Hiérarchiser les dimensions de la qualité des données : analyse comparative entre la littérature et les praticiens en technologies de l'information

par

Martin Goulet

Essai présenté au CeFTI
en vue de l'obtention du grade de maître en technologies de l'information
(Maîtrise en génie logiciel incluant un cheminement de type cours en technologies de l'information)

FACULTÉ DES SCIENCES
UNIVERSITÉ DE SHERBROOKE

Longueuil, Québec, Canada, février 2012

Sommaire

Depuis l'avènement de l'informatique et plus particulièrement des bases de données, les entreprises se sont mises à accumuler de très grandes quantités de données. Elles sont utilisées par plusieurs de leurs processus et de leurs employés. Ces données peuvent être présentes dans un ou plusieurs systèmes d'information. Il peut s'agir de données se rapportant à la comptabilité, aux clients, aux produits, aux fournisseurs de même que sur une foule d'autres sujets propres à chaque entreprise. Ces données ont parfois été entrées (saisies) à des époques différentes dans des systèmes différents. Dans certains systèmes, les données qui y ont été saisies ont pu faire l'objet de contrôles rigoureux. Tandis que dans d'autres systèmes, elles ont pu y être saisies de manière moins rigoureuse. Vu la grande diversité des processus et des sources de données, il peut être aisé de penser que la qualité des données comprises dans ces systèmes peut varier d'un système à l'autre. Puisque les données sont presque toujours vitales pour les entreprises, il devient impératif de connaître le niveau de qualité des données comprises dans ces systèmes.

La qualité des données est un sujet qui prend de plus en plus d'ampleur au sein des entreprises. Elle devient un incontournable qui permet aux entreprises de tirer le plein potentiel des données qu'elles utilisent tant dans leurs opérations courantes que dans leur prise de décision stratégique.

Afin de pouvoir mesurer et contrôler la qualité des données qui sont comprises dans les différents systèmes des entreprises, qu'il s'agisse d'entrepôt de données, de centre de données ou de base de données, il importe de mettre en place les processus nécessaires. Dans le cadre d'un projet en qualité des données, les dimensions sont utilisées afin de définir les mesures qui permettront de connaître le niveau de qualité des données. Puisque la littérature contient un grand nombre de dimensions et vu la somme d'effort et les coûts nécessaires à la maîtrise et à la mise en place d'une dimension, il devient primordial pour les gestionnaires de données

d'être en mesure d'identifier les dimensions qui auront le plus d'impacts sur la qualité de leurs données, mais lesquelles? Cet essai se penche sur cette question par l'entremise de trois objectifs.

Le premier objectif consiste à analyser la littérature pour identifier les dimensions les plus importantes. À la suite de cette analyse, une compilation des dimensions présentes dans la littérature est présentée. Le deuxième objectif consiste à définir l'importance devant être accordée à ces dimensions selon les praticiens en technologies de l'information. Le troisième objectif consiste à comparer les résultats de ces deux analyses afin de voir si la littérature s'accorde ou non avec les praticiens. C'est lors de cette dernière analyse que les critères permettant d'identifier les dimensions les plus importantes seront définis. Ces critères prennent en considération chacune des analyses effectuées. Seules deux dimensions se conforment aux exigences définies. Il s'agit de : « actualisation et disponibilité » et « interprétable ».

Remerciements

Pour parvenir à compléter la rédaction d'un document de ce type, il faut employer plusieurs stratégies de gestion du temps et se fixer un horaire qui se doit d'être respecté scrupuleusement. Outre, cette organisation, plusieurs sacrifices et concessions doivent être pris en compte. Il va sans dire, que tous les gens qui nous sont proches doivent également composer avec les contraintes inhérentes à cette somme d'effort. C'est pourquoi je tiens à remercier ma conjointe, Isabelle, qui a su faire preuve d'une grande patience dans le respect de mes horaires et qui a su du même coup m'encourager à poursuivre afin que j'atteigne le but que je me suis fixé. Je tiens aussi à remercier mes parents, mon frère et ma sœur qui m'ont également formulé plusieurs encouragements et m'ont permis de parvenir à mon but.

Il est aussi important de souligner le soutien de mes directeurs qui ont su m'orienter et me fournir des conseils judicieux ce qui m'a permis de concentrer mes efforts sur l'objectif de mon essai. Leur point de vue fut très apprécié puisqu'il m'a permis d'obtenir une vision externe de mon essai et ainsi avoir un œil plus critique sur le contenu de celui-ci. Leur bagage académique et professionnel n'aurait pu être remplacé d'aucune façon. Je tiens à souligner tout le temps qu'ils ont mis à la lecture de mon essai et à la rédaction de leurs judicieux commentaires.

Merci également à toutes les personnes à qui il m'a été donné de discuter de mon projet de rédaction. C'est parfois dans le cadre de ces discussions qu'il m'a été possible de me questionner sur la pertinence de tel ou tel aspect de mon essai.

Je tiens aussi à remercier l'Université de Sherbrooke, qui a su mettre à profit les ressources dont elle dispose, telles que la bibliothèque, les locaux de rencontre, le personnel et la documentation en ligne. Ces ressources ont en partie contribué à l'accomplissement de ce travail de recherche et de rédaction.

Table des matières

Liste des tableaux

Liste des figures

Glossaire

Assurance qualité	Ensemble des actions préétablies et systématiques nécessaires pour donner la confiance appropriée en ce qu'un produit ou service satisfera aux exigences données relatives à la qualité [17].
Donnée	Représentation d'une information, codée dans un format permettant son traitement par ordinateur [17].
Gestion des données	La gestion des données comprend des activités, telles que le développement, l'exécution et la supervision des processus, des politiques et des pratiques qui visent à contrôler, à publier et à améliorer la valeur des données et des informations de l'entreprise [16].
Information	Élément de connaissance concernant un phénomène et qui, pris dans un contexte déterminé, a une signification particulière [17].
Mesure (métrique)	Action d'utiliser une unité de mesure pour évaluer et mesurer le niveau de qualité des données [17].
Qualité des données	Valeur des données, fondée sur une appréciation de leur exactitude, de leur actualité, de leur précision, de leur exhaustivité, de leur pertinence et de leur accessibilité, en vue de leur utilisation [17].
Nuplet	Dans un système de gestion de base de données, il s'agit d'une ligne qui peut être composée d'un ou plusieurs attributs [17].

Liste des sigles, des symboles et des acronymes

BD	Base de données
CeFTI	Centre de formation en technologies de l'information
DAMA	*The Data Management Association :* Association en gestion des données
DMBOK	*The Data Management Body of Knowledge* : Le corpus des connaissances en gestion des données
ICIS	Institut canadien d'information sur la santé
ISO	*International Organization for Standardization* : Organisation internationale de normalisation
MDM	*Master Data Management :* Gestion des données de référence
MIT	*Massachusetts Institute of Technology* : Institut des technologies du Massachusetts
PME	Petites et moyennes entreprises
POSMAD	*Plan Obtain Store and Share Maintain Apply Dispose* : Planifier, acquérir, stocker, partager, maintenir, utiliser, supprimer ou archiver.
ProDQM	*Proactive Data Quality Management* : Gestion proactive de la qualité des données
QD	Qualité des données
RDQA	*Routine Data Quality Assessment Tool* : Outil d'évaluation de la qualité des données
SI	Système d'information
TDQM	*Total Data Quality Management* : Gestion globale de la qualité des données
TIQM	*Total Information Quality Management* : Gestion globale de la qualité de l'information

Introduction

Les données qu'une entreprise possède sont le fruit d'un ensemble de processus plus ou moins complexes. Ces données proviennent des différents systèmes d'information de l'entreprise et sont utilisés par les gestionnaires, mais aussi par les autres employés, tels que les gens de la production, de la comptabilité ou des ventes. C'est donc dire que ces données ont une très haute importance dans les tâches courantes des utilisateurs. Malheureusement, ce ne sont pas tous les processus d'entrée de données qui ont été conçus pour contrôler et valider les données qui y sont entrées. Vu ce manque de contrôle et de validation, il s'avère nécessaire pour les entreprises qui désirent utiliser ces données qu'elles en évaluent leur qualité.

La gestion de la qualité des données est une discipline qui prend de plus en plus d'ampleur. La tendance des entreprises à valoriser et à utiliser leurs données est sans cesse croissante, puisqu'elles cherchent à pouvoir en retirer des bénéfices. Ces bénéfices peuvent être stratégiques, concurrentiels, monétaires, organisationnels ou de tout autre type propre à l'entreprise.

L'évaluation de la qualité des données consiste en une élaboration de mesures qui permettent d'évaluer et de connaître le niveau de qualité des données et des informations présentes dans les différents systèmes de l'entreprise. Cette évaluation fait partie de la gestion de la qualité des données. Les mesures effectuées dans un contexte d'évaluation de la qualité des données permettent d'établir si les données sont entre autres : intègres, intégrales, complètes, précises et disponibles. Les différents aspects utilisés pour mesurer la qualité des données se nomment des dimensions. Le détail des activités qui permettent d'évaluer le niveau de qualité des données est décrit au chapitre 1.

Certaines statistiques publiées sur le site Gartner permettent de penser que les entreprises ont intérêt à mettre en place des processus qui vont leur permettre d'évaluer la qualité de leurs

données [7]. Par exemple, 25 pour cent des données critiques des grandes entreprises sont imparfaites. Les compagnies européennes ont positionné la piètre qualité des données comme étant leur deuxième principale préoccupation en intelligence d'affaires. Une étude auprès de 600 personnes impliquées en intelligence d'affaires a démontré que 35 pour cent de ses utilisateurs considèrent la piètre qualité des données comme faisant partie d'une de leurs trois principales préoccupations pour les prochains 12 à 18 mois. Ce qui place cette problématique au second rang des plus grands défis [7]. Les problèmes de qualité des données coûtent aux entreprises américaines plus de 600 milliards par année [22]. En somme, ces statistiques permettent de comprendre que les techniques de mesure et de contrôle de la qualité des données sont nécessaires aux entreprises d'aujourd'hui.

La connaissance du niveau de qualité des données peut donc être considérée avec une très haute importance, et ce, tout particulièrement pour les entreprises qui désirent obtenir des bénéfices par l'utilisation de leurs données. Les gestionnaires y verront un avantage puisqu'ils seront en mesure d'évaluer la qualité des données qui sont utilisées pour la confection de leurs rapports. Une méconnaissance de la qualité des données peut faire en sorte de diminuer l'intérêt des gestionnaires envers les systèmes d'information qui contiennent ces données et ainsi passer à côté de leur source principale d'informations. S'ils ne sont pas en mesure de connaître le niveau de qualité des données qui sont utilisées pour leurs rapports, ils ne pourront pas prendre de décisions en connaissance de cause. Cette situation pourrait entraîner des décisions inappropriées, incomplètes ou trompeuses qui auraient été biaisées par la non-qualité des données. Il va sans dire que des impacts financiers tels que la perte de clients, un approvisionnement inapproprié et des dépenses supplémentaires sont aussi envisageables. Puisque dans un contexte comme celui-ci, la qualité des données vient directement influencer la qualité des décisions qui sont prises, il devient essentiel de connaître la qualité des données comprises dans les différents systèmes de l'entreprise. Les dimensions étant à la base de plusieurs méthodes de mesures de la qualité des données, il est primordial de procéder à une sélection judicieuse de ces dimensions.

La littérature contient peu de techniques qui permettent de procéder à l'identification des dimensions pertinentes pour une entreprise donnée. La plupart du temps, cette identification est réalisée par les utilisateurs. Cette situation complexifie le processus de sélection puisque les gestionnaires de données doivent trouver ou définir la technique qui répond le mieux à leurs besoins. Le choix des dimensions étant une étape cruciale dans la mise en place des techniques d'évaluation et de mesure de la qualité des données [2], cet essai se penche sur cette problématique par l'entremise de trois objectifs. Le premier objectif consiste à analyser la littérature pour identifier les dimensions les plus importantes. Le deuxième objectif consiste à définir l'importance accordée par les praticiens pour ces dimensions. Le troisième objectif consiste à comparer les résultats de ces deux analyses afin de voir si la littérature s'accorde ou non avec les praticiens.

Au chapitre 1, les fondements conceptuels ainsi que les principaux termes utilisés en gestion de la qualité des données seront décrits. Certaines techniques de sélection présentes dans la littérature seront également décrites. Au chapitre 2, une analyse du contenu de la littérature exposera les dimensions trouvées dans celle-ci. En regard de leur fréquence de citation, une hiérarchie sera mise de l'avant. Au chapitre 3, il sera décrit la méthode utilisée afin d'obtenir le niveau de priorisation de chacune des dimensions en regard des praticiens en technologies de l'information. Au chapitre 4, les résultats obtenus par la littérature et par les praticiens seront comparés afin d'identifier les dimensions ayant la plus haute importance. En conclusion, il sera question de faire un retour sur les trois objectifs et de mentionner les contributions de ce travail.

Chapitre 1

Fondements conceptuels

Puisque les dimensions revêtent une grande importance dans le vaste monde de la qualité des données, ce chapitre propose une description de leur nature et de leur utilisation dans ce domaine. Viendra ensuite un survol des cadres de référence et de certaines techniques de sélection des dimensions qui sont présentes dans la littérature.

1.1 Donnée vs information

Les termes « donnée » et « information » sont souvent utilisés dans un contexte de qualité des données. Cette section vise à les définir et à les distinguer.

Une « donnée » est en quelque sorte la représentation d'une information qui est codée dans un format numérique, parfois sous forme analogique, permettant son utilisation et son traitement par ordinateur [17]. Emmagasinées dans un système informatique, les données ne sont pas des informations. Elles deviendront des informations dès qu'elles seront décodées dans leur contexte d'utilisation.

Quant au terme « information », il est utilisé pour identifier une donnée qui a fait l'objet d'une interprétation. Cette information prendra toute sa signification dans le contexte auquel elle est destinée [17].

A priori, une donnée hors contexte ne fournit pas d'information. Il s'agit simplement d'une valeur quelconque permettant de qualifier une portion d'une entité. Par exemple une valeur numérique (adresse, quantité, âge,…), un qualificatif (couleur, sexe, ancienneté,…), une valeur booléenne (0, 1, vrai, faux, oui, non). Une fois mise en contexte, elle devient une

information. Une information peut contenir une ou plusieurs données. Par exemple, le nombre « 34 » (une donnée), il y a 34 étudiants inscrits dans le cours INF 735 (une information).

Certains auteurs utilisent le terme « donnée » alors que d'autres utilisent le terme « information ». Les définitions associées à chacun de ces termes permettent d'en comprendre les différences. Toutefois, les auteurs s'entendent pour dire que dans un contexte de qualité des données, les termes données et informations désignent tous les deux une entité suffisamment semblable pour en confondre la définition [24]. Dans le cadre de cet essai, l'auteur a choisi d'utiliser le terme « données ». Les données étant ni plus ni moins « la matière première » de l'information.

La figure 1.1, illustre le lien entre les données, l'information et la prise de décision.

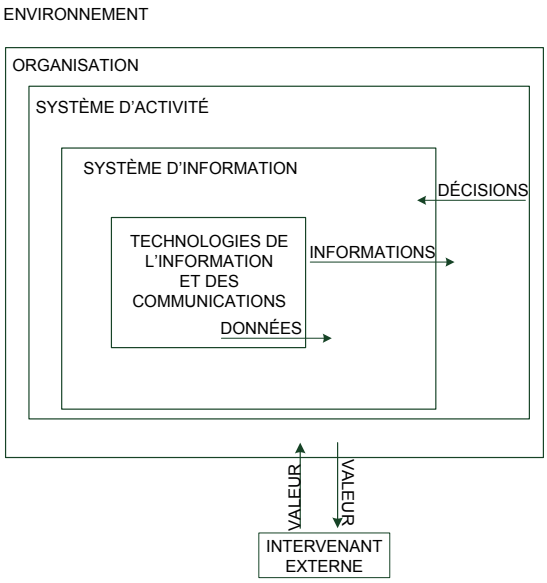


Figure 1.1 Données et informations

Traduction libre
Source : Paul B.-D., (2009), p. 9

Beynon-Davies illustre avec cette figure que les données sont d'abord traitées afin de générer de l'information qui sera par la suite utilisée pour la prise de décision [3].

1.2 Les données

Les données sont au centre de l'exploitation de l'entreprise. Elles sont utilisées par tous et sont généralement requises dans toutes les sphères de l'entreprise, tel qu'illustré à la figure 1.2. Chaque employé de l'entreprise doit consulter les données du système d'information lors de ses tâches courantes et il doit aussi les mettre à jour, en ajouter ou en supprimer. C'est pourquoi les données font l'objet d'une très grande sollicitation de la part des utilisateurs, et ce, de manière constante. Elles sont à la base d'une séquence primordiale pour la survie des entreprises. En effet, dans un contexte de gestion des connaissances, les données sont la source des informations, qui elles sont la source des connaissances qui ensuite deviennent la sagesse telles qu'illustrées à la figure 1.3. C'est donc dire que la qualité de tout ce qui utilise à la source ces données peut être directement liée à la qualité de celles-ci. Une piètre qualité des données apportera une qualité médiocre de l'information d'où en découlera une mauvaise qualité de la connaissance.

