la qualité requise.

Dans l'exemple de la Figure 4, l'utilisateur ne s'intéresse qu'à la fiabilité et à la fraîcheur de la donnée en privilégiant la fiabilité (Fiabilité : 70%, Fraîcheur : 30%). Il choisit également une stratégie de recommandation. L'objet synthétique qui lui est finalement proposé, est composé de valeurs issues de sources hétérogènes (S1 et S3) pour leur précision puis fraîcheur optimales.

5.2. Résultats des différentes stratégies de recommandation

Dans les cas de valeurs multi-sources, les différentes stratégies de recommandation (LA, EA, AND, SG) proposent un ordonnancement des valeurs selon leur qualité relative. La Figure 5 présente le résultat de chaque stratégie.

Dans l'exemple, les valeurs proposées par la source S1 pour la longueur et la largeur de ***P*** arrivent en tête pour la recommandation quel que soit le modèle de calcul choisi par l'utilisateur.

Dans le cas où il n'existe qu'une seule valeur disponible pour renseigner l'attribut, la valeur sera proposée si son seuil de qualité est supérieur à celui du minimum de qualité. C'est le cas de la valeur proposée par la source S3 pour le poids du produit ***P*** (Figure 4).

Si l'utilisateur ne choisit pas explicitement sa stratégie de recommandation, ou bien s'il ne spécifie pas le poids des critères de qualité, la recommandation par défaut sera celle choisie par le veilleur (dans l'exemple, Fiabilité : 60%, Précision : 20%, Crédibilité : 15% et Fraîcheur : 5%).

