



HAL
open science

De la Qualité de l'Information Administrative et Scientifique

Thierry Bontems, Sabine Goulin

► **To cite this version:**

Thierry Bontems, Sabine Goulin. De la Qualité de l'Information Administrative et Scientifique. 8. Ecole Inter-organismes "Qualité en Recherche et en Enseignement Supérieur", Montpellier (France), QUARES, Sep 2010, Montpellier, France. pp.180. <halshs-01402579>

HAL Id: halshs-01402579

<https://shs.hal.science/halshs-01402579v1>

Submitted on 24 Nov 2016

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire HAL, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.



Distributed under a Creative Commons CC BY-NC-ND 4.0 - Attribution - Non-commercial use - No Derivative Works - International License

Sommaire

1.	Introduction	3			
2.	Enjeux pour la recherche et l'enseignement supérieur	7			
3.	Quelques définitions.....	9			
4.	Quatre facteurs qui qualifient l'information.....	10			
	4.1 Accessibilité.....	10			
	4.2 Utilité.....	10			
	4.3 Crédibilité	11			
	4.4 Interprétabilité	12			
5.	Six dimensions définissant la qualité des données 13				
	5.1 Pertinence	13			
	Suivi des besoins.....	13			
	Examen du programme	14			
	Etablissement des priorités	14			
	Suivi de la performance	14			
	5.2 Exactitude	14			
	Conception.....	15			
	Mise en œuvre.....	15			
	Evaluation de l'exactitude.....	15			
	5.3 Rapidité de diffusion de l'information	16			
	5.4 Accessibilité.....	16			
	5.5 Interprétation.....	17			
	5.6 Cohérence	17			
6.	Indicateurs et mesures.....	18			
7.	Qui est concerné ?.....	19			
8.	Gouvernance	19			
	8.1 Direction de l'unité et directions métiers	20			
	8.2 Comité qualité des données	20			
9.	Conclusion	20			

1. Introduction

A la préhistoire, le principal vecteur de sauvegarde de l'information reste l'image apposée sur le mur grotte. Le périmètre de diffusion est extrêmement restreint et s'adresse à la famille essentiellement. L'apparition dans les civilisations antiques de la communication par écrit fait passer cette information au niveau d'une région et/ou d'un clan. C'est avec l'arrivée de l'imprimerie dès le XI^{ème} siècle en Chine, après le XV^{ème} en Europe qu'on constate une accélération de l'accroissement du volume d'information ainsi qu'une diffusion au niveau des pays et des nations. A partir de cette époque, la quantité d'information disponible va

multiplier des médias disponibles, la photographie, la radio, le cinéma, la télévision, le téléphone, autant de vecteurs qui continuent à démultiplier la quantité d'information. Parallèlement à cet accroissement du nombre de ces vecteurs de communication, on constate un accroissement de la rapidité et de la mondialisation des échanges de données. L'arrivée d'internet à la fin du XX^{ème} siècle va être le catalyseur de cette accélération et de cette mondialisation des échanges. Aujourd'hui nous assistons à un phénomène de convergence de l'ensemble de ces vecteurs de communication pour arriver à un outil multi communicant centralisateur, le mobile, sur lequel on peut réunir l'ensemble des médias précédemment cité, nous avons quitté le monde du multimédia