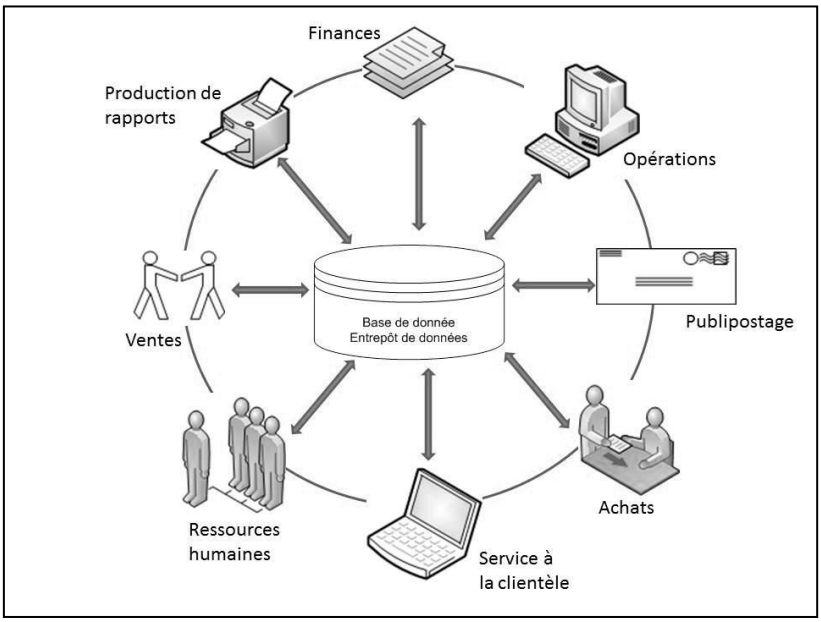


Figure 1.2 Les données au cœur de l'entreprise

Figure 1.3 Données — Informations — Connaissances

Inspiré de : Rhem A. J., (2006) et de Redman T., (1998)

Le cycle de vie des données regroupe un ensemble d'activités qui permettent de représenter et de suivre l'état d'une donnée depuis sa planification jusqu'à sa suppression (ou archivage). C'est en sachant dans quelle phase se situe la donnée qu'il sera possible d'effectuer les activités appropriées sur celle-ci.

Selon McGilvray, le cycle de vie des données est composé de six activités. Ces activités sont réalisées de manière chronologique, telle que présenté au tableau 1.1. McGilvray utilise

l'acronyme « POSMAD » pour faire référence à ces activités [16]. Cet acronyme est utilisé pour désigner chacune des phases de la version anglaise (*Plan*, *Obtain*, *Store and share*, *Maintain*, *Apply* et *Dispose*).

Tableau 1.1 POSMAD Cycle de vie des données

Phase	Définition	Exemple d'activité
Planifier (*Plan*)	Préparation en vue d'emmagasiner les données.	Déterminer les objectifs, planifier l'architecture, développer les standards et les définitions.
Acquérir (*Obtain*)	Acquisition des données.	Création des données, achat, chargement, importation, etc.
Stocker et partager (*Store and share*)	Placer les données dans un système informatique ou sur support rigide et les rendre disponibles selon la méthode établie.	Placer les données dans une base de données ou dans un fichier. Partager les informations à l'aide du réseau, des courriels, etc.
Maintenir (*Maintain*)	S'assurer que les données fonctionnent toujours correctement.	Mettre à jour, modifier, manipuler, analyser, uniformiser, etc.
Utiliser (*Apply*)	Utiliser les données pour atteindre les objectifs requis.	Accéder aux données, les utiliser. Ce qui implique de faire des transactions, de créer des rapports, prendre des décisions administratives, exécuter des traitements automatisés, etc.
Disposer (*Dispose*)	Retirer les données du système lorsqu'elles ne sont plus requises.	Archiver les données ou les supprimer.

Traduction libre
Source : McGilvray, D., (2008), p. 24

Ce qu'il est important de remarquer dans ce cycle c'est qu'il débute avant l'acquisition des données et qu'il se termine après l'utilisation de celles-ci. L'utilisation du cycle de vie des données peut s'avérer très utile dans un contexte de gestion des données puisqu'il permet de mieux cerner les actions à entreprendre envers les données en regard de l'évolution des données dans le temps. Les actions concernant l'acquisition, le maintien ou la suppression des

données sont bien différentes l'une de l'autre de par les intervenants requis, les méthodes et les moyens nécessaires pour accomplir l'action envisagée. Outre la représentation en tableau, tel que proposé par McGilvray, une représentation linéaire du cycle de vie des données peut également être utilisée, tel qu'illustré à la figure 1.4.

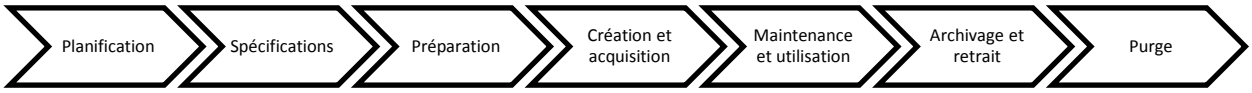


Figure 1.4 Cycle de vie des données

Traduction libre
Source : *The DAMA Guide to The Data Management Body of Knowledge*, (2009), p.4

Le cycle de vie des données représenté sous sa forme linéaire permet de mieux schématiser les différentes étapes. Puisque dans ce cas-ci il a été créé par une autre organisation, les étapes ne sont pas toutes identiques à celles du tableau 1.1. Mais il est possible de dire que dans leur ensemble elles offrent quand même plusieurs similitudes que ce soit pour le nom donné aux activités ou pour les actions dont elles sont composées.

1.3 Qualité des données

La qualité des données est constituée d'un large éventail de pratiques et de techniques comportant des outils de mesurage, d'évaluation et de contrôle qui permettent de mieux connaître les données comprises dans les systèmes en regard des différents critères (dimensions) jugés pertinents pour les données concernées. La connaissance du niveau de qualité des données peut être l'évènement déclencheur qui permettra de justifier la mise en place d'un programme d'assurance qualité qui sera utilisé afin d'en maintenir ou d'en augmenter le niveau de qualité des données en regard des différents facteurs déterminés lors de l'évaluation de la qualité des données.

Le terme « contrôle de qualité » a été popularisé par Deming au cours des années 1950. Il a défini quatorze recommandations de gestion. Ces recommandations ont par la suite été le fondement de plusieurs auteurs. Ce sont les entreprises japonaises qui furent les premières à utiliser ces recommandations. En 1954, Juran a également contribué à établir des bases pour la gestion de la qualité en proposant 10 étapes pour établir un contrôle de qualité. Au même moment, Ishikawa de son côté, préconise une amélioration constante de la qualité basée sur un diagramme de cause à effet qu'il nomme « fishbone »[1]. Viennent ensuite les années 1970, durant lesquelles les États-Unis ont mis en place des pratiques qui par la suite, au cours des années 1980, sont devenues le TQM (*Total Quality Management*). Cette méthode est le fruit d'un amalgame créé par les grands auteurs, tels que Deming, Juran et Crosby. Puis, c'est au tournant des années 2000 que la gestion des systèmes d'information a vu le jour avec l'IQM (*Information Quality Management*) [13].

Dans la littérature, le terme qualité des données ne comporte pas de définition précise. Il est plutôt défini comme étant un but qui lui est utilisé pour définir des orientations et des techniques qui permettront de l'atteindre. Toutefois, dans la littérature il y a présence de certaines formulations telles que : une donnée de qualité est une donnée qui convient aux usages auxquels les utilisateurs sont en droit de s'attendre[2]. La définition fournie par l'Office québécois de la langue française est celle présentée au glossaire. Elle est étroitement liée aux dimensions.

Les mesures effectuées dans un contexte de qualité des données permettent d'établir si les données sont justes et valables, si les utilisateurs peuvent y faire confiance, si elles seront disponibles au moment opportun, à l'endroit approprié et aux utilisateurs finaux qui doivent utiliser ces données pour leur travail. Que ce soit du point de vue des décideurs, de l'administration, de la production ou des ventes.

---

1 Diagramme en forme d'arêtes de poisson

2 Traduction libre de […we define « data quality » as data that are fit for the use by data consumers] [25]

Pour bien comprendre ce qu'est une donnée de qualité, il peut être intéressant de savoir reconnaître une donnée qui est dite de mauvaise qualité. Celles-ci peuvent être reconnues par l'absence d'une partie de la donnée, par exemple, le numéro d'immeuble absent d'un champ d'adresse, une donnée n'ayant pas la bonne valeur ou une donnée qui n'est pas à jour [25]. Un champ obligatoire ayant une valeur nulle constitue aussi une donnée de mauvaise qualité.

Des données de qualité auront pour effet d'améliorer la prise de décisions tout en fournissant une réponse plus rapidement et précisément que pour un système d'information dont la qualité des données ne serait pas connue [12]. Cette situation permet également une augmentation des revenus et une amélioration dans la fluidité des opérations [6].

Pour qu'une donnée soit dite de qualité, elle devra répondre à certains critères qui auront été établis dans les premières étapes du processus. Ces critères peuvent aller d'un extrême à l'autre. Ces extrêmes signifient qu'une donnée sera représentée par une qualité totale ou par une qualité « telle quelle » (*as is*). Le juste milieu entre ces deux extrêmes, est d'avoir des données qui répondent à certains critères dans lesquels il est possible d'y retrouver certains défauts, qui seront en somme jugés mineurs ou sans impact réel sur leurs utilisations [6]. Cette situation peut être illustrée par la nécessité d'avoir le contenu d'un champ d'adresse parfait en regard des dimensions utilisées, ou permettre une certaine souplesse en permettant, par exemple, que si une virgule située entre le numéro d'immeuble et le nom de la rue est manquante, le courrier soit tout de même acheminé au bon destinataire.

Les problèmes de qualité des données ne sont pas exclusivement liés aux TI, mais il s'agit plutôt d'une problématique qui est en grande partie liée aux processus de l'entreprise. Les solutions mises de l'avant afin d'améliorer la qualité des données ne devraient donc pas être exclusivement basées sur la correction des données contenues dans les systèmes de

l'entreprise, mais devraient faire en sorte d'améliorer les processus de l'entreprise afin que les données qui sont entrées et manipulées dans les systèmes soient de meilleure qualité [4].

La non-qualité des données est un constat qui peut avoir lieu à la suite de certaines mesures ou de certaines situations qui permettent de déceler que les données ne représentent pas celles du monde réel qu'elles devraient représenter.

La non-qualité des données ne prend pas sa source dans des systèmes spécifiques. Elle peut être présente partout où il y a présence de données. Cependant, il est possible de déceler des problèmes de non-qualité, lorsqu'il est nécessaire de reprendre du travail, de mettre en place des activités de correction des données, de gérer les plaintes des clients, de faire l'objet de retour, etc.

Les concepts d'évaluation de la non-qualité des données ont dû évoluer avec le temps afin de s'adapter aux différents besoins des entreprises et à la réalité à laquelle elles doivent faire face. Dans les années 60, une des premières approches pour mesurer la qualité des données consistait à créer des modèles mathématiques théoriques qui servaient à connaître la duplicité des données statistiques. Viennent ensuite les années 80 où l'on mettra l'accent sur les méthodes pouvant servir à contrôler les données manufacturières comprises dans les systèmes afin de détecter et d'éliminer les problèmes liés à la qualité des données. Ensuite, c'est seulement au début des années 90 que les informaticiens ont commencé à considérer les problèmes de qualité des données en des termes tels que : définir, mesurer et améliorer la qualité des données numériques comprises dans les systèmes. Ces systèmes peuvent être des entrepôts de données, des bases de données ou tout autre système d'information existant [2].

De façon générale, la non-qualité des données occasionne des coûts financiers supplémentaires résultant de certaines situations, telles que la reprise de travail, la non-disponibilité des données requises au moment voulu et la perte de confiance. Ce qu'English nomme « le coût

des processus »[3]. Elle peut aussi, tout simplement, engendrer des pertes de revenus par des processus non concurrentiels et par la perte de contrats vu l'absence de données de qualité. Ce qu'English nomme « le coût des opportunités »[4] [4].

1.4 Évaluation de la qualité des données

L'évaluation de la qualité des données est constituée de mesures qui sont utilisées afin de parvenir à connaître le niveau de qualité des données comprises dans les différents systèmes, habituellement une base de données, un centre de données ou un entrepôt de données. Les autres sources, telles que des fichiers plats, par exemple Microsoft Excel ou Microsoft Project, peuvent également être mesurées. Ces mesures sont effectuées en regard de certaines caractéristiques, c'est-à-dire les dimensions, qui auront été définies préalablement à cette étape. Les principaux composants de l'évaluation de la qualité des données sont les dimensions et les mesures associées à chacune d'elle.

Les processus de contrôle et de mesures qui permettent d'évaluer la qualité des données sont inclus dans un grand processus qui vise à mettre en place une approche globale afin de permettre à l'entreprise d'atteindre le niveau de qualité souhaité pour leurs données. Dans l'approche de McGilvray, ces processus se situent en troisième position [16], tel qu'illustré sur la figure 1.5. Il est nécessaire que ces processus soient réalisés dans les premières étapes puisque ce sera à partir des informations recueillies lors de ces étapes qu'il sera possible de définir et d'évaluer avec certitude les activités suivantes. Il est essentiel de noter que les méthodes permettant d'évaluer la qualité des données pourront être réutilisées ultérieurement dans le processus général. C'est-à-dire qu'il deviendra nécessaire d'identifier la qualité des données avant et après la mise en place des autres activités. Il faut penser entre autres à mesurer la qualité des données avant l'implantation des processus de contrôle qui préviendront

---

[3] Traduction libre de « process costs »
[4] Traduction libre de « opportunity costs »

les problèmes de qualité futurs de celles-ci, ensuite, mesurer la qualité des données, une fois les processus et méthodes correctement implantés. L'analyse de ces mesures permettra de connaître l'impact du projet sur la qualité des données et d'en évaluer sa rentabilité.

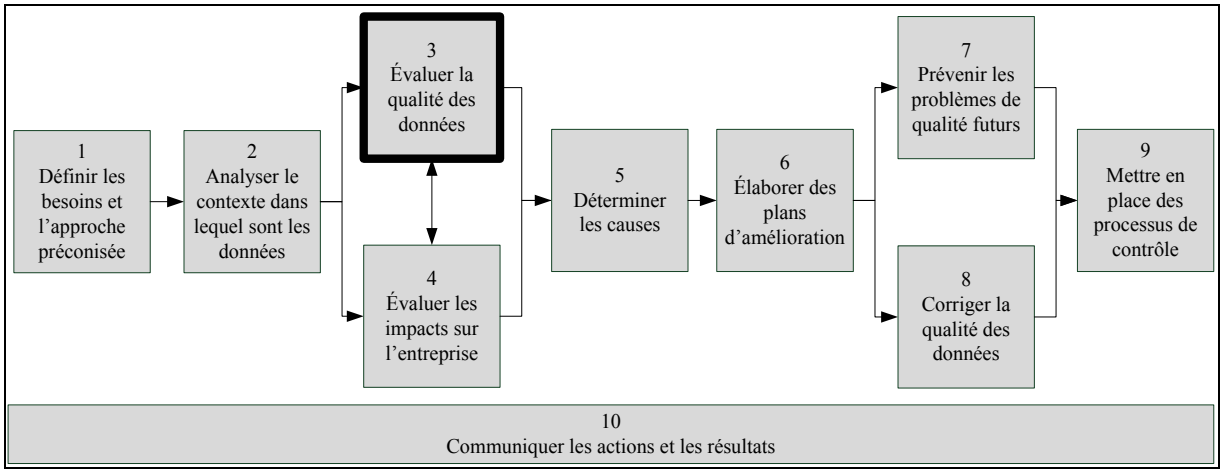


Figure 1.5 Les 10 étapes en qualité des données

Traduction libre
Source : McGilvray, D., (2008), p. 108

L'étape d'évaluation de la qualité des données comprend trois activités. La première étant la sélection des dimensions pertinentes, en second, vient la mesure de la qualité des données associées à chacune de ces dimensions puis en troisième, si plusieurs dimensions ont été mesurées, il faut procéder à la synthèse des résultats obtenus [16]. C'est donc en lien avec la première activité, la sélection des dimensions pertinentes, que le contenu de cet essai prend tout son sens.

1.5 Les dimensions en qualité des données

Une dimension en qualité des données correspond à une caractéristique d'une donnée qui permet de la classifier et d'en définir les besoins au niveau de sa qualité. Les dimensions sont principalement utilisées pour définir, mesurer et gérer la qualité de celles-ci [16]. Elles

permettent entre autres d'aborder la qualité des données sous un certain angle. Elles sont représentées par des termes, tels que validité, complétude ou précision. L'annexe 3 présente une liste non exhaustive des dimensions répertoriées dans la littérature. Une dimension est définie par un ensemble d'attributs qui représentent un certain aspect de la donnée [25]. La définition attribuée à chacune des dimensions peut être de nature théorique ou de nature opérationnelle [1]. Une suite logique est d'attribuer une définition théorique lors de l'élaboration du projet et ensuite, lorsque les utilisateurs et les différents intervenants auront contribué au projet, les définitions seront revues afin de correspondre davantage au monde réel, soit la nature opérationnelle des données.

Depuis plusieurs années, certaines entreprises ont mis en place des pratiques et des techniques de mesure et de contrôle en qualité des données. Certaines de ces entreprises ont rendu disponible sur le web la méthode qu'elles ont mise en place et qui leur a permis de planifier, d'évaluer, d'analyser, de mesurer et de contrôler la qualité de leurs données. Il existe également des auteurs qui ont publié des articles et des livres portant sur certains aspects liés à cette qualité. Notons entre autres les documents de l'ICIS (Institut canadien d'information sur la santé), l'IMF (*International Monetary Fund*), l'ICES (*Institute for Clinical Evaluative Sciences*) et l'EPA (*United States Environmental Protection Agency*).

De manière générale, la littérature mentionne l'importance des dimensions dans un projet en qualité des données. Les dimensions sont un aspect essentiel à la qualité des données. Elles procurent une façon de mesurer et de gérer la qualité des données et des informations [16].

Vu l'importance accordée aux dimensions, il va de soi que procéder à une sélection des dimensions pertinentes revêt une importance primordiale. D'autant plus que la littérature contient un très grand nombre de dimensions ayant été décrites au fil des années. Ces dimensions demeurent un atout essentiel pour procéder à une analyse qualitative des données. La sélection des dimensions est le point de départ de tout processus lié à la qualité des

données. Ce sont les dimensions sélectionnées qui permettront d'établir des concepts plus matures en vue de définir les caractéristiques d'une donnée de qualité, de définir des mesures pertinentes et de mettre en place des processus de contrôle fiables [2].

Le contenu des textes formulés par ces auteurs permet de bien comprendre l'importance que peut avoir la sélection des dimensions dans un contexte de qualité des données. Ce contexte pourrait être décrit plus précisément comme étant les étapes permettant l'évaluation du niveau de qualité des données. Malgré l'importance associée à la sélection des dimensions, ces auteurs fournissent peu de pistes ou de techniques permettant d'en faire la sélection. Il est habituellement laissé à la discrétion de l'équipe chargée de mettre en place ou de gérer les processus de contrôle de qualité des données. Cette sélection est la plupart du temps faite de manière intuitive [24].

Une autre des problématiques se rapportant aux dimensions consiste en une divergence des définitions pouvant être associées à chacune de ces dimensions. Ces définitions sont habituellement étroitement liées au contexte de l'entreprise dans laquelle elles seront utilisées. Le tableau 1.2 représente bien ces divergences en ce qui a trait à la dimension de l'exhaustivité.

Tableau 1.2 Définitions de la dimension « exhaustivité »

Référence	Définition
Wand 1996	La capacité d'un système d'information pour représenter tous les états importants du système du monde réel devant être représentés.
Wang 1996	La mesure permettant de connaître l'ampleur, la profondeur et la portée suffisante pour représenter adéquatement la tâche à accomplir.
Redman 1996	Le niveau de représentation des données comprises dans la collection de données.
Jarke 1999	Le pourcentage de l'information du monde réel qui est compris dans les sources ou les entrepôts de données.
Bovee 2001	Qualifie l'information ayant toutes les pièces nécessaires pour représenter une entité.

Référence	Définition
Naumann 2002	Il est le quotient du nombre de valeurs non NUL dans une source et la taille de la relation universelle.
Liu 2002	Toutes les valeurs qui sont censées être collectées selon la théorie de collecte utilisée.

Traduction libre
Source : Batini, C., (1998), p. 41

Tout en considérant les critères précédents, les dimensions sont également sujettes à une fluctuation du niveau d'importance qu'elles peuvent avoir en regard du type d'entreprise dans laquelle elles seront utilisées et selon le point de vue qu'aura l'utilisateur final sur les données qu'il se doit de manipuler. Il faut penser au-delà des dimensions telles que la précision et l'intégrité afin d'inclure les autres aspects des données qui peuvent dans certains cas avoir une plus grande importance pour les utilisateurs. Les dimensions telles que la disponibilité et la crédibilité peuvent faire partie de ces considérations [24].

Certains auteurs, en utilisant une approche empirique, tendent à regrouper les dimensions en catégories. Ces catégories sont généralement au nombre de quatre. Intrinsèque, contextuel, représentationnel et accessibilité [2]. Ces catégorisations peuvent parfois porter à confusion puisque certaines dimensions peuvent être présentes dans plus d'une catégorie. Il s'agit tout de même d'une approche intéressante qui peut simplifier la sélection des dimensions. La figure 1.9 illustre bien cette catégorisation.

Les dimensions sont utilisées afin de permettre une meilleure gestion de la qualité en ayant pour effet de canaliser les efforts dans la direction que doit prendre le projet en qualité des données puisque ce sont elles qui vont faire en sorte de définir ce qui doit être mesuré. Chacune des dimensions représente des aspects spécifiques des données. Plus précisément, un type de défaut que peuvent contenir les données [25]. Les dimensions constituent par le fait même un facteur essentiel à la réussite d'un tel projet. C'est en utilisant les dimensions qu'il est possible d'atteindre efficacement les buts recherchés. Comme dans le cas d'un service à la

clientèle, la dimension de la disponibilité « *timeliness* » peut être considérée comme très importante puisqu'elle couvre un des aspects majeurs de ce service. Elle permet de représenter le temps que prend une donnée à devenir disponible au service à la clientèle une fois la transaction effectuée. Par exemple, si le système nécessite un délai de 48 heures pour que la donnée soit disponible, il peut être juste de croire que si le client contacte le service à la clientèle pendant cette période, il ne sera pas en mesure d'obtenir les renseignements qu'il désire. Bien entendu, les efforts seront concentrés sur les dimensions qui auront été identifiées de grande importance, mais ceci ne vient pas amoindrir la place que peuvent avoir les autres dimensions. Toutefois, il permet de mettre en place des mesures efficaces qui permettront de contrôler adéquatement cet aspect de la donnée. D'autres dimensions pourront être utilisées ultérieurement pour évaluer la qualité des données sous d'autres angles, et ainsi améliorer la qualité globale des données.

C'est donc dire que chaque dimension nécessitera des outils, des techniques et des processus qui lui sont propres. Elles nécessitent également une somme d'efforts différents et permettent d'atteindre des buts distincts [16]. Les dimensions sont au cœur de toutes les activités liées à la qualité des données. Toutes les étapes qui gravitent autour utilisent les dimensions sélectionnées afin de définir leurs caractéristiques. Il peut s'agir, du contenu des tests de validation, des mesures, des rapports découlant de ces tests, des actions correctives à mettre en place, des données visées. La figure 1.6, illustre le caractère central des dimensions.

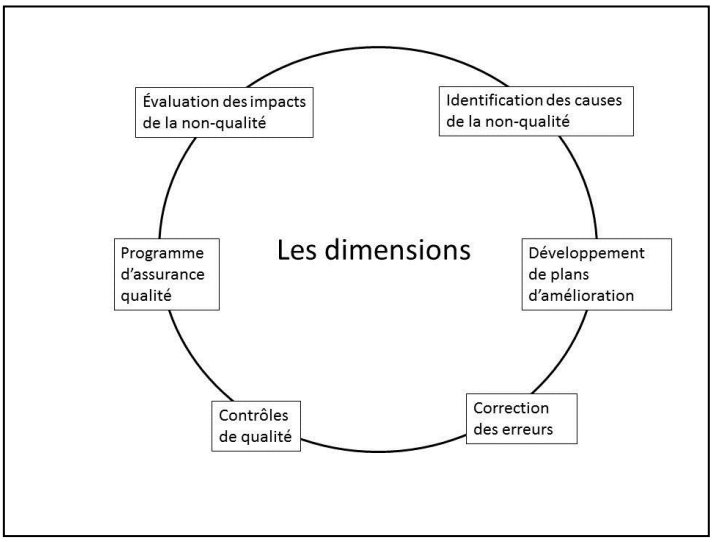


Figure 1.6 Les dimensions au centre des processus qualité

Plusieurs organismes, entreprises, individus ou groupes d'individus ont défini des cadres de référence qui permettent de mettre en place des principes de gestion de la qualité des données selon une méthode clairement définie. Chacun de ces cadres contient des solutions qui lui sont propres, mais plusieurs contiennent également des méthodes communes qui peuvent, de ce fait, être considérées comme des incontournables en qualité de données. Le but recherché par ces cadres de référence est habituellement de contrôler la qualité des données comprises dans les systèmes d'information. Pour en arriver à ce contrôle, plusieurs activités sont nécessaires. Comme ce ne sont pas tous les cadres de référence qui utilisent les dimensions ou le terme dimension pour parvenir à gérer la qualité des données, les cadres de référence décrits dans les pages qui suivent ne sont pas tous basés sur les dimensions. Malgré l'absence de ce terme, ils constituent des avenues intéressantes en fournissant plusieurs idées et concepts sur le fonctionnement de certains principes répandus en qualité de donnée. Voici un bref survol de certains cadres de référence.

1.5.1 TDQM

Cette méthode datant des années 90 a été développée au MIT (*Massachusetts Institute of Technology)*. Elle est entre autre basée sur les quatre étapes de la roue de Deming. Ces étapes sont la planification, la réalisation, la vérification et la réaction. La roue de Deming est principalement utilisée dans un processus d'amélioration continue. Dans le cadre de cette méthode, les quatre étapes ont été remplacées par : définir, mesurer, analyser et améliorer. Deming est considéré comme étant le fondateur de la méthode TDQM [9], mais aussi, de manière plus étendue, le fondateur des principes utilisés en qualité de données. La première étape de cette méthode comprend des phases telles que : 1) identifier les dimensions clés; 2) donner des définitions précises et significatives pour chacune des dimensions; 3) définir les mesures de ces dimensions; 4) développer un algorithme pour calculer la qualité des données [24]. Quinze dimensions sont utilisées et réparties en quatre catégories (intrinsèque, accessibilité, contextuelle et représentationnelle). Certaines dimensions sont présentes dans plus d'une catégorie.

Cette méthode est donc étroitement liée à l'utilisation des dimensions dans un contexte de qualité des données. Les dimensions sont à la base des principes mis de l'avant dans cette méthode. Elles sont utilisées dans plusieurs étapes du processus afin de parvenir à cibler les données devant être mesurées et pour lesquelles l'entreprise pourra en retirer le plus de bénéfices. La partie concernant la sélection des dimensions sera décrite un peu plus loin.

1.5.2 proDQM

L'approche proDQM (*proactive Data Quality Management)*, basée sur le TQM (*Total Quality Management)*, est constituée de deux étapes principales. La première étape consiste en une planification de la qualité, qui permet de répertorier les requis et les attentes vis-à-vis des données, ainsi que les critères de qualité, la classification et la priorisation des processus. La

seconde étape consiste en une élaboration des contrôles de qualité à proprement parler, ce qui inclut une vérification des procédés de délivrance des données afin de savoir s'ils respectent les spécifications établies précédemment. proDQM inclut également des mesures qui permettent de connaître la qualité des données [8]. Cette approche contient cinq catégories qui définissent différentes manières d'évaluer la qualité des données en regard de leur nature. Les cinq catégories sont :

- la vue transcendante,
- la vue basée sur les produits,
- la vue basée sur les utilisateurs,
- la vue basée sur la production et
- la vue basée sur la valeur.

Ces cinq catégories peuvent être utilisées selon une méthode séquentielle. Cette méthode consiste à définir, dans un premier temps, les requis pour le niveau de vue basée sur les utilisateurs. En partant de ces requis, la vue basée sur les produits pourra être définie ce qui aura pour effet de poursuivre à la vue suivante, soit celle basée sur la production et les processus. L'auteur dérive de ces trois niveaux de vue, deux facteurs de qualité. Soit le facteur qualité de conception et le facteur qualité de conformité [8]. La qualité de conception est utilisée pour savoir si les requis sont adéquatement représentés dans la conception du produit et par les spécifications de celui-ci. Tandis que le facteur qualité de conformité est utilisé pour évaluer la différence comprise entre le produit final et les spécifications d'origine.

Cette méthode utilise une approche basée sur les vues plutôt que sur les dimensions. L'auteur considère que l'utilisation des dimensions est trop théorique et conceptuelle en étant basée sur le niveau interne des systèmes. Cette approche cherche à inclure l'aspect subjectif des utilisateurs en ce qui a trait aux requis en qualité des données [8]. Toutefois, il utilise les termes actualisation, interprétable, utilisable et plausible pour décrire certains aspects des

données, utilisés notamment pour mesurer la qualité des données. Ce qui s'apparente beaucoup à la terminologie utilisée pour nommer les dimensions.

1.5.3 TIQM

L'auteur de cette méthode, Larry English, a entre autre été inspiré par le livre de Deming « *Out of the Crisis* », qui contient une liste de 14 recommandations en gestion. Six Sigma constitue une influence majeure dont s'est inspiré English pour créer une correspondance de ses processus sur les méthodes : 1) définir; 2) mesurer; 3) analyser; 4) améliorer; 5) contrôler. Cette méthode se caractérise par cinq étapes distinctes de mesures et d'amélioration ainsi que par un procédé général qui permet de suivre et de gérer l'ensemble du projet. Tel qu'illustré sur la figure 1.7, English utilise des catégories pour identifier les problèmes de qualité. Dans chacune de ces catégories, il utilise des attributs qui semblent très similaires aux dimensions. Toutefois, il n'utilise pas le terme dimension. Les attributs tels que la précision, la cohérence, la validité, la complétude et l'unicité sont fréquemment utilisés par English, principalement dans le processus « *P2 Assess Information Quality* » [5]. Il mentionne qu'il ne s'agit pas d'un processus séquentiel nécessitant de débuter par la première étape, mais qu'il est tout à fait possible de démarrer le projet depuis n'importe quelle étape.

Cette méthode offre certains avantages tels que la possibilité d'obtenir une formation. Elle est aussi accompagnée d'une documentation et elle peut être utilisée avec n'importe quels logiciels, techniques ou outils de qualité. Elle prétend pouvoir couvrir l'ensemble des processus requis en gestion de la qualité.

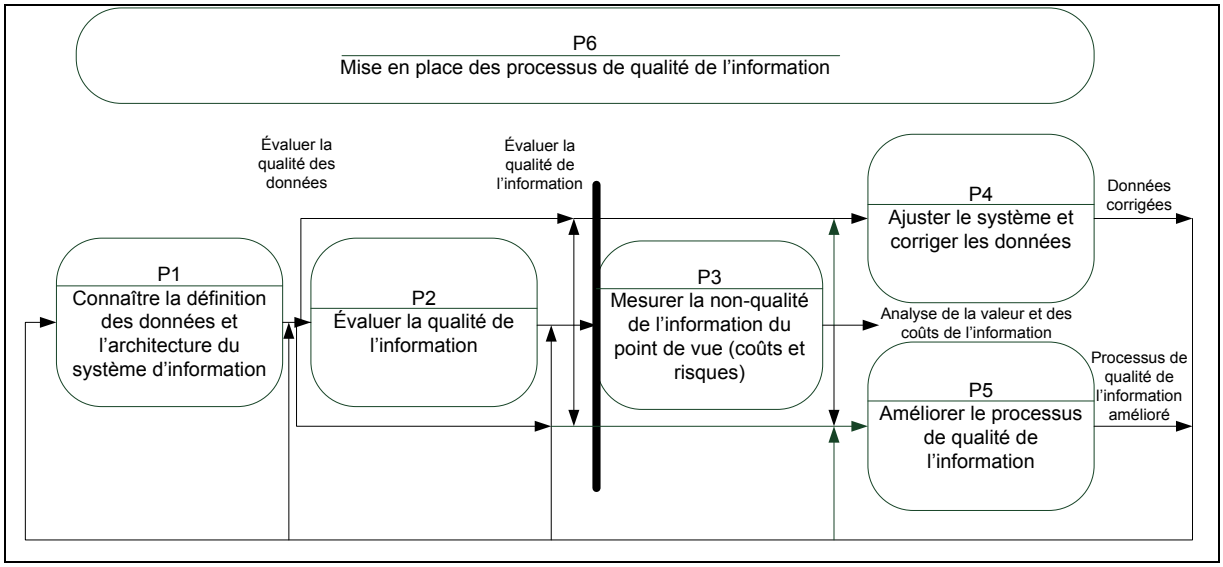


Figure 1.7 Les 6 processus du TIQM

Traduction libre
Source : English, L., (septembre 2002), p. 5

1.5.4 DAMA-DMBOK

L'association (*DAMA International*) est une organisation internationale qui a été mise sur pied dans le but de devenir une référence majeure pour tous les professionnels spécialisés dans la gestion des données. Cette association propose des standards qui vont permettre d'uniformiser les termes, les méthodes et les outils utilisés en gestion de données. Ces notions sont comprises dans un ouvrage publié en 2009, qui se nomme « *The DAMA guide to the Data Management Body of Knowledge* ». La préparation de ce corpus des connaissances a été basée sur deux autres corpus : le PMBOK (*Project Management Body of Knowledge*) et le SWEBOK (*Software Engineering Body of Knowledge*), publié par IEEE (*Institute of Electrical and Electronics Engineers*) [21]. Le corpus des connaissances de DAMA peut être accompagné par l'ouvrage « *The DAMA Dictionary of Data Management* ». Ce dictionnaire fournit environ 800 définitions de termes utilisés en gestion de données.

DAMA utilise plusieurs dimensions qui permettent de mesurer la qualité des données afin d'obtenir des évaluations de haut niveau. Ces dimensions sont au nombre de onze. Afin de mesurer la qualité des données associées à ces dimensions, DAMA utilise six mesures. Le résultat de ces mesures permet de connaître, en regard des dimensions, les endroits où il y a présence de données de mauvaise qualité. La roue de Deming est ici utilisée pour définir l'approche de gestion de la qualité des données. C'est pourquoi chacune des douze activités présentées à la Figure 1.8 se situent dans l'un des cycles (planification, réalisation, vérification et réaction).

De façon plus générale, le guide « DAMA-DMBOK » a pour but de façonner un consensus pour toutes les fonctions liées à la gestion des données informatiques [21].

Le guide contient 430 pages qui permettent de couvrir de manières très étendues plusieurs aspects concernant la gestion des données. Un des aspects couverts est la gestion de la qualité des données. La figure 1.8 illustre en détail toutes les étapes et tous les intervenants dans les pratiques et les techniques de gestion de la qualité selon le DAMA. Il s'agit d'une approche qui semble très complète et basée sur des processus éprouvés. Il est intéressant de constater qu'à l'intérieur de ces deux documents, un grand nombre d'informations utiles à une bonne gestion de la qualité des données sont présentées. La collaboration d'un grand nombre de professionnels du domaine a permis la confection de ces ouvrages.

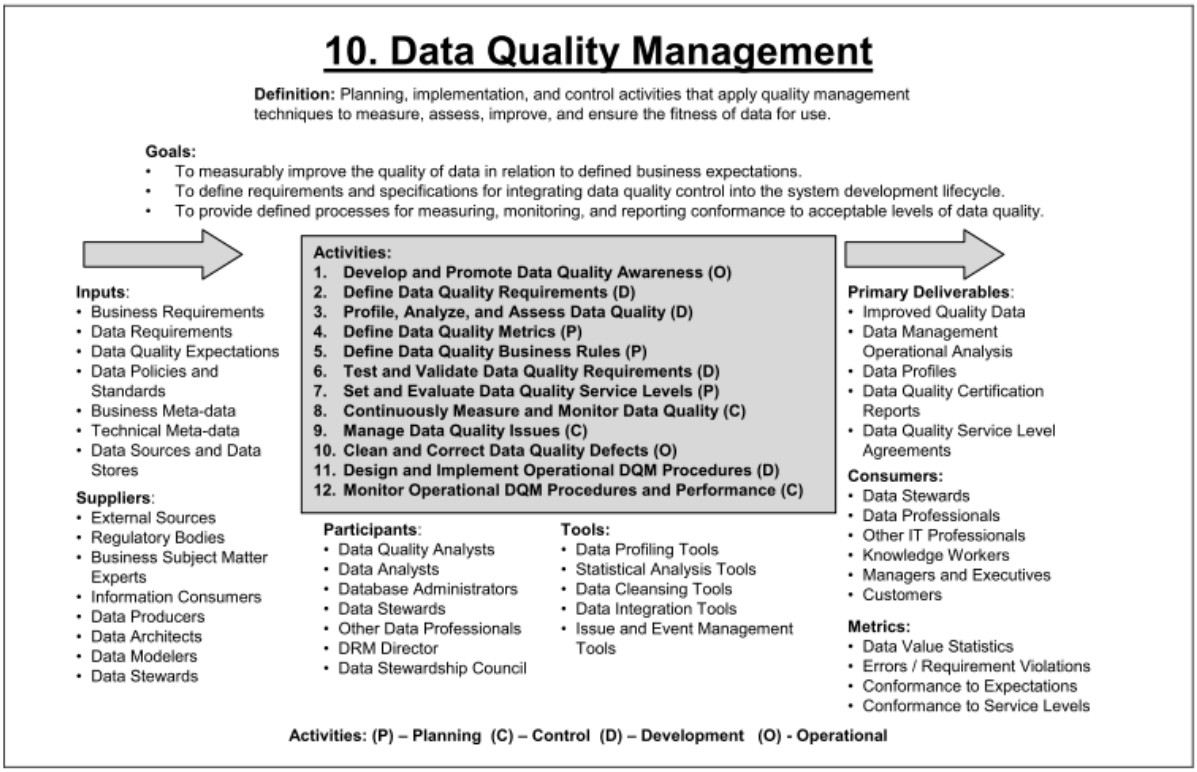


Figure 1.8 DAMA Gestion de la qualité des données

Source : *The DAMA Guide to The Data Management Body of Knowledge*, (2009), p. 292

1.6 Méthodes de sélection des dimensions

Puisque la problématique est d'abord liée à l'identification des dimensions qui sont jugées pertinentes à un projet de qualité des données, il a été nécessaire de prendre connaissance des techniques de sélection actuellement disponibles.

Avec les années, différentes approches ont été proposées par des auteurs. Ces approches ont des points communs puisqu'elles sont habituellement basées sur des auteurs d'importance pour le sujet, tels que Deming, Redman, Wang et Ballou, lesquels ont émis des hypothèses majeures sur le sujet. Voici un bref survol de ces approches.

1.6.1 TDQM

Cette méthode, élaborée par Richard Y. Wang du MIT propose une approche qui a fait intervenir près de 400 utilisateurs de données. Elle est composée de deux questionnaires qui, une fois complétés, ont fait l'objet d'une étude réalisée en deux étapes.

Le premier questionnaire a été rempli par 25 utilisateurs de données et 112 étudiants possédant de l'expérience en utilisation de données. Ce questionnaire a permis d'identifier 179 attributs de qualité. Le deuxième questionnaire a été confectionné à partir de ces 179 attributs. Il a ensuite été remis à 1500 finissants au MBA d'une université américaine, lesquels utilisent fréquemment les données pour la prise de décisions. Sur ces 1500 finissants, 355 ont retourné le questionnaire complété. Chacun des finissants devait attribuer une cote d'importance sur une échelle de 1 à 9 pour chacun des attributs, où « 1 » signifie très important et « 9 » pas important. Cet exercice a permis de diminuer la liste à 118 attributs en éliminant les attributs jugés « pas importants ».

Une fois ces deux questionnaires complétés, une analyse des résultats obtenus dans le deuxième questionnaire est réalisée. Il s'agit d'une étude de type empirique qui a pour effet de regrouper les 118 attributs en quatre familles de facteurs (intrinsèque, conceptuel, représentationnel et accessibilité). Ces familles de facteurs sont imposées par Wang. Ce regroupement a pour effet d'éliminer les attributs qui ne correspondent à aucune des familles de facteurs. Une fois regroupés, ces attributs sont maintenant nommés comme étant des dimensions. Les 20 dimensions ainsi identifiées font ensuite l'objet d'une analyse de facteur de stabilité. Celle-ci a permis d'éliminer les dimensions qui n'ont pas obtenu un facteur de stabilité concluant. C'est une fois ce travail complété que les 15 dimensions jugées les plus importantes sont connues. À cette étape, elles sont présentées dans un cadre conceptuel, tel que montré à la figure 1.9. C'est à l'aide de ce cadre conceptuel que les gestionnaires de données devront procéder à une hiérarchisation des dimensions qui reflètent le mieux leurs

besoins en qualité.

Wang mentionne qu'à la suite d'une analyse des pointages obtenus, il a été en mesure d'identifier les deux dimensions les plus importantes. Il s'agit de « précision » et « exactitude ». Toutefois, il met en garde les gestionnaires de données indiquant qu'à elles seules ces dimensions ne sont peut-être pas suffisantes pour couvrir les besoins des entreprises. C'est pourquoi il faut impérativement analyser l'impact que peuvent avoir sur la qualité des données les 15 dimensions présentées dans le cadre conceptuel [15][25].

Ce qui est intéressant avec cette méthode c'est qu'elle a permis de passer des 179 attributs initiaux à l'identification de 15 dimensions considérées comme étant les plus importantes par cette étude. Un autre aspect intéressant de cette méthode est dû au fait qu'elle prend racine à la suite d'un questionnaire remis aux utilisateurs. Ce qui semble être un bon départ pour représenter adéquatement le point de vue des utilisateurs. Toutefois, si les utilisateurs ont peu ou pas de connaissances en qualité des données, il pourrait en résulter un nombre très important d'attributs, lesquels nécessiteront beaucoup d'efforts pour en synthétiser la liste. Cette méthode nécessite en soi beaucoup de temps puisque plusieurs utilisateurs interviennent à différents moments dans le processus. Par contre, les résultats peuvent êtres intéressants puisqu'ils permettent de modifier la vision de l'équipe de qualité des données pour qu'elle soit davantage centrée utilisateur. En espérant que cette vision soit celle requise pour le projet.

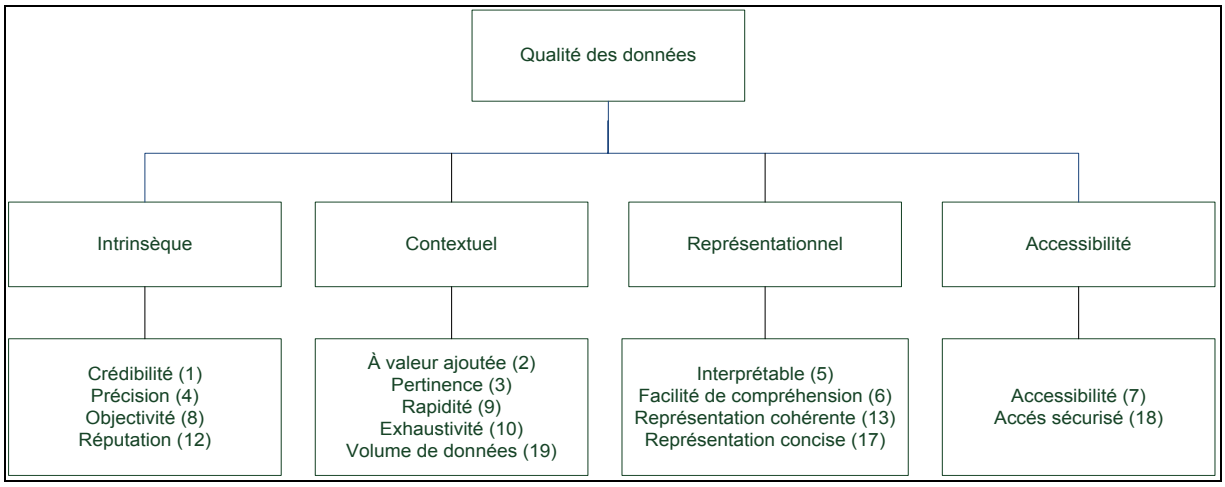


Figure 1.9 Regroupement des dimensions selon TDQM

Traduction libre
Source : Wang, R., Strong, D., (1996), p. 20

1.6.2 Danette McGilvray

Une approche publiée en 2008 utilise un court questionnaire basé sur 12 dimensions qui ont été préalablement sélectionnées par l'auteure. Pour chacune de ces dimensions, deux questions doivent être posées. 1) Dois-je évaluer ces données? 2) Puis-je évaluer ces données? Si la réponse à ces deux questions est oui, la dimension peut-être retenue pour passer à l'étape suivante. La première question sous-entend qu'il est pertinent de mesurer la qualité d'une dimension seulement si celle-ci permet d'atteindre les besoins d'affaires. La deuxième question sous-entend que cette dimension peut être mesurée avec la structure déjà en place. Elle sous-entend également que si le coût est trop élevé ou si les techniques requises sont non-disponibles, il n'est pas réaliste de penser pouvoir mesurer cette dimension. Une fois la sélection effectuée, il faut procéder à la mesure du niveau de qualité des données pour chacune de ces dimensions. Une fois les résultats connus, une deuxième étape est proposée. Cette deuxième étape sert à mesurer l'impact de ces dimensions sur l'entreprise. Les aspects qualitatifs et quantitatifs seront mesurés afin de connaître l'impact qu'a la qualité ou la non-

qualité des données sur l'entreprise [16].

Cette méthode est intéressante du fait qu'elle est très simple et ne nécessite pas l'intervention de plusieurs personnes. Toutefois, elle se limite aux 12 dimensions choisies par McGilvray. Ce qui ne sera pas nécessairement représentatif pour toutes les entreprises et leur contexte particulier. Il peut constituer un bon point de départ afin de savoir si la mise en place d'un projet de qualité des données est pertinente et peut apporter de la valeur à l'entreprise. Les gestionnaires de données pourraient ajouter des dimensions qu'ils jugeraient appropriées et qui sont absentes de la liste fournie par McGilvray. Cette technique de sélection est en somme très rapide et nécessite peu de ressources humaines. La partie qui constitue la mesure du niveau de qualité des données est certainement la partie la plus complexe et la plus délicate de cette méthode.

1.6.3 ICIS

L'Institut canadien d'information sur la santé est considéré comme un pionnier dans les techniques de gestion de la qualité des données dans le milieu de la santé [12]. Il a mis en place, dans les années 2000, un cadre sur la qualité des données. Ce cadre comprend un outil d'évaluation qui permet de mesurer et de documenter les limites et les forces comprises dans les banques de données de l'ICIS. L'Institut canadien d'information sur la santé mentionne que :

> « *Pour parvenir à obtenir de l'information de qualité supérieure, il faut déterminer les causes fondamentales des anomalies, prévenir les erreurs dans les processus d'information, établir les exigences en matière de qualité de l'information et contrôler les processus d'information.* » ([10], p. 3)

Cet outil d'évaluation utilise cinq dimensions pour lesquelles 19 caractéristiques et 61 critères ont été définis. Les dimensions utilisées sont l'exactitude, l'actualité, la comparabilité, la facilité d'utilisation et la pertinence. Pendant l'exercice, chacun des critères se voit assigner

une cote. Les cotes (respecté, non respecté, inconnu ou non applicable) sont celles généralement utilisées. Ces cotes sont ensuite utilisées afin de confectionner un plan d'action qui servira à corriger les lacunes. Ce qui est intéressant dans cette démarche c'est que les cinq dimensions utilisées ont été hiérarchisées afin de donner des niveaux d'importance différents à chacune des dimensions. Ce document attribue à la dimension « pertinence », le plus haut niveau d'importance puisqu'il mentionne que si les autres dimensions sont respectées, mais que la pertinence n'est pas au rendez-vous, alors la donnée est futile et de peu d'importance. La figure 10 illustre les cinq dimensions au sein du cycle de travail de la qualité des données, tel que représenté par l'ICIS.

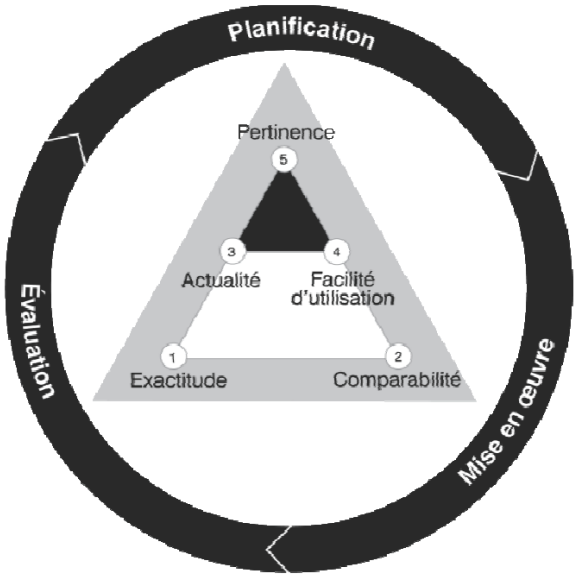


Figure 1.10 Cycle de travail de l'ICIS

Source : ICIS, (2009), p. 8

L'ensemble de la démarche proposée par l'ICIS a fortement inspiré le département de la santé de la Nouvelle-Zélande qui en a utilisé les fondements en y apportant quelques modifications afin de le conformer à leur contexte.

Cette approche est intéressante puisqu'elle est basée sur un petit nombre de dimensions. De plus, les dimensions font entre elles l'objet d'une certaine hiérarchisation. Ce qui permet de bien orienter les efforts en vue d'atteindre les objectifs. Comme cette technique est conçue pour s'appliquer à un cadre bien précis (la santé), il est difficile de critiquer le nombre de dimensions résultantes avec le choix restreint que ça implique. Ces dimensions étant destinées à un but bien précis. La hiérarchisation permet de mettre en place des pratiques et des techniques évolutives qui nécessitent de bien maîtriser une dimension avant de passer à la suivante. Ceci peut faire en sorte d'obtenir une grande maîtrise pour chacune des dimensions et ainsi en retirer un apport maximal.

1.7 Conclusions

Les fondements conceptuels ont maintenant été définis, mais plus particulièrement la nécessité de procéder à une sélection des dimensions qui seront les plus enclines à atteindre les objectifs du projet de qualité des données. Les chapitres qui suivent fourniront un complément d'information sur ce sujet. Tels que présentés, les critères permettant de procéder à cette sélection doivent prendre en considération certains facteurs, tels que le cycle de vie des données, le type de données manipulées et le type d'entreprise dans lequel nous œuvrons. Cette sélection peut être assistée par les techniques présentes dans la littérature, dont celles présentées dans ce chapitre. Puisque les définitions données aux dimensions varient d'un auteur à l'autre et que les termes utilisés pour nommer ces dimensions sont aussi divergents, le chapitre 2 présente une analyse de la littérature.

Chapitre 2

Analyse de la littérature

Les grands principes qui régissent la qualité des données étant maintenant connus, il faut procéder à l'analyse des dimensions présentes dans la littérature. Ceci constitue une des étapes utilisées en vue de connaître celles qui auront le plus d'impacts favorables à l'atteinte des objectifs du projet en qualité des données. Voici plus en détail les facteurs qui ont été considérés lors de cette analyse, la méthode utilisée et les résultats de cette analyse.

Vu l'importance qu'elles ont, la sélection d'une ou des dimensions devient un moment décisif du projet puisque tous les efforts futurs seront réalisés dans le but d'atteindre les spécifications définies par ces dimensions. Dans son livre sur la qualité des données, Batini mentionne l'importance de définir ces dimensions dès le début du projet : « *Choosing dimensions to measure the level of quality of data is the starting point of any DQ-related activity.* »[5] ([2], p.12). Cette sélection n'est pas aussi simple qu'il pourrait paraître de prime abord. Il existe de multiples dimensions, telles que présentées à l'annexe 3, et pour chacune de ces dimensions, comme citées précédemment, il existe également plusieurs définitions qui permettent de mieux refléter les différents contextes représentés par les données. Cette sélection doit être effectuée en connaissance de cause. Malheureusement, il n'existe pas de méthode formelle qui permet de procéder à la sélection. Certaines dimensions, telles que la précision (*accuracy*) et l'exactitude (*correctness*) peuvent être considérées prioritaires par un bon nombre d'utilisateurs finaux [8]. Mais ce ne sont pas tous les utilisateurs ou tous les gestionnaires qui vont avoir une même vision sur le choix des dimensions pouvant être jugées prioritaires pour l'entreprise.

---

[5] Choisir les dimensions afin de mesurer le niveau de qualité des données est le point de départ de toute activité liée à la qualité des données. (Traduction libre)

Dans un contexte de gestion de la qualité des données, l'aspect visant la sélection des dimensions est étroitement lié aux définitions que peut avoir chacune de ces dimensions. Ces définitions sont sujettes à une variation en fonction du point de vue utilisé pour les définir. Par exemple, un point de vue de conception en système d'information ou un point de vue qui vise à analyser la qualité des données comprises dans un système peuvent être à la source de cette discordance. Ce que Batini nomme une définition « théorique » et une définition « opérationnelle » [1]. Un autre aspect pouvant impacter les définitions concerne le type d'entreprise dans lequel seront utilisées les dimensions. S'il s'agit d'une entreprise de type manufacturière, les principales données concerneront entre autres choses, des inventaires, des fournisseurs, des processus, tandis que dans le cas d'une institution d'enseignement, les données serviront plutôt à gérer des programmes, des cours, des enseignants et des étudiants. Pour sa part, un concessionnaire automobile aura d'autres besoins pour la gestion de ses données qui seront de l'ordre du service à la clientèle, du suivi des rappels et de la couverture des garanties. Il va sans dire que le point de vue utilisé pour définir les dimensions est grandement influencé par le secteur d'activité de l'entreprise. Ce que Batini nomme la nature contextuelle de la qualité [1]. Dans le présent ouvrage, les définitions données à chacune des dimensions sont puisées dans la littérature. Il s'agit des définitions généralement admises pour chacune d'elle.

Ce chapitre présente donc une analyse des dimensions présentées dans la littérature.

2.1 Approche d'analyse

Pour permettre d'identifier les dimensions qui semblent les plus importantes en regard de la littérature, certaines étapes ont été nécessaires. La première étape consiste en une compilation des dimensions présentes dans la littérature. Une fois cette compilation terminée, un regroupement des dimensions ayant des définitions similaires a été effectué. À la suite de ce

regroupement et de la compilation des fréquences, une hiérarchisation des dimensions est établie.

2.2 Dimensions de la littérature

Tel que mentionné précédemment et tel que présenté à l'annexe 3, la littérature regorge de qualificatifs utilisés pour nommer les dimensions. Plusieurs de ces qualificatifs peuvent parfois sembler redondants et certains peuvent être d'une utilisation rarissime.

Certaines dimensions, telles que l'intégrité, l'intégralité, l'exactitude et la complétude sont fréquemment citées comme étant des incontournables en qualité de données. Mais ceci ne veut pas dire qu'elles conviennent à toutes les situations ou qu'il s'agisse des dimensions les plus pertinentes. D'autres dimensions peuvent s'avérer plus appropriées pour un projet donné.
Dans un premier temps, il a été question de compiler les dimensions présentes dans la littérature en répertoriant les termes et les définitions données par les auteurs ainsi que la référence à l'auteur et la fréquence de citation de chacune des dimensions répertoriées.

À l'aide de cette compilation, un regroupement de ces multiples termes a été effectué afin de jumeler sous un même terme les dimensions jugées similaires. Ainsi, des termes généraux ont pu être définis lesquels se nomment « dimension résultante ». Ces regroupements ont été effectués en regard des définitions données par les auteurs pour chacune des dimensions qu'ils ont citées. Voici un exemple qui illustre le genre de regroupement qui a été effectué. Wang définit le terme crédibilité ainsi : la mesure avec laquelle les données sont considérées comme vraies et crédibles. Il définit aussi le terme réputation de cette manière : une mesure avec laquelle les données sont hautement considérées en termes de source et de contenu. De son côté, Batini définit le terme confiance comme étant un indice de la qualité de la source. Les définitions données à ces trois termes sont très similaires puisqu'ils décrivent une même vue sur les données. C'est pourquoi elles ont été regroupées sous la dimension résultante

« perception, pertinence et confiance ».

C'est à la suite de ce regroupement qu'il a été possible de connaître la somme des fréquences de citation de chacune des dimensions résultantes. C'est la valeur de cette fréquence qui sera utilisée pour connaître le niveau d'importance que la littérature accorde à chacune des dimensions. La fréquence signifie le nombre de citations dont la dimension a fait l'objet. Il est à noter que les termes ont été comptabilisés une seule fois par ouvrage. C'est-à-dire que si un ouvrage cite plusieurs fois le terme « actualisation », il a été comptabilisé une fois. Le résultat de cette compilation est présenté au tableau 2.1 en ordre décroissant de fréquence.

Ce tableau constitue une version épurée de celui présenté à l'annexe 3. Dans cette annexe, la totalité des termes utilisés par les auteurs ainsi que les définitions et les références aux auteurs est présentée. Dans plusieurs des cas, la dimension résultante est une traduction libre de celui proposé par McGilvray [16]. Ce tableau synthétise plus de 90 entités « dimension – définitions » réparties dans plus de 20 références. Il est également présenté en ordre décroissant de fréquence. Les dimensions ayant obtenu la plus grande fréquence sont présentées au début du tableau. Il est à noter que les dimensions utilisées par English, auteur de l'approche TIQM, ont été comptabilisées. Mais aucune définition liée à English n'est apparente dans le tableau puisqu'aucune des définitions données par cet auteur n'a été rendue publique.

Tableau 2.1 Tableau des dimensions

Dimension résultante	**Fréquence**
Actualisation et disponibilité	12
Perception, pertinence et confiance	11
Exhaustivité	10
Cohérence et synchronisation	8
Précision	8

Dimension résultante	**Fréquence**
Interprétable	6
Principes de base de l'intégrité des données	6
Accessibilité	5
Duplication	5
Conformité	5
Facilité d'utilisation et adaptables	4
Sécurité	3
À valeur ajoutée	2
Objectivité	2
Portabilité	2
Représentation concise	2
Convenance	1
Décroissance des données	1
Format précis	1
Format flexible	1
Habileté à représenter des valeurs nulles	1
La variété des données et sources de données	1
Libre d’erreurs	1
Rapport coût-efficacité	1
Utilisation efficace de la mémoire	1
Transactionnel	1
Spécifications de données	1
Volatilité	1

2.3 Limites

Une première limite porte sur le nombre de références utilisées dans cette analyse. Un plus grand nombre de références permettrait probablement d'obtenir des résultats encore plus représentatifs du contenu de la littérature.

Une deuxième limite porte sur l'étape de regroupement des dimensions. Celle-ci a été réalisée par l'auteur de cet essai et a été validée informellement par les directeurs de cet essai. Aucune validation formelle des regroupements n'a été réalisée avec des spécialistes. Il est probable que d'autres personnes seraient arrivées à un regroupement différent.

2.4 Conclusion

L'analyse de la littérature a fourni une liste épurée des dimensions présentes dans la littérature. Cette liste épurée est le résultat d'un regroupement des dimensions en regard des définitions données par les auteurs. Ce regroupement a ensuite permis de hiérarchiser les dimensions selon leur fréquence de citation. Dans le but de comparer la théorie et la pratique, le prochain chapitre présente une analyse réalisée avec des praticiens en TI.

Chapitre 3

Analyse des praticiens

Ce chapitre présente une analyse des dimensions en qualité des données par les praticiens en technologies de l'information. Cette analyse a pour but de connaître le point de vue de praticiens quant à l'importance accordée à chacune des dimensions. Il y sera décrit la méthode utilisée pour sonder les praticiens et celle utilisée pour compiler ces résultats dans un format qui permet d'en faciliter l'analyse.

3.1 Cueillette de données

Cet exercice comprend un questionnaire remis à des étudiants du CeFTI, dans lequel il leur a été demandé d'associer des niveaux de priorité à chacune des dimensions. Ces niveaux de priorité seront établis sans égard pour la littérature. Ce qui signifie que les résultats seront seulement tributaires de la définition donnée à chacune des dimensions et des aptitudes de l'étudiant envers le domaine concerné soit : la qualité des données. La manière utilisée pour procéder à cette cueillette et la méthode d'analyse des données recueillies constituent un point crucial de cette analyse.

L'outil utilisé pour procéder à cette cueillette est un questionnaire dont une grande partie de son contenu est basé sur la compilation effectuée lors de l'analyse de la littérature.
Ce questionnaire a été confectionné en ligne avec les outils disponibles gratuitement sur le site de Google Docs. L'hyperlien permettant d'accéder au questionnaire a ensuite été inclus dans un courriel d'invitation transmis aux étudiants. La page web ainsi accédée peut être complétée en ligne par les étudiants et soumise immédiatement une fois complétée.

3.2 Description du questionnaire

La première partie du questionnaire contient une mise en situation qui est utilisée afin d'évaluer la pertinence des réponses obtenues en fournissant une connaissance démographique sur les répondants. Cette partie offre aussi l'avantage de permettre au participant de se détacher des tâches en cours et ainsi mieux se concentrer sur le questionnaire, une introduction en quelque sorte. Il pourra alors se forger peu à peu une idée du contenu qui sera abordé dans le corps du questionnaire. Plusieurs de ces questions étant en lien direct avec le sujet traité.

C'est dans cette partie que nous retrouvons deux brèves questions. La première étant « Depuis combien d'années travaillez-vous en TI? ». Ce qui permet de situer rapidement le positionnement du répondant dans un contexte TI. Vient ensuite « Quel est votre plus haut niveau académique en TI? ». Celle-ci permet encore là de positionner le répondant, mais cette fois-ci, en regard de son niveau d'études. Bien entendu, dans le cadre de cet essai, le questionnaire sera remis à d'autres étudiants de niveau universitaire. Dans le cadre normal des entreprises, les répondants pourraient avoir des niveaux scolaires variés.

Les résultats de cette section peuvent être interprétés ainsi. Dans un premier temps, si le répondant indique qu'il a très peu d'expérience en TI et qu'il ne connaît pas ce que sont les dimensions, il pourrait être à propos de considérer les réponses obtenues avec ce répondant comme ayant une faible valeur. Comparativement à un questionnaire dont le répondant indique qu'il a déjà travaillé sur un projet en qualité des données et qu'il a une très grande expérience en TI. Les réponses obtenues par ce questionnaire pourraient être considérées de grande valeur. Bien entendu, il s'agit d'une évaluation subjective, mais qui peut tout de même représenter de façon assez réelle la situation de chaque participant.

Les données recueillies dans cette mise en situation peuvent être ou ne pas être considérées ultérieurement. Leur utilisation est à la discrétion de l'enquêteur (celui qui émet le sondage). Les résultats obtenus avec cette partie du questionnaire peuvent être utilisés pour évaluer la pertinence globale de tous les questionnaires reçus ou pour évaluer individuellement chacun des questionnaires soumis. Cette deuxième approche peut être plus complexe à implémenter puisqu'elle doit utiliser un facteur d'ajustement qui sera utilisé pour corriger la pondération soumise par le répondant, et ce, pour chacun des questionnaires reçus. Dans ce contexte-ci, c'est l'utilisation d'une pondération globale qui a été retenue. Celle-ci est plus rapide à utiliser et permet d'avoir une évaluation globale des aptitudes des répondants envers le domaine de l'étude. Le chapitre 4 décrit l'analyse effectuée à partir de cette section du questionnaire.

La deuxième partie du questionnaire contient l'ensemble des 28 dimensions compilées dans la littérature, tel que présenté au tableau 2.1. Par contre, afin d'en faciliter la compréhension et d'en améliorer la lisibilité, seule une définition a été conservée pour chacune des dimensions. Le répondant doit sélectionner le niveau de priorité qui lui semble le plus approprié. Chaque dimension est pondérée en regard d'une échelle commune. Cette échelle est basée sur l'échelle de Likert. Les cinq échelons sont :

1. Très importante
2. Importante
3. Peu importante
4. Pas importante
5. Ne sait pas

L'échelle de Likert est très intéressante vu sa simplicité. Il s'agit d'une approche répandue pour ce genre de questionnaire. Elle est facilement compréhensible par les répondants et ne nécessite pas de connaissances particulières. Cette échelle offre des possibilités comportant habituellement de cinq à sept niveaux. En plus d'être simple pour les répondants, elle est aussi

très simple à administrer puisque les réponses sont comprises dans une plage de cinq valeurs. Ce qui a pour effet d'éliminer les réponses à développement. De plus, une seule réponse (sélection) est possible pour chacune des questions. L'annexe 4 contient l'intégralité du formulaire web qui a été distribué aux étudiants du CeFTI. Afin d'obtenir un questionnaire dont le contenu permet de représenter adéquatement le point de vue des étudiants, il a été nécessaire de procéder à deux itérations.

Dans un premier temps, il a fallu procéder à l'envoi d'un courriel aux étudiants de la cohorte de Sainte-Thérèse. Cette démarche a été utilisée comme « prétest » afin de valider la forme et la compréhension du questionnaire par ceux-ci. La compilation des résultats a permis de définir et d'affiner la méthode d'analyse des résultats. Le premier envoi a été effectué en mai 2011. Par la suite, une fois le questionnaire ajusté, le deuxième envoi qui constitue la version définitive a été transmis en juillet 2011 à tous les étudiants présents sur la liste de distribution du CeFTI.

3.2.1 Premier questionnaire transmis

Dans un premier temps, en guise de prétest, le questionnaire a été transmis à la liste de distribution de la cohorte de Sainte-Thérèse. Ce qui signifie environ 25 étudiants. Cet envoi a permis de récolter sept questionnaires complétés.

Le traitement des données recueillies par ce questionnaire a causé certains problèmes qui ont été constatés une fois le processus de compilation débuté. En effet, une échelle de type Likert contient habituellement un champ central ayant une valeur équivalant à « neutre (ou indifférent) » et un autre champ situé à l'une des limites de l'échelle qui équivaut à « pas important ». Dans le contexte de ce travail, l'utilisation de ces deux champs a fait en sorte que deux des réponses auraient pu être comptabilisées avec la même valeur. Dans d'autres contextes, qui ne sont pas ceux de cet essai, l'échelle de Likert serait tout à fait juste et

n'apporte pas nécessairement l'imprécision qui a été relevée après coup. Le champ « neutre » est habituellement utilisé par les répondants dans les cas où la situation ne s'applique pas ou pour laquelle le répondant ne veut pas se prononcer. Tandis que le champ « pas importante » devrait être utilisé par le répondant pour définir une situation où l'énoncé doit être qualifié de « pas importante » en opposition à l'énoncé « très importante ». Dans ce cas-ci, les répondants n'ont aucunement utilisé le champ « pas importante » tandis que le champ « neutre » a été utilisé 63 fois. Ceci laisse croire que cette valeur a été utilisée à outrance et que certains répondants auraient plutôt voulu indiquer « pas importante ». Une vérification en ce sens a été effectuée auprès des premiers répondants, lesquels ont confirmé l'ambiguïté lors de l'utilisation de ces champs. C'est à la suite de ces constatations qu'il s'est avéré judicieux de corriger l'échelle du questionnaire et de procéder à un second envoi.

Grâce aux résultats obtenus lors de ce premier envoi, la méthode d'analyse consistant entre autre à élaborer le tableau de compilation qui fournit les valeurs requises au calcul du coefficient de corrélation des rangs de Spearman a pu être confectionnée et testée. D'emblée, la présence de dimensions ayant un rang de valeur ex aequo a nécessité d'insérer deux colonnes supplémentaires au tableau. Une fois que tous les champs du tableau ont été compilés, les valeurs ont pu être insérées dans l'algorithme de calcul du coefficient. Le résultat obtenu a ensuite fait l'objet d'une analyse afin d'en déterminer sa pertinence à identifier les dimensions les plus importantes. C'est à la suite de cette analyse qu'il s'est avéré nécessaire de trouver des méthodes complémentaires pour interpréter les résultats du questionnaire de manière plus concluante.

3.2.2 Deuxième questionnaire transmis

Une fois la deuxième version du questionnaire corrigée, un envoi de masse a été effectué à tous les étudiants présents sur la liste de distribution du CeFTI. Les étudiants ont été sollicités par un courriel qui expliquait l'importance d'avoir un grand nombre de répondants et les buts

recherchés par la diffusion de ce questionnaire. Le traitement des données recueillies lors de ce deuxième questionnaire a permis d'obtenir un résultat beaucoup plus juste. En effet, puisque le nombre de répondants est supérieur au « prétest » et que l'échelle utilisée pour les réponses a été précisée, les réponses ainsi obtenues offrent une meilleure cohérence. L'envoi de cette deuxième version du questionnaire a permis de recevoir 10 questionnaires complétés. Il a tout de même été nécessaire de procéder à une relance auprès des étudiants afin d'augmenter le nombre de répondants. À la suite de cette relance, 7 questionnaires complétés ont été reçus. Ce qui totalise 17 questionnaires complétés par les étudiants.

3.3 Résultats

Une fois tous les questionnaires reçus, j'ai compilé les résultats afin d'en faciliter leur analyse et ainsi pouvoir en tirer des constats. La compilation de la première partie du questionnaire, qui est intitulée « description des répondants », est compilée et présentée dans le tableau 3.1. Dans un but de simplification, les réponses obtenues par cette section ont été compilées de manière globale pour l'ensemble des répondants. Cette pondération permet de qualifier la valeur de l'étude. Les totaux apparaissant au centre du tableau permettent de constater qu'une grande partie des répondants ont de bonnes connaissances en qualité des données puisque plus de 50 % d'entre eux sont en accord ou fortement en accord. Tandis que les totaux au bas du tableau permettent de constater qu'ils ont aussi une bonne expérience en TI puisque plus de 58 % d'entre eux ont une expérience supérieure à 10 ans de travail en TI et que les niveaux académiques en TI sont dans une proportion de 70,5 %.

Les résultats obtenus dans cette première partie du questionnaire sont très encourageants puisque les répondants sont majoritairement du domaine des TI et possèdent de bonnes connaissances envers le domaine concerné par l'étude. Ce qui permet de croire que le résultat de cette étude peut-être très représentatif d'une situation réelle dans un contexte d'entreprise.

Tableau 3.1 Description des répondants

	Poids				
Questions	**5**	**4**	**3**	**2**	**1**
Section : « Connaissances en qualité des données »	**Fortement en désaccord**	**En désaccord**	**Neutre**	**En accord**	**Fortement en accord**
J'utilise fréquemment, dans mes tâches courantes, des données comprises dans un système, tel qu'une base de données, un entrepôt de données…	1	1	2	3	10
Je possède de très bonnes connaissances en termes de « qualité des données ».	1	3	4	5	4
J'ai déjà fait partie d'une équipe ayant comme objectif la gestion de la qualité des données.	6	3	3	3	2
Je connais ce que sont les dimensions et à quoi elles servent.	2	1	6	5	3
Totaux :	**10**	**8**	**15**	**16**	**19**
Section : « Expérience TI »	**Je ne travaille pas en TI**	**0 à 5 ans**	**5 à 10 ans**	**10 à15 ans**	**15 ans ou plus**
Depuis combien d'années travaillez-vous en TI?	4	2	1	3	7
Section « Scolarité TI »	**Je n'ai pas d'études en TI**	**Collégial**	**Univer-sitaire (1er cycle)**	**Univer-sitaire (2e ou 3e cycle)**	**Autre**
Quel est votre plus haut niveau académique en TI?	1	0	2	12	2
Totaux :	**15**	**10**	**18**	**31**	**28**

La deuxième partie du questionnaire, qui est la raison d'être de celui-ci, a été elle aussi compilée sous forme de tableau, tel que présenté au tableau 3.2.

Les chiffres inscrits dans les colonnes (très importante, importante, peu importante, pas importante et ne sait pas) indiquent la fréquence des réponses. Par exemple, si le chiffre « 2 » est inscrit dans le champ « peu importante », ceci signifie que deux répondants ont qualifié

cette dimension de « peu importante ». La somme des réponses de chacune des dimensions correspond au nombre de répondants ayant participé à ce sondage.

Afin de quantifier les dimensions en regard du niveau d'importance qu'elles ont obtenue, il faut nécessairement donner une pondération à chacune des colonnes. En ce sens, le niveau d'importance « très importante » s'est vu attribuer à un poids de quatre tandis que le niveau « ne sait pas », un poids de zéro. Le poids total de chacune des dimensions est calculé en fonction du poids assigné à chacune des colonnes et de la fréquence indiquée dans chacune d'elles[6]. Par exemple, pour la dimension « à valeur ajoutée », le calcul du poids total est effectué ainsi : (9x4) + (8x3) = 60. C'est la valeur de ce poids qui sera utilisé dans la section traitant de l'analyse comparative, plus précisément lors du calcul du coefficient de corrélation des rangs de Spearman.

D'un coup d'œil rapide au tableau, il est facile de constater qu'aucune des dimensions n'a été identifiée comme « très importante » par tous les répondants. À l'opposé, aucune dimension n'a été totalement rejetée, soit identifiée comme « pas importante » par tous les répondants. Cette situation justifie une analyse plus approfondie des résultats afin d'en connaître la tendance. C'est ce que l'on verra au Chapitre 4.

Tableau 3.2 Compilation des résultats du sondage

Poids	**4**	**3**	**2**	**1**	**0**	
Dimension	**Très importante**	**Importante**	**Peu importante**	**Pas importante**	**Ne sait pas**	**Poids total**
Libre d'erreurs	13	3		1		62
Principes de base de l'intégrité des données	11	5	1			61

[6] Poids total = (fréquence colonne « très importante » * poids) + (fréquence colonne « importante » * poids) + (fréquence colonne « peu importante * poids) + (fréquence colonne « pas importante » * poids)

Poids	4	3	2	1	0	
Dimension	**Très importante**	**Importante**	**Peu importante**	**Pas importante**	**Ne sait pas**	**Poids total**
À valeur ajoutée	9	8				60
Actualisation et disponibilité	10	6	1			60
Sécurité	10	6	1			60
Interprétable	9	6	2			58
Facilité d'utilisation et adaptable	8	8		1		57
Accessibilité	6	9	2			55
Cohérence et synchronisation	8	7	1		1	55
Habileté à représenter des valeurs nulles	6	8	3			54
Conformité	6	7	4			53
Format flexible	5	9	3			53
Perception, pertinence et confiance	7	7	2		1	53
Exhaustivité	4	10	3			52
Précision	7	7	1		2	51
Portabilité	4	9	2		2	47
Rapport coût-efficacité	4	6	5	2		46
Convenance	1	11	3	2		45
La variété des données et sources de données	4	5	6	1	1	44
Transactionnel	3	8	4		2	44
Spécifications de données	4	8	1	1	3	43
Volatilité	1	10	3	3		43
Objectivité	2	8	4	2	1	42
Duplication	5	2	7		3	40
Représentation concise	2	5	8	1	1	40
Format précis	6	5			6	39

Poids	4	3	2	1	0	
Dimension	**Très importante**	**Importante**	**Peu importante**	**Pas importante**	**Ne sait pas**	**Poids total**
Utilisation efficace de la mémoire	1	6	5	3	2	35
Décroissance des données		8	4	1	4	33
Somme :	**156**	**197**	**76**	**18**	**29**	

3.4 Limites

Une limite est le nombre de répondants. En effet, il est raisonnable de penser que les résultats pourraient être plus précis avec un nombre plus élevé de répondants.

De plus, les résultats obtenus par le questionnaire auraient pu être différents si par exemple, les étudiants avaient eu la chance d'assister à une présentation sur la qualité des données ou faire quelques lectures sur la qualité des données, ce qui aurait pu leur fournir une base commune et peut-être leur permettre d'aborder de façon différente l'importance de chacune des dimensions. Des exemples de données auraient pu être fournis pour chacune des dimensions, ce qui aurait permis à l'étudiant de mieux visualiser la portée de chacune des dimensions. Toutefois, cette avenue n'a pas été utilisée afin de ne pas alourdir le processus puisque notamment, un des buts recherchés était que le questionnaire soit le plus succinct possible et qu'il puisse être facilement compréhensible par un vaste auditoire.

Chapitre 4

Analyse comparative

Le but de cette analyse comparative est de savoir s'il y a une concordance entre les résultats obtenus lors de l'analyse de la littérature et ceux obtenus lors de l'analyse des praticiens en TI. La première analyse porte sur la corrélation des rangs tandis que les trois analyses comparatives suivantes sont de type statistique.

4.1 Coefficient de corrélation des rangs de Spearman

Le coefficient de corrélation des rangs de Spearman permet d'établir un degré de corrélation entre deux ensembles de valeurs basé sur le rang qui leur est assigné. Dans ce cas-ci, le premier ensemble de valeurs est celui représentant les fréquences attribuées dans la littérature, tel que compilé dans le tableau 2.1 et le deuxième ensemble de valeurs correspond au poids total accordé par les praticiens, tel que présenté au tableau 3.2. C'est à partir de ces valeurs qu'il est possible d'assigner un rang à chacune des dimensions et ainsi procéder au calcul du coefficient de corrélation des rangs, tel que proposé par Spearman [20]. Spearman offre la possibilité d'utiliser deux formules. Puisque seule une de ces formules permet de traiter convenablement les valeurs lorsqu'il y a présence de valeurs ex aequo, c'est cette formule qui a été utilisée. Dans ce cas-ci, les valeurs ex aequo se retrouvent dans les colonnes « rang (X_i et Y_i) », tel que démontré au tableau 4.1. Chaque colonne doit être traitée séparément. Ce qui veut dire que si la colonne (X_i) contient des dimensions ayant un même rang, ce qui est le cas ici, il faut ajouter la colonne « rang moyen (x_i) » qui permet de calculer un rang moyen pour chacune des dimensions ayant le même rang. Ce rang moyen est ensuite assigné à chacune des dimensions ayant le même rang (X_i). Le même principe s'applique pour la colonne (y_i).

Afin de mieux comprendre le fonctionnement de cette formule, voici, quelques explications sur les principes qui la régissent. Le tableau 4.1 contient les données requises par le calcul de Spearman, tel que détaillé à l'annexe 2. Les valeurs surlignées en gris sont celles offrant la plus grande corrélation de rang.

Tableau 4.1 Coefficient de corrélation des rangs de Spearman

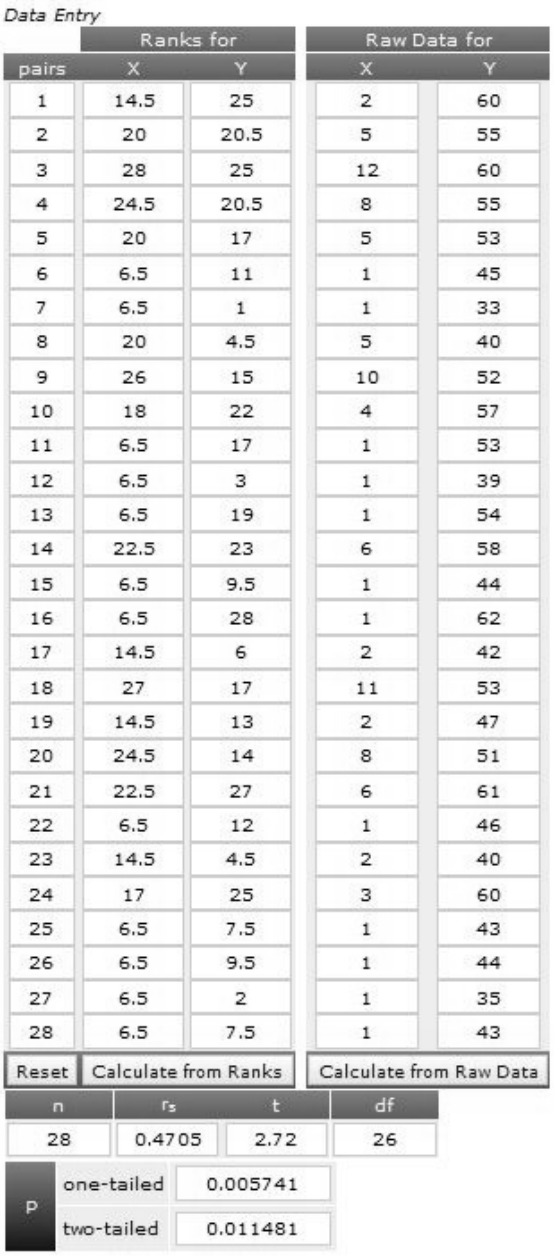
Dimension	Littérature	Praticiens en TI	Somme	Rang	Rang moyen (x_i)	Rang	Rang moyen (y_i)	Écart de rang	Écart de rang au carré
	X_i	Y_i		x_i	r_i	y_i	s_i	d_i	d_i^2
À valeur ajoutée	2	60	62	16	14,5	24	25	10,50	110,25
Accessibilité	5	55	60	21	20	20	20,5	0,50	0,25
Actualisation et disponibilité	12	60	72	28	28	25	25	3,00	9
Cohérence et synchronisation	8	55	63	25	24,5	21	20,5	4,00	16
Conformité	5	53	58	20	20	16	17	3,00	9
Convenance	1	45	46	12	6,5	11	11	4,50	20,25
Décroissance des données	1	33	34	11	6,5	1	1	5,50	30,25
Duplication	5	40	45	19	20	4	4,5	15,50	240,25
Exhaustivité	10	52	62	26	26	15	15	11,00	121
Facilité d'utilisation et adaptables	4	57	61	18	18	22	22	4,00	16
Format flexible	1	53	54	10	6,5	17	17	10,50	110,25
Format précis	1	39	40	9	6,5	3	3	3,50	12,25
Habileté à représenter des valeurs nulles	1	54	55	8	6,5	19	19	12,50	156,25
Interprétable	6	58	64	23	22,5	23	23	0,50	0,25
La variété des données et sources de donne	1	44	45	7	6,5	9	9,5	3,00	9
Libre d'erreurs	1	62	63	6	6,5	28	28	21,50	462,25
Objectivité	2	42	44	15	14,5	6	6	8,50	72,25
Perception, pertinence et confiance	11	53	64	27	27	18	17	10,00	100
Portabilité	2	47	49	14	14,5	13	13	1,50	2,25
Précision	8	51	59	24	24,5	14	14	10,50	110,25
Principes de base de l'intégrité des donnée	6	61	67	22	22,5	27	27	4,50	20,25
Rapport coût-efficacité	1	46	47	5	6,5	12	12	5,50	30,25
Représentation concise	2	40	42	13	14,5	5	4,5	10,00	100
Sécurité	3	60	63	17	17	26	25	8,00	64
Spécifications de données	1	43	44	4	6,5	7	7,5	1,00	1
Transactionnel	1	44	45	3	6,5	10	9,5	3,00	9
Utilisation efficace de la mémoire	1	35	36	2	6,5	2	2	4,50	20,25
Volatilité	1	43	44	1	6,5	8	7,5	1,00	1

La démarche chronologique utilisée pour compléter ce tableau est la suivante. Les valeurs comprises dans les colonnes « littérature (X_i) » et « praticiens en TI (Y_i) » proviennent respectivement de la littérature et du questionnaire. Ensuite les rangs (x_i) et (y_i) ont été déterminés. Pour ce faire, il est nécessaire de placer en ordre croissant les dimensions selon la fréquence qu'elles ont obtenue. Ensuite, en utilisant ce classement, un rang numérique

croissant est associé à chacune des dimensions. La même méthode doit être utilisée pour calculer le rang des données présentes dans la colonne « praticiens en TI (Y_i) ». Puisque les colonnes « littérature (X_i) » et « praticiens en TI (Y_i) » contiennent des valeurs ex aequo, il a été nécessaire d'utiliser les colonnes de « rang moyen (r_i) et (s_i) ». Cette valeur est calculée dans toutes les situations pour lesquelles les dimensions ont une fréquence ou un pointage identique. Par exemple, si deux dimensions ont une même fréquence, il faut faire la moyenne du rang assigné à chacune de ces dimensions. C'est cette valeur qui constituera le rang moyen. L'écart compris entre les rangs moyens est représenté dans la colonne « écart de rang (d_i) ». Tandis que la dernière colonne « écart de rang au carré (d_i^2) » contient cette même valeur à la puissance deux.

Le Collège Vassar, New York, États-Unis fournit sur son site Internet un outil de calcul du coefficient de corrélation des rangs de Spearman. Certaines données du tableau 4.1 ont été insérées afin de générer la valeur du coefficient. La valeur ainsi calculée est de 0,4705. Le tableau 4.2 constitue une capture d'écran de ce site. Elle reflète les données utilisées et les résultats obtenus. Cette valeur a par la suite été validée par un autre site Web utilisant les formules de Justine Ho de l'Université Paul Sabatier, Toulouse, France. Bien entendu, la valeur obtenue par ces deux outils est la même.

Tableau 4.2 Calcul du coefficient de corrélation des rangs de Spearman



Data Entry

	Ranks for		Raw Data for	
pairs	X	Y	X	Y
1	14.5	25	2	60
2	20	20.5	5	55
3	28	25	12	60
4	24.5	20.5	8	55
5	20	17	5	53
6	6.5	11	1	45
7	6.5	1	1	33
8	20	4.5	5	40
9	26	15	10	52
10	18	22	4	57
11	6.5	17	1	53
12	6.5	3	1	39
13	6.5	19	1	54
14	22.5	23	6	58
15	6.5	9.5	1	44
16	6.5	28	1	62
17	14.5	6	2	42
18	27	17	11	53
19	14.5	13	2	47
20	24.5	14	8	51
21	22.5	27	6	61
22	6.5	12	1	46
23	14.5	4.5	2	40
24	17	25	3	60
25	6.5	7.5	1	43
26	6.5	9.5	1	44
27	6.5	2	1	35
28	6.5	7.5	1	43

Reset | Calculate from Ranks | Calculate from Raw Data

n	r_s	t	df
28	0.4705	2.72	26

P		
	one-tailed	0.005741
	two-tailed	0.011481

Source : Capture d'écran de la page : http://faculty.vassar.edu/lowry/corr_rank.html

La valeur du coefficient de corrélation des rangs de Spearman résultant du calcul est comprise entre -1 et 1. Plus le coefficient s'approche de -1, plus les valeurs comparées sont diamétralement opposées. Plus la valeur du coefficient est près de 1, plus les valeurs comparées sont identiques. Ce qui signifie qu'il y a une très forte corrélation entre les données comparées. Tandis que la valeur zéro signifie qu'il n'y a pas de corrélation entre les variables [20].

Dans le cas qui nous intéresse, le coefficient de corrélation des rangs est de 0,4705. Ce qui signifie qu'il existe une corrélation partielle entre la littérature et les praticiens. Cette information permet de penser qu'il pourrait être judicieux de susciter plus de répondants afin d'avoir un meilleur échantillon. Malgré cette corrélation partielle, certaines dimensions ont une valeur de rang très similaire. En référence au tableau 4.1, voici les neuf dimensions offrant la plus forte corrélation.

Tableau 4.3 Dimensions ayant une forte corrélation

Écart de rang (d_i)	Dimension
0 à 1,0	• Accessibilité • Interprétable • Spécification des données • Volatilité
1,1 à 2,0	• Portabilité
2,1 à 3,0	• Actualisation et disponibilité • Conformité • Variété des données et sources de données • Transactionnel

Il est important de noter que le fait d'avoir obtenu une corrélation élevée ne signifie pas pour autant qu'il s'agit de dimensions qui sont hautement importantes. Bien que ces dimensions aient obtenu un rang très similaire dans chacune des deux compilations. Par exemple, la

dimension « volatilité » est cinquième avant-dernière dans le classement général, tandis qu'« actualisation et disponibilité » se classe au premier rang. C'est pourquoi il est nécessaire de procéder à des analyses comparatives afin de trouver une tendance. Cette situation s'explique par le fait que si la dimension se retrouve au dernier rang dans la littérature et qu'elle se retrouve aussi au dernier rang dans l'évaluation des praticiens alors l'écart de corrélation sera très faible ou même nul. Mais dans les deux cas, il s'agit d'une dimension qualifiée de « pas ou peu importante ».

4.2 Analyse comparative

En agençant différemment les résultats obtenus, il est possible de faire ressortir les dimensions ayant les plus hautes fréquences (6 et plus) et qui ont aussi un pointage élevé. Ce pointage a été déterminé de façon à ce qu'il ait tout au plus un écart de 10 par rapport au pointage maximal de 68 qui peut être obtenu. Le pointage maximal correspond à la multiplication du nombre de répondants et du poids associé à la colonne « très importante » c'est-à-dire (17 x 4). En utilisant ces notions, les dimensions qui ont obtenu une fréquence supérieure à 6 et un pointage supérieur à 58 et pour lesquelles il y a une forte corrélation peuvent être répertoriées. Deux dimensions se retrouvent dans cette situation. Il s'agit de « actualisation et disponibilité » et « interprétable ». Ces dimensions peuvent donc, de manière générale, être considérées comme de grande importance. Celles qui ne sont pas citées sont dans l'une des situations suivantes :

- la fréquence est inférieure à six,
- le pointage est inférieur à 58 et
- l'écart de rang est supérieur à trois.

Voici deux méthodes alternatives qui permettent de procéder à une analyse rapide en utilisant seulement les fréquences et le pointage de chacune des dimensions.

Une première méthode consiste à utiliser la somme de ces deux valeurs (fréquence et pointage) afin de connaître celles qui ont le plus d'importance. La dimension « actualisation et disponibilité » a obtenu 72, ce qui lui confère le plus haut pointage. Vient ensuite « principes de base de l'intégrité des données » avec 67. Tandis que les dimensions « interprétable » et « perception, pertinence et confiance » ont obtenu un pointage ex aequo de 64. Ce qui, en regard de cette méthode, identifie ces quatre dimensions comme étant les plus importantes.

Une deuxième méthode consiste à ne pas effectuer de parallèle entre ces deux compilations et utiliser seulement les pointages obtenus dans les colonnes « très importante » et « importante ». De cette façon, il est possible d'en déduire les dimensions suivantes :

- à valeur ajoutée,
- accessibilité,
- actualisation et disponibilité,
- cohérence et synchronisation,
- facilité d'utilisation et adaptable,
- interprétable,
- libre d'erreurs,
- principes de base de l'intégrité des données et
- sécurité.

Ce sont celles qui ont été identifiées comme étant les plus importantes par les répondants puisqu'au moins 88 % de ceux-ci les ont qualifiées avec l'un de ces niveaux. Soit 15 répondants ou plus. Avec cette méthode, la palme revient à la dimension « à valeur ajoutée » pour laquelle 17 répondants l'ont qualifiée de très importante ou d'importante.

4.3 Limites

Les résultats de l'analyse faisant l'objet du présent chapitre dépendent des résultats des deux analyses faisant l'objet des chapitres précédents. En effet, si les résultats d'une de ces deux analyses sont faussés, les résultats de cette analyse le sont également. Cette situation s'illustre facilement par l'exemple donné en lien avec la figure de Rhem, figure 1.3, qui mentionne que si les données sont de mauvaise qualité, l'information qui en découle sera nécessairement de mauvaise qualité. À son tour, la connaissance qui découle de l'information ne pourra que refléter la qualité de celle-ci.

Conclusion

Atteinte des objectifs

La méconnaissance du niveau de qualité des données comprises dans les différents systèmes informatiques des entreprises n'étant pas souhaitable, il devient pertinent pour les entreprises de mettre en place des processus et des méthodes qui permettent d'évaluer celle-ci. Ceci permettra aux entreprises d'atteindre un niveau de connaissances de la qualité de leurs données qui pourra être jugé satisfaisant. Pour y parvenir, des processus d'évaluation de la qualité des données doivent être en place. Puisque les dimensions s'avèrent être des outils couramment utilisés à cette fin et que leur sélection devient un aspect stratégique et primordial à un projet de qualité des données, voici comment cet essai satisfait aux trois objectifs.

Le premier objectif de cet essai est de permettre l'identification des dimensions les plus importantes en regard de la littérature. Le résultat de l'analyse effectuée dans cette optique indique que les cinq dimensions les plus importantes sont : 1) actualisation et disponibilité; 2) perception, pertinence et confiance; 3) exhaustivité; 4) cohérence et synchronisation; 5) précision. Ce premier objectif n'étant toutefois pas suffisant pour démarrer le projet en qualité des données, une deuxième analyse fut utilisée afin de parachever celle-ci. Celle-ci correspond au deuxième objectif de cet essai qui est d'identifier les dimensions les plus importantes en regard des praticiens. Le résultat de cette analyse indique que les dimensions les plus importantes sont : 1) à valeur ajoutée; 2) accessibilité; 3) actualisation et disponibilité; 4) cohérence et synchronisation; 5) conformité. Puisque les résultats obtenus à la suite de ces deux analyses n'ont pas une concordance parfaite, il a ensuite été question de procéder à des analyses comparatives qui ont permis d'identifier les dimensions qui semblent les plus importantes. C'est à la suite de cette non-concordance que fut construit le troisième objectif. Celui-ci consiste à comparer les résultats obtenus lors des deux premières analyses afin de voir si la littérature concorde avec les praticiens. En combinant les résultats des analyses

comparatives effectués, les dimensions 1) actualisation et disponibilité; 2) interprétable; ont pu être identifiées comme étant les plus importantes.

À la lumière de ces analyses, il va sans dire que ces deux dimensions devraient occuper une place importante dans un projet en qualité des données. Les autres dimensions ne sont toutefois pas à négliger puisqu'elles pourraient refléter d'avantage les besoins en qualité des données des systèmes qui requièrent une évaluation de la qualité des données qui les composent.

Contributions

Les recherches et les travaux effectués dans le cadre de ce travail ont permis de réaliser certaines contributions. Ils sont en lien direct avec les objectifs mentionnés précédemment.
La principale contribution concerne la compilation d'une liste épurée de dimensions présentes dans la littérature. Cette liste synthétise plus de 90 dimensions/définitions présentes dans vingt ouvrages de littérature en un tableau énumérant les 28 dimensions résultantes. Le produit issu de cette compilation a permis de hiérarchiser les dimensions afin d'identifier celles qui sont les plus importantes en regard de la littérature. Ce classement a été effectué en fonction des fréquences de citation de chacune d'elles. Cette contribution est le fruit de la première analyse.

À l'aide de cette compilation, un questionnaire s'adressant à des praticiens en technologies de l'information a pu être confectionné. C'est par ce questionnaire qu'il a ensuite été possible de connaitre le niveau d'importance accordée aux dimensions par les praticiens en TI. Le questionnaire constitue une autre contribution de ce travail. La troisième contribution concerne la méthode utilisée dans les différentes analyses comparatives qui ont permis de comparer les résultats entre la littérature et les praticiens en TI. Afin d'affirmer ou d'infirmer les résultats présentés à la suite de ces analyses, il pourrait être intéressant de consulter des spécialistes en qualité des données. La connaissance de ces spécialistes pourrait permettre de raffiner les

méthodes d'analyse afin de concevoir un processus de sélection précis pouvant être utilisé en entreprise.

Les différentes méthodes de sélection des dimensions et les différents cadres de qualité des données présentés dans cet essai sont de grandes sources d'informations qui fournissent des principes, des méthodes et des outils pertinents pour un tel projet. Tout gestionnaire de données devrait en prendre connaissance afin d'en juger le bien-fondé en vue d'une utilisation future.

Liste des références

[1] Batini, C., Cinzia, C., Chiara, F. et Andrea M., *Methodologies for Data Quality Assessment and Improvement*, ACM Computing Surveys, vol. 41, n^{o} 3, juillet 2009, p. 16:1-16:52.

[2] Batini, C., *Data Quality Concepts, Methodologies and Techniques*, 1re éd., Springer, New York, 1998, 249 p.

[3] Beynon-Davies, P., *Database Systems*, 3 éd. Palgrave Macmillan, Royaume-Unis, 2009, 601 p.

[4] English, L., *The Essentials of Information Quality Management*, http://www.information-management.com, 24 mars 2011.

[5] English, L., *Total Information Quality Management – A Complete methodology for IQ Management*, http://www.information-management.com, 17 février 2011.

[6] Even, A. et Shankaranarayanan G., *Dual Assessment of Data Quality in Customer Database*, ACM Journal of Data and Information Quality, vol. 1, n^{o} 3, article 15, décembre 2009.

[7] Gartner, http://www.gartner.com/it/page.jsp?id=501733, 31 octobre 2010.

[8] Helfert, M., Herrmann, C., *Proactive Data Quality Management for Data Warehouse Systems*, Librairie de l'Université de Toronto, 4^{e} conférence internationale sur la conception et la gestion des entrepôts de données, année 2002, p. 97-106.

[9] Huda, S. M. K., *Assessment of Deming's Philosophy with Respect to its Link to the Current Scenario in Pakistan Construction Industry*, NED University of Engineering & Technology Conference, Pakistan, 2008, p. 247-259.

[10] Institut canadien d'information sur la santé, *Le cadre de la qualité des données de l'ICIS 2009*, http://www.cihi.ca/CIHI-ext-portal/pdf/internet/DATA_QUALITY_FRAMEWORK_2009_FR, 7 décembre 2010.

[11] Journal du net, http://www.journaldunet.com/solutions/0701/070110-qr-edqm.shtml, 2 juin 2011.

[12] Kerr, K., Norris, T. et Stockdale, R., *Data Quality and Decision Making : A Healthcare Case Study*, 18e Conférence australienne sur les systèmes d'information, no 177, décembre 2007, p. 1016-1026.

[13] Levis, M., Helfert, M. et Brady, M., *Information quality management: Review of an evolving research area*, Massachuset Institute of Technologies, 2007, 15 p.

[14] Loshin, D., *Monitoring Data Quality Performance Using Data Quality Metrics*, Informatica, http://www.it.ojp.gov/documents/Informatica_Whitepaper_Monitoring_DQ_Using_Metrics.pdf, 17 février 2011.

[15] Madnick, S. E., Wang, R. Y., Lee, Y. W. et Zhu, H., *Overview and Framework for Data and Information Quality Research*, ACM Journal of Data and Information Quality, vol. 1, no 1, article 2, juin 2009.

[16] McGilvray, D., *Ten Steps to Quality Data and Trusted Information*, 1er éd., Morgan Kaufmann, USA, 2008, 325 p.

[17] Office québécois de la langue française, http://granddictionnaire.com/, 8 avril 2011.

[18] Pipino, L., L., Lee, Y. W. et Wang, R. Y., *Data Quality Assessment, Communication of the ACM*, vol. 45, no 4, avril 2002, p.211-218

[19] Redman, T.C., *Data Quality for the Information Age*, 1re éd., Artech House, Boston, MA., 1996, 261 p.

[20] Sprent, P., *Applied Nonparametric Statistical Methods*, 1 éd. Chapman & Hall, Londre, 1989, 297 p.

[21] The Data Management Association*, The DAMA Guide to The Data Management Body of Knowledge*, 1re éd., DAMA International, New Jersey, États-Unis, 2009, 406p.

[22] The Data Warehousing Institute, *Data Quality Report*, http://tdwi.org/research/2002/02/tdwis-data-quality-report.aspx?sc_lang=en, 20 novembre 2010.

[23] Wand, Y. et Wang, R., *Anchoring Data Quality Dimensions in Ontological Foundations*, Communications of the ACM, vol. 39, no 11, novembre 1996, p. 86-95.

[24] Wang R. Y., Ziad M. et Lee Y. W., *Data Quality*, 1re éd., Kluwer Academic Publisher, New York, USA, 2002, 162 p.

[25] Wang, R. et Strong, D., *Beyond Accuracy : What Data Quality Means to Data Consumers*, Journal of Management Information Systems, vol. 12, n° 4, printemps 1996, p. 5-44.

Annexe 1
Bibliographie

J. Rhem, A., *UML for Developing Knowledge Management Systems*, 1 éd. Auerbach Publications, États-Unis, 2006, 269 p.

Annexe 2

Formules du coefficient de corrélation des rangs de Spearman

La formule du coefficient de corrélation des rangs de Spearman utilisée lorsqu'il y a des valeurs ex aequo est celle-ci :

$$\rho = \frac{\sum_i r_i s_i - C}{\sqrt{\{[\sum_i r_i^2 - C][\sum_i s_i^2 - C]\}}} \qquad C = \tfrac{1}{4} n(n+1)^2$$

Cette équation nécessite de connaître les variables suivantes :

r_i = Rang moyen de X_i

s_i = Rang moyen de Y_i

n = Nombre de dimensions compilées

X_i = Fréquences de chaque dimension en regard de la littérature

Y_i = Pointage de chaque dimension en regard des praticiens (étudiants)

Annexe 3

Tableau synthèse des dimensions

Dimension résultante	Dimension de l'auteur	Définition	Auteur	Fréquence
Actualisation et disponibilité	Actualisation	Comment rapidement les données sont mises à jour. [2]	Carlo Batini	12
	Actualisation	La mesure qui indique que les données sont suffisamment à jour pour la tâche à accomplir. [18]	Richard Wang	
	Actualisation		Larry English	
	Actualisation	La mesure qui indique si les données sont à jour. Une donnée est considérée comme étant à jour en dépit d'une discordance possible causée par les changements liés au temps en regard de la bonne valeur. [19]	Thomas Redman	
	Actualisation	La mesure de la « fraîcheur » des données et de leur justesse face à d'éventuels changements liés au temps. La mesure du ratio de rafraichissement attendue qui permet de dire qu'une donnée est à jour et de déterminer quand elle sera périmée. [21]	DAMA	
	Actualisation et disponibilité	La mesure qui indique que l'information est à jour en regard de la donnée du monde réel qu'elle se doit de représenter. [14]	David Loshin	
	Actualisation et disponibilité	Une mesure du degré avec lequel les données sont à jour et disponibles pour une utilisation spécifique et dans les délais requis. [16]	Danette McGilvray	
	Actualité	L'actualité désigne principalement le caractère courant ou à jour des données au moment de leur diffusion selon l'écart entre la fin de la période de référence à laquelle les données se rapportent et la date à laquelle les données deviennent accessibles aux utilisateurs. [10]	ICIS	
	Disponibilité	La mesure de temps entre le moment où l'information est requise et le moment où elle est disponible. [21]	DAMA	

Dimension résultante	Dimension de l'auteur	Définition	Auteur	Fréquence
	Opportunité	Les données actualisées sont disponibles pour la tâche à accomplir. [2]	Carlo Batini	
	Synchronisation	Concerne une intégration juste des données ayant une date de modification différente. Une mesure de la disponibilité des données à partir de plusieurs sources. [2]	Carlo Batini	
Perception, pertinence et confiance	Confiance	Indique la qualité de la source. [1]	Carlo Batini	11
	Crédibilité	Considère si une certaine source fournit des données qui peuvent être considérées comme vraies, réelles et crédibles. [2]	Carlo Batini	
	Crédibilité	La mesure avec laquelle les données sont considérées comme vraies et crédibles. [18]	Richard Wang	
	Fiabilité (crédibilité)	Pour représenter si une source fournit des données qui véhiculent la bonne information. [2]	Carlo Batini	
	Perception, pertinence et confiance	Une mesure de la perception et de la confiance en la qualité des données, l'importance, la valeur et la pertinence des données aux besoins des entreprises. [16]	Danette McGilvray	
	Pertinence	La vue doit fournir les données nécessaires à l'application. [19]	Thomas Redman	
	Pertinence	La mesure avec laquelle les données sont applicables et utiles pour la tâche à accomplir. [24]	Richard Wang	
	Pertinence	La pertinence décrit de quelle façon une banque de données répond aux besoins actuels et futurs des utilisateurs. [10]	ICIS	
	Pertinence	Les données sont applicables et utiles pour la tâche à accomplir. [2]	Carlo Batini	
	Réputation	Une mesure avec laquelle les données sont hautement	Richard Wang	

Dimension résultante	Dimension de l'auteur	Définition	Auteur	Fréquence
		considérées en termes de source et de contenu. [24]		
	Réputation	Quel est le niveau de confiance de la source de donnée. [2]	Carlo Batini	
Exhaustivité	Couverture de données	Une mesure de la disponibilité et de l'exhaustivité des données par rapport à l'univers de données ou de la population d'intérêt. [16]	Danette McGilvray	10
	Exhaustivité, Complétude	Une mesure qui indique si les données sont d'une ampleur suffisante, d'une étendue et d'une portée requise pour la tâche à accomplir. Il existe trois types d'exhaustivité, celle du schéma, des colonnes et de la population. [2]	Carlo Batini	
	Exhaustivité, Complétude	Indique que certains attributs doivent être affectés de valeurs dans un ensemble de données. [14]	David Loshin	
	Exhaustivité, Complétude	Une mesure qui indique si des données sont manquantes, d'une ampleur et d'une profondeur suffisante en regard de la tâche à accomplir. [24]	Richard Wang	
	Exhaustivité, Complétude	Reflète l'absence de données d'attributs. [6]	G. Shankaranarayanan	
	Exhaustivité, Complétude	Une mesure qui indique que les attributs ont toujours une valeur comprise dans un certain ensemble de données. Elle est aussi utilisée pour mesurer si toutes les lignes appropriées pour un ensemble de données sont présentes. [21]	DAMA	
	Exhaustivité, Complétude		Larry English	
	Exhaustivité, Complétude	Chaque item de données nécessaires doit y être présent. [19]	Thomas Redman	

Dimension résultante	Dimension de l'auteur	Définition	Auteur	Fréquence
	Volume de données	Une mesure qui indique si le volume de données est approprié pour la tâche à accomplir. [24]	Richard Wang	
	Volume de données	La quantité ou le volume de données disponible est approprié. [1]	Carlo Batini	
Cohérence et synchronisation	Cohérence	Les valeurs des données pour un ensemble de données sont compatibles avec les valeurs dans un autre ensemble de données. [14]	David Loshin	8
	Cohérence	Identifie la violation des règles de sémantique définies pour un ensemble de données qui peut correspondre à des uplets, des tables relationnelles ou des fichiers plats. [2]	Carlo Batini	
	Cohérence	La cohérence vise à assurer que les valeurs données à un ensemble de données sont compatibles avec les valeurs comprises dans un autre ensemble de données. La cohérence peut être définie entre un ensemble de valeurs d'attribut et un autre ensemble d'attributs dans le même enregistrement (la cohérence au niveau des enregistrements), entre un ensemble de valeurs d'attributs et un autre attribut figurant dans les registres différents (cohérence entre enregistrements), ou entre un ensemble de valeurs d'attribut et l'attribut du même ensemble pour un même enregistrement à différents points dans le temps (cohérence temporelle). [21]	DAMA	
	Cohérence		Larry English	
	Cohérence des valeurs	Signifie que deux ou plusieurs données ne sont pas en conflit les unes avec les autres. [19]	Thomas Redman	

Dimension résultante	Dimension de l'auteur	Définition	Auteur	Fréquence
	Cohérence et synchronisation	Une mesure de l'équivalence des informations stockées ou utilisées dans différents magasins de données, les applications et les systèmes, ainsi que les procédés utilisés pour la fabrication de données. [16]	Danette McGilvray	
	Représentation cohérente	Utiliser une méthode de représentation uniforme pour une même donnée. Par exemple, la représentation d'une adresse sera toujours, un champ contenant le numéro d'immeuble et le nom de la rue. [2]	Carlo Batini	
	Représentation cohérente	La mesure avec laquelle les données sont toujours présentées selon un même format. [24]	Richard Wang	
Précision	Exactitude	La dimension de l'exactitude porte sur la conformité de l'information contenue dans la banque de données, ou qui en découle, à la réalité qu'elle doit mesurer. [10]	ICIS	8
	Précision	Est la différence entre une valeur v et une valeur v' étant considérée comme la représentation juste du monde réel. [2]	Carlo Batini	
	Précision	Une mesure de l'exactitude du contenu des données (ce qui nécessite une source de référence faisant autorité pour être identifié et accessible). [16]	Danette McGilvray	
	Précision	Le degré avec lequel les données représentent correctement les objets du monde réel qu'elles sont destinées à modéliser. [14]	David Loshin	
	Précision		Larry English	
	Précision	La mesure dans laquelle les données sont considérées comme exactes, fiables et certifiées exemptes d'erreurs. [25]	Richard Wang	
	Précision	Le degré avec lequel les données représentent correctement les entités de la vie réelle qu'ils modélisent. Il s'agit habituellement de comparer la concordance entre la donnée	DAMA	

Dimension résultante	Dimension de l'auteur	Définition	Auteur	Fréquence
		modélisée et une source de donnée de référence qui contient la bonne valeur. [21]		
	Précision	Les données peuvent être considérées comme porteuses de la bonne information. [2]	Carlo Batini	
Interprétable	Facilité de compréhension	La mesure avec laquelle les données sont claires, sans ambiguïté et faciles à comprendre. [24]	Richard Wang	6
	Facilité de compréhension	La mesure avec laquelle les données sont utiles et offrent un avantage lors de leur utilisation. [2]	Carlo Batini	
	Intelligibilité	Concerne la documentation et les métadonnées qui sont disponibles pour interpréter correctement le sens et les propriétés des sources de données. [1]	Carlo Batini	
	Intelligibilité	La mesure avec laquelle les données sont dans un langage approprié, ont des symboles, des unités et des définitions claires. [24]	Richard Wang	
	Intelligibilité	La mesure dans laquelle les données sont faciles à comprendre. [23]	Richard Wang	
	Interprétable	Un bon format est celui qui aide l'utilisateur à interpréter correctement ses valeurs. [19]	Thomas Redman	
Principes de base de l'intégrité des données	Comparabilité	La dimension de la comparabilité évalue le degré de cohérence des bases de données au fil du temps et leur utilisation des conventions standard (comme des éléments de données ou des périodes de référence), qui rendent ces bases comparables à d'autres bases de données. [10]	ICIS	6
	Intégrité	La sécurité de l'information, à savoir la protection de l'information afin d'éliminer les modifications non autorisées, non prévues ou non intentionnelles dans le but de prévenir la corruption ou la falsification des données. [2]	Carlo Batini	

Dimension résultante	Dimension de l'auteur	Définition	Auteur	Fréquence
	Intégrité référentielle	Utilisation d'identificateurs uniques pour les objets de l'environnement afin de simplifier la gestion des données. [14]	David Loshin	
	Intégrité référentielle	L'intégrité référentielle survient lorsque toutes les références destinées à une donnée dans une colonne d'une table et les données d'une autre colonne de la même table ou d'une autre table sont valables. S'applique aussi aux clés étrangères. Si une clé primaire apparait comme une clé étrangère, le document auquel est associée cette clé existe réellement. [21]	DAMA	
	Principes de base de l'intégrité des données	Une mesure de l'existant, de la validité, de la structure, du contenu et autres caractéristiques des bases de données. [16]	Danette McGilvray	
	Traçabilité	Une mesure dans laquelle les données sont bien documentées, vérifiables et dont la source est identifiable. [23]	Richard Wang	
Accessibilité	Accessibilité	Une mesure de la capacité de l'utilisateur à accéder aux données de sa propre culture avec les technologies disponibles. [2]	Carlo Batini	5
	Accessibilité	Une mesure de la disponibilité des données ou de leur facilité et de leur rapidité à être accessibles. [24]	Richard Wang	
	Accessibilité	Les valeurs des données doivent pouvoir être facilement obtenues. [19]	Thomas Redman	
	Accessibilité		Larry English	

Dimension résultante	Dimension de l'auteur	Définition	Auteur	Fréquence
	Flexibilité	Une mesure de l'extensibilité des données, de leur adaptabilité et de leur facilité à être utilisées pour d'autres besoins. [25]	Richard Wang	
Conformité	Adéquation	Un format est plus approprié qu'un autre s'il est plus approprié aux besoins de l'utilisateur. [19]	Thomas Redman	5
	Conformité	Indique si les instances des données sont soit emmagasinées, échangées ou présentées dans un format qui est compatible avec le domaine des valeurs et avec les autres valeurs ayant des attributs similaires. [14]	David Loshin	
	Conformité		Larry English	
	Conformité	Une mesure du coût requis pour la collecte des données afin de déterminer s'il est raisonnable. Vérifie que la valeur des données est conforme au domaine de valeur qui comprend le type des données, la précision, le format, le choix des valeurs prédéfinies, la plage acceptée, etc. [21]	DAMA	
	Qualité de présentation	Une mesure qui indique la façon dont l'information est présentée et recueillie auprès de ceux qui l'utilisent. Le format et la structure permettent l'utilisation adéquate des données. [16]	Danette McGilvray	
Duplication	Duplication	Une mesure des duplications indésirables existant au sein ou à travers des systèmes pour des champs particuliers, des enregistrements ou un ensemble de données. [16]	Danette McGilvray	5
	Duplication	La duplication se produit lorsqu'une entité du monde réel est présente deux fois ou plus dans une source de données. [2]	Carlo Batini	
	Duplication		Larry English	

Dimension résultante	Dimension de l'auteur	Définition	Auteur	Fréquence
	Unicité	L'unicité se réfère à l'exigence que les entités modélisées dans l'entreprise soient capturées et représentées une seule fois dans l'architecture de l'application en cause. [14]	David Loshin	
	Unicité	L'unicité est une représentation unique d'une donnée pour un ensemble de données défini. L'utilisation d'une clé sera nécessaire pour référer à cette donnée unique. [21]	DAMA	
Facilité d'utilisation et adaptables	Convivialité		Larry English	4
	Facilité d'utilisation	La facilité d'utilisation désigne la facilité avec laquelle on peut comprendre les données d'une banque de données et y accéder. [10]	ICIS	
	Facilité de manipulation	La mesure avec laquelle les données sont faciles à manipuler et à appliquer aux différentes tâches. [24]	Richard Wang	
	Facilité d'utilisation et adaptables	Une mesure qui indique le niveau avec lequel les données peuvent être consultées et utilisées ainsi que la mesure avec laquelle les données peuvent être mises à jour, entretenues et gérées. [16]	Danette McGilvray	
Sécurité	Sécurité	Une mesure de l'accès aux données afin qu'elle soit limitée de manière appropriée afin d'en assurer sa sécurité. [24]	Richard Wang	3
	Sécurité	Représente le besoin de contrôler les accès et l'utilisation des données. [21]	DAMA	
	Sécurité	Une mesure de l'accès aux données en terme de restriction afin qu'elles soient conservées en toute sécurité. [2]	Carlo Batini	
À valeur ajoutée	À valeur ajoutée	Les données sont bénéfiques et l'on en retire des avantages lors de leur utilisation. [1]	Carlo Batini	2

Dimension résultante	Dimension de l'auteur	Définition	Auteur	Fréquence
	À valeur ajoutée	Une mesure de la convivialité et des avantages offerts par leur utilisation. [24]	Richard Wang	
Objectivité	Objectivité	Prend en compte l'impartialité de la source de données de l'approvisionnement. [1]	Carlo Batini	2
	Objectivité	Une mesure de l'impartialité, exempte de préjugé et objective. [24]	Richard Wang	
Portabilité	Portabilité	La possibilité de réutiliser le format d'une donnée dans une variété de situations. [19]	Thomas Redman	2
	Portabilité	Le format peut être appliqué à un aussi large éventail de situations que possible. [2]	Carlo Batini	
Représentation concise	Représentation concise	Les données sont représentées de manière compacte sans être surchargées. [2]	Carlo Batini	2
	Représentation concise	Une mesure de représentation des données afin qu'elles soient compactes et sans surcharge. [24]	Richard Wang	
Convenance	Convenance	Un format est plus approprié qu'un autre s'il est plus adapté aux besoins des utilisateurs. [2]	Carlo Batini	1
Décroissance des données	Décroissance des données	Une mesure du taux de décroissance des données. [16]	Danette McGilvray	1
Format flexible	Format flexible	Les changements dans les besoins des utilisateurs et du support d'enregistrement peuvent être facilement adaptés. [2]	Carlo Batini	1
Format précis	Format précis	La possibilité de faire la distinction entre les éléments du domaine qui doivent être considérés par les utilisateurs. [2]	Carlo Batini	1
Habileté à représenter des valeurs nulles	Habileté à représenter des valeurs nulles	La capacité de distinguer soigneusement (sans ambiguïtés) des valeurs nulles et par défaut à partir des valeurs applicables dans ce domaine. [2]	Carlo Batini	1

Dimension résultante	Dimension de l'auteur	Définition	Auteur	Fréquence
La variété des données et sources de données	La variété des données et sources de données	Une mesure de la disponibilité des données à partir de plusieurs sources. [23]	Richard Wang	1
Libre d'erreurs	Libre d'erreurs	Une mesure dans laquelle les données sont exactes et fiables. [24]	Richard Wang	1
Rapport coût-efficacité	Rapport coût-efficacité	Une mesure du coût requis pour la collecte des données afin de déterminer s'il est raisonnable. [23]		1
Spécifications de données	Spécifications de données	Une mesure de l'existant, de l'exhaustivité, de la qualité et de la documentation des normes de données, des modèles de données, des règles métier, des métadonnées et des données de référence. [16]	Danette McGilvray	1
Transactionnel	Transactionnel	Une mesure du degré avec lequel les données permettent de produire l'opération commerciale ou les résultats souhaités. [16]	Danette McGilvray	1
Utilisation efficace de la mémoire	Utilisation efficace de la mémoire	L'efficacité dans la représentation physique. Une icône est moins efficace qu'un code. [2]	Carlo Batini	1
Volatilité	Volatilité	La fréquence avec laquelle les données varient dans le temps. Une donnée qui ne change pas dans le temps (par exemple la date de naissance) a une volatilité de valeur 0. [2]	Carlo Batini	1

Traduction libre
Source : Voir la liste des références

Annexe 4

Questionnaire ayant servi à l'étude comparative

Questionnaire sur l'importance des dimensions dans un contexte de qualité des données

*Obligatoire

Contexte

Ce questionnaire s'inscrit dans le cadre d'un essai en TI qui porte sur le thème de la qualité des données.

La qualité des données est utilisée pour savoir si une donnée est une source fiable d'information dans le but de répondre adéquatement à l'utilisation auquel elle est dédiée.

Une dimension est utilisée pour représenter certaines caractéristiques des données. Elle sert entre autre, à mesurer la qualité des données en regard des critères qui la caractérisent. Par exemple, la dimension « sécurité » sert à mesurer le niveau de sécurité de la donnée en des termes tel que la gestion des accès, la confidentialité etc. mais elle ne cherche pas à mesurer si la donnée est juste (précision) ou si la donnée est à jour (actualisation).

Objectifs

Ce questionnaire vise à évaluer l'importance perçue de différentes dimensions en qualité des données.

Description des répondants

Connaissances en qualité des données

	Fortement en désaccord	En désaccord	Neutre	En accord	Fortement en accord
J'utilise fréquemment, dans mes tâches courantes, des données comprises dans un système, tel qu'une base de données, un entrepôt de données…	○	○	○	○	○
Je possède de très bonnes connaissances en termes de « qualité des données ».	○	○	○	○	○
J'ai déjà fait partie d'une équipe ayant comme objectif la gestion de la qualité des données.	○	○	○	○	○
Je connais ce que sont les dimensions et à quoi elles servent.	○	○	○	○	○

Expérience TI

	Je ne travaille pas en TI	0 à 5 ans	5 à 10 ans	10 à15 ans	15 ans ou plus
Depuis combien d'années travaillez-vous en TI?	○	○	○	○	○

Scolarité TI

	Je n'ai pas d'études en TI	Collégial	Universitaire (1er cycle)	Universitaire (2e ou 3e cycle)	Autre
Quel est votre plus haut niveau académique en TI?	○	○	○	○	○

Évaluation de l'importance des dimensions

Pour chacune des dimensions, veuillez indiquer le niveau d'importance qui vous semble le plus juste. Plusieurs dimensions peuvent avoir le même niveau d'importance. Pour chacune des dimensions, il vous suffit de sélectionner la case d'option correspondant à votre réponse. Une fois le questionnaire complété, vous devez cliquer sur "Envoyer" au bas du formulaire.

À valeur ajoutée *

	Très importante	Importante	Peu importante	Pas importante	Ne sait pas
Les données sont bénéfiques et on en retire des avantages lors de leur utilisation.	○	○	○	○	○

Accessibilité *

	Très importante	Importante	Peu importante	Pas importante	Ne sait pas
Une mesure de la capacité de l'utilisateur à accéder aux données de sa propre culture avec les technologies disponibles.	○	○	○	○	○

Actualisation et disponibilité *

	Très importante	Importante	Peu importante	Pas importante	Ne sait pas
La mesure qui indique que l'information est à jour en regard de la donnée du monde réel qu'elle se doit de représenter.	○	○	○	○	○

Cohérence et synchronisation *

	Très importante	Importante	Peu importante	Pas importante	Ne sait pas
Une mesure de l'équivalence des informations stockées ou utilisées dans les différents magasins de données, les applications et les systèmes.	○	○	○	○	○

Conformité *

	Très importante	Importante	Peu importante	Pas importante	Ne sait pas
Indique si les instances des données sont soit emmagasinées, échangées ou présentées dans un format qui est compatible avec le domaine des valeurs et avec les autres valeurs ayant des	○	○	○	○	○

	Très importante	Importante	Peu importante	Pas importante	Ne sait pas
attributs similaires.					

Convenance *

	Très importante	Importante	Peu importante	Pas importante	Ne sait pas
Un format est plus approprié qu'un autre s'il est plus adapté aux besoins des utilisateurs.					

Décroissance des données *

	Très importante	Importante	Peu importante	Pas importante	Ne sait pas
Une mesure du taux de décroissance des données (retour à une valeur antérieur).					

Duplication *

	Très importante	Importante	Peu importante	Pas importante	Ne sait pas
Une mesure des duplications indésirables existant au sein ou à travers des systèmes pour des champs particuliers, des enregistrements ou un ensemble de données.					

Exhaustivité *

	Très importante	Importante	Peu importante	Pas importante	Ne sait pas
Une mesure qui indique si les données sont d'une ampleur suffisante, d'une étendue et d'une portée requise pour la tâche à accomplir. Il existe trois types d'exhaustivité, celle du schéma, des colonnes et de la population.	○	○	○	○	○

Facilité d'utilisation et adaptables *

	Très importante	Importante	Peu importante	Pas importante	Ne sait pas
Une mesure qui indique le niveau avec lequel les données peuvent être consultées et utilisées ainsi que la mesure avec laquelle les données peuvent être mises à jour, entretenues et gérées.	○	○	○	○	○

Format flexible *

	Très importante	Importante	Peu importante	Pas importante	Ne sait pas
Les changements dans les besoins des utilisateurs et du	○	○	○	○	○

	Très importante	Importante	Peu importante	Pas importante	Ne sait pas
support d'enregistrement peuvent être facilement adaptés.					

Format précis *

	Très importante	Importante	Peu importante	Pas importante	Ne sait pas
La possibilité de faire la distinction entre les éléments du domaine qui doivent être considérés par les utilisateurs.	○	○	○	○	○

Habileté à représenter des valeurs nulles *

	Très importante	Importante	Peu importante	Pas importante	Ne sait pas
La capacité de distinguer soigneusement (sans ambiguïtés) des valeurs nulles et par défaut à partir des valeurs applicables dans ce domaine.	○	○	○	○	○

Interprétable *

	Très importante	Importante	Peu importante	Pas importante	Ne sait pas
Concerne la documentation et les métadonnées qui sont disponibles	○	○	○	○	○

	Très importante	Importante	Peu importante	Pas importante	Ne sait pas
pour interpréter correctement le sens et les propriétés des sources de données.					

La variété des données et sources de données *

	Très importante	Importante	Peu importante	Pas importante	Ne sait pas
Une mesure de la disponibilité des données à partir de plusieurs sources.	○	○	○	○	○

Libre d'erreurs *

	Très importante	Importante	Peu importante	Pas importante	Ne sait pas
Une mesure dans laquelle les données sont exactes et fiables.	○	○	○	○	○

Objectivité *

	Très importante	Importante	Peu importante	Pas importante	Ne sait pas
Prends en compte l'impartialité de la source de données de l'approvisionnement.	○	○	○	○	○

Perception, pertinence et confiance *

	Très importante	Importante	Peu importante	Pas importante	Ne sait pas
Considère si une certaine source	○	○	○	○	○

	Très importante	Importante	Peu importante	Pas importante	Ne sait pas
fournit des données qui peuvent être considérées comme vraies, réelles et crédibles.					

Portabilité *

	Très importante	Importante	Peu importante	Pas importante	Ne sait pas
La possibilité de réutiliser le format d'une donnée dans une variété de situations.	○	○	○	○	○

Précision *

	Très importante	Importante	Peu importante	Pas importante	Ne sait pas
Est la différence entre une valeur v et une valeur v' étant considérée comme la représentation juste du monde réel.	○	○	○	○	○

Principes de base de l'intégrité des données *

	Très importante	Importante	Peu importante	Pas importante	Ne sait pas
La sécurité de l'information, à savoir la protection de l'information afin d'éliminer les modifications non autorisées, non	○	○	○	○	○

	Très importante	Importante	Peu importante	Pas importante	Ne sait pas
prévues ou non intentionnelles dans le but de prévenir la corruption ou la falsification des données.					

Rapport coût-efficacité *

	Très importante	Importante	Peu importante	Pas importante	Ne sait pas
Une mesure du coût requis pour la collecte des données afin de déterminer s'il est raisonnable.	○	○	○	○	○

Représentation concise *

	Très importante	Importante	Peu importante	Pas importante	Ne sait pas
Les données sont représentées de manière compacte sans être surchargées.	○	○	○	○	○

Sécurité *

	Très importante	Importante	Peu importante	Pas importante	Ne sait pas
Une mesure de l'accès aux données afin qu'elle soit limitée de manière appropriée afin d'en assurer sa sécurité.	○	○	○	○	○

Spécifications de données *

	Très importante	Importante	Peu importante	Pas importante	Ne sait pas
Une mesure de l'existant, de l'exhaustivité, de la qualité et de la documentation des normes de données, des modèles de données, des règles métier, des métadonnées et des données de référence.	○	○	○	○	○

Transactionnel *

	Très importante	Importante	Peu importante	Pas importante	Ne sait pas
Une mesure du degré avec lequel les données permettent de produire l'opération commerciale ou les résultats souhaités.	○	○	○	○	○

Utilisation efficace de la mémoire *

	Très importante	Importante	Peu importante	Pas importante	Ne sait pas
L'efficacité dans la représentation physique. Une icône est moins efficace qu'un code.	○	○	○	○	○

Volatilité *

	Très importante	Importante	Peu importante	Pas importante	Ne sait pas
La fréquence avec laquelle les données varient dans le temps. Une donnée qui ne change pas dans le temps (par exemple la date de naissance) a une volatilité de valeur 0.	○	○	○	○	○

Questions / commentaires?

Fin du questionnaire

Merci pour le temps que vous avez pris afin de répondre à ce questionnaire. Vos réponses me seront d'une grande utilité dans la poursuite de ma rédaction.

Envoyer

Fourni par Google Documents