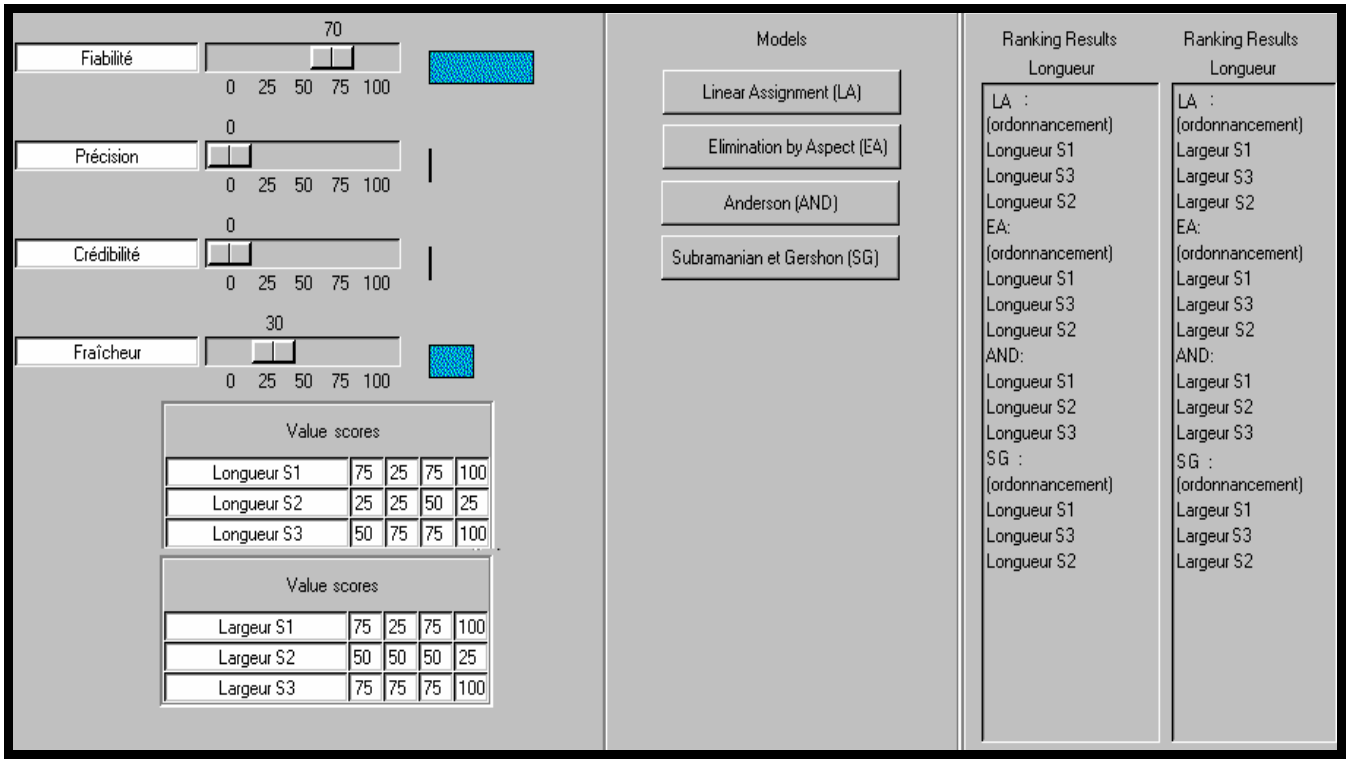

Figure 5. Résultats de recommandation pour l'exemple

6. Conclusion et perspectives

Les objectifs de cet article étaient les suivants :

- tout d'abord, analyser les travaux existants et introduire le thème de la qualité des données multi-sources à la convergence de différents domaines de recherche,
- puis, proposer une méthodologie de mise en oeuvre de la qualité relative des données : cette méthodologie introduit la qualité des données multi-sources dans la conception, le développement et la maintenance d'un système spécifique, et dans toute la chaîne de traitement de l'information,
- présenter ensuite des outils pour la recommandation des données basée sur leur qualité relative, (notamment, un mode de représentation, des modèles de calcul de scores et d'ordonnancements et des algorithmes de recommandation),
- enfin, valider la démarche et ses outils en les appliquant au domaine de la veille technologique.

Cette démarche aussi coûteuse qu'elle puisse paraître, est très adaptée à ce domaine d'application, où toutes les décisions reposent déjà sur l'expertise de l'information. Chaque information est qualifiée et évaluée par un veilleur, expert du domaine. En proposant une recommandation des données, le système d'information permet de présenter la valeur ajoutée par le veilleur et d'en préserver la justification.

Les orientations que nous poursuivons actuellement, concernent :

- l'assistance au choix de la stratégie de recommandation, pour guider l'utilisateur et cibler le meilleur modèle de calcul,
- l'anticipation du système pour une recommandation en fonction des profils d'utilisateurs (recommandation par défaut, acquisition implicite des profils),
- la recommandation floue (fonctions de calcul disjonctives/conjonctives pour l'agrégation des évaluations des critères de qualité).

Les perspectives de recherche sur la qualité des données multi-sources et leur recommandation sont nombreuses. Elles dépassent largement le cadre de la veille technologique. Elles concernent notamment la recommandation personnalisée des données semi-structurées et l'extension de la notion d'ordonnancement de pertinence *(relevancy-ranking).* Ceci constitue une intéressante perspective qui motive nos travaux dans le cadre du projet QIRi@D *(Quality and Information Retrieval in electronic Document)* [BDM98].

7. Références

[And90] Anderson E., Choice models for evaluation and selection of software packages, Journal of Management Information Systems, vol. 6, p. 123-138, 1990

[Bas95] Bash R. (Ed.), Electronic informartion delivery : ensuring quality and value, 1995

[BDM98] Berti L., Damoiseaux J. L., Murisasco E., Combining the power of query languages and search engines for on-line document and information retrieval : the QIRi@D Environment", Proc. of the Intl. Workshop on Principles of Digital Document Processing (PODDP '98), St Malo, 1998

[Ber98] Berti L., From data source quality to information quality : the relative dimension, Proc. of the 3rd Intl. Conf. on Information Quality, MIT, Cambridge, p.247-263, 1998

[BG98] Berti L., Graveleau D., Contribution à la définition d'un vigiciel : quelle modélisation de l'information factuelle, événementielle et référentielle ?, Actes du colloque Veille Stratégique, Scientifique et Technologique (VSST '98), Toulouse, , octobre 1998

[Bon96] Bonjour E., La qualité et la mise à jour des Bases de données Techniques utilisées en GPAO, Université de Franche-Comté, 1996

[BP85] Ballou D., Pazer H., Modeling data and process quality multi-input, multi-output information systems, Management Science, vol. 31, no. 2, p. 150-162, 1985

[Bro90] Brodie M. L., Data quality in information systems, Information and Management, vol. 3, p. 245-258, 1980

[CP87] Cavallo R., Pitarelli M., The theory of probabilistic databases, Proc. of the 13th Intl. Conf. on Very Large Data Bases, p. 71-81, 1987

[CP98] Chengalur-Smith I., Pipino L. (Ed.), Proc. of the 3rd Intl. Conf. on Information Quality, MIT, Cambridge, 1998