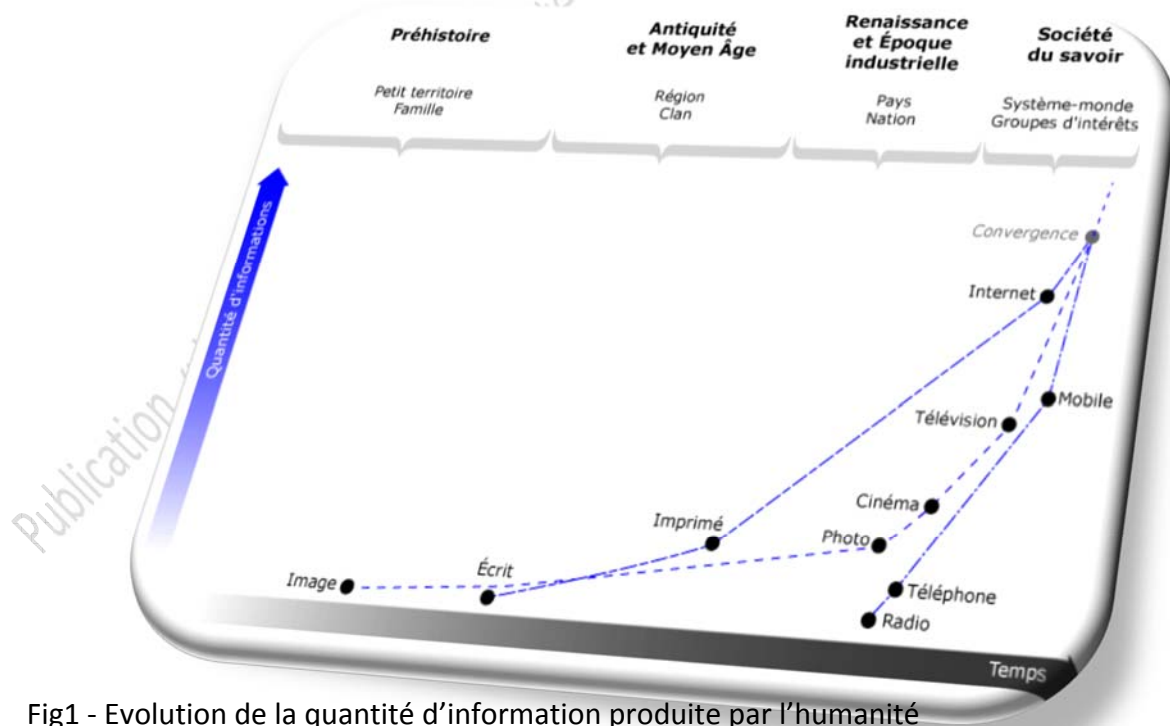


Fig1 - Evolution de la quantité d'information produite par l'humanité

s'accroître de façon exponentielle. L'avènement de l'ère industrielle voit la

pour rentrer dans le monde de la trans-médiatisation (cf fig. 1). Néanmoins si la

multiplication des médias semble être ralentie, la quantité d'informations disponibles et la rapidité de diffusion de ces dernières ne paraissent pas prêtes de ralentir. Plus d'informations ont été générées durant les 30 dernières années que pendant les 500 qui les ont précédées. A titre d'exemple, les nouvelles technologies de type RFID génèrent un volume de données 10 à 100 fois supérieur à celui des technologies de type code-barres.

La question se pose aujourd'hui de mettre en place un référentiel permettant de garantir la « qualité des données ».

La non-qualité à un coût, Aux Etats-Unis une étude ⁽¹⁾ de 2002 a estimé le coût de la mauvaise qualité des données à plus de 600 milliards de dollars par an, en comparaison, c'est le PIB² de la suisse (20^{ème} rang mondial) Nous verrons comment par exemple le retour sur investissement (ROI)³ lors de l'utilisation de bases de données altérés ou inexacts peut varier du simple au double.

Autres exemples de non-qualité aux conséquences désastreuses tant sur le plan financier que sur le plan humain : lors de l'introduction en bourse de la petite société JCom le 8 décembre 2005, la maison de courtage Mizuho Securities a placé 610.000 titres à 1 yen au lieu de vendre une action à 610.000 yens. Cette erreur grossière a coûté 286 millions d'euros à Mizuho Securities, a provoqué la chute des titres des maisons de courtage, et a suscité la démission des

principaux dirigeants de la Bourse de Tokyo⁴. En 1999 la perte par la NASA du satellite Mars Climate Orbiter a coûté au contribuable américain, 125 millions de dollars. L'enquête conclura que certains paramètres ont été calculés en paramètres anglo-saxons et transmis en l'état à l'équipe de navigation qui les attendaient en système métrique⁵.

Dans la même lignée, lors de la conception de l'A380, il a été prouvé que les difficultés concernant la phase d'industrialisation ont porté essentiellement sur le câblage électrique d'une partie du fuselage. La conception avait été menée avec des logiciels de versions différentes le coût pour EADS : des mois de retard de livraison et des pénalités inhérentes⁶.

Autre mésaventure, certes moins onéreuse pour la société mais au combien délicate pour les intéressés, Trois touristes norvégiens attendus à Rhodes (Grèce) atterrissent à Rodez (France) suite à une erreur de saisie de leur commande internet⁷. Cette mésaventure peut arriver à chacun de nous. Il est tout à fait possible aujourd'hui de réserver sur le site de certaines compagnies aériennes lowcost, un vol dont le retour se programme avant l'aller.

Les conséquences de ce genre d'erreurs peuvent avoir des conséquences humaines

¹ « data quality and the bottom line » -The Data Warehousing Institute 2002

²

	Produit	Intérieur	Brut :
	http://www.journaldunet.com/economie/magazine/classement-pib.shtml		

³ ROI : Return of Investissement – Retour sur investissement

⁴ http://www.senat.fr/rap/r06-017/r06-017_mono.html
-<http://www.mysharetrading.com/news/other-trading-news/trading-suspended-in-j-com-after-30-billion-yen-typo-blunder-by-mizuho-securities.htm>

⁵ <http://edition.cnn.com/TECH/space/9909/30/mars.metric/>

⁶ Audition de M. Louis Gallois-coprésident exécutif d'EADS du 22 novembre 2006

⁷ http://