[CPPR98] Calabretto S., Pinon J. M., Poullet L., Richez M. A., De la qualité de l'information à la qualité de la documentation, Document Numérique, vol.12, no.1, p. 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

[DeM89] De Michiel L.G., Resolving database incompatibility : an approach to performing relational operations over mismatched domains, IEEE Transactions on Knowledge and Data Engineering, vol. 4, p. 485-493, 1989

[GJ98] Goodchild M., Jeansoulin R. (Ed.), Data quality in geographic information : from error to uncertainty, Hermès, 1998

[FC94] Fritz C., Carter B., A classification and summary of software evaluation and selection methodologies, Technical Report, Mississippi State University, 1994

[FLR94] Fox C., Levitin A., Redman T., The notion of data and its quality dimensions, Information Processing and Management, vol. 30, no. 1, 1994

[FPSSU96] Fayyad U., Piatetsky-Shapiro G., Smyth P., Uthurusamy R. (Ed.), Advances in Knowledge Discovery and Data Mining, AAAI Press, MIT Press, 1996

[Jac92] Jacso P., CD-ROM software, dataware and hardware : evaluation, selection and installation, 1992

[KS92] Kashyap V., Sheth A., So far (schematically) yet so near (semantically), Proc. of the IFIP Database Semantics Conf. on Interoperable Database Systems, 1992

[Lau86] Laudon K. C., Data quality and due process in large interorganizational record systems, Communications of the ACM, p. 4-11, 1986

[Lee92] Lee S. K., An extended relational database model for uncertain and imprecise information, Proc. of the 18th Intl. Conf. on Very Large Data Bases, p.614-621, 1992

[LU90] Liepins G., Uppuluri V., Data quality control : theory and pragmatics, M. Dekker, New-York, 1990

[LSS94] Lim E. P., Srivastava J., Shekhar S., Resolving attribute incompatibility in database integration : An evidential reasoning approach, Proc. of the 10th Intl. Conf. on Data Engineering, 1994

[PAGM96] Papakonstantinou Y., Abiteboul S., Garcia-Molina H., Object fusion in mediator systems, Proc. of the 22nd Intl. Conf. on Very Large Database Systems, 1996

[PC93] Parsaye K., Chignell M., Intelligent Database Tools and Applications, Wiley & Sons, 1993

[Red96] Redman T., Data quality for the information age, Artech House Publishers, 1996

[Rot96] Rothenberg J., Metadata to support data quality and longevity, Proc. of the 1st IEEE Metadata Conf., 1996

[RV97] Resnick P., Varian H. R., Recommender systems, Communications of the ACM, vol. 40, no. 3, p. 56-89, 1997

[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

[Sad91] Sadri F., Reliability of answers to queries in relational databases, IEEE Transactions on Knowledge and Data Engineering, vol. 3, no. 2, p. 245-252, 1991

[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 2rd Intl. Conf. on Information Quality, MIT, Cambridge, 1997

[SLW97] Strong D., Lee Y., Wang R., Data quality in context, Communications of the ACM, vol. 40, no. 5, p. 103-110, 1997

[SM91] Siegel M., Madnick S. E., A metadata approach to resolving semantic conflicts, Proc. of the 17th Intl. Conf. on Very Large Data Bases, p. 133-145, 1991

[TB98] Tayi G. K., Ballou D. P., Examining Data Quality, Communications of the ACM, vol. 41, no. 2, p. 54-57, 1998

[TCY93] Tseng F. S., Chen A. L., Yang W. P., Answering heterogeneous database queries with degrees of uncertainty, Distributed and Parallel Databases, vol. 1, p. 281-302, 1993

[UW91] Urban S.D., Wu J., Resolving semantic heterogeneity through the explicit representation of data model semantics, SIGMOD Record, vol. 20, no. 4, p. 55-58, 1991

[VN95] Vadaparty K.V., Naqvi S. A., Using constraints for efficient query processing in nondeterministic databases, IEEE Transactions on Knowledge and Data Engineering, vol. 7, no. 6, p. 850-864, 1995

[WM90] Wang R., Madnick S. E., A polygen model for heterogeneous database systems : the source tagging perspective, Proc. of the 16th Intl. Conf. on Very Large Data Bases, p. 519-538, 1990

[WKM93] Wang R., Kon H. B., Madnick S. E., Data quality requirements analysis and modeling, Proc. of the 9th Int. 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, vol. 7, no. 4, p. 670-677, 1995

[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, Communications of the ACM, vol. 41, no. 2, p. 58-65, 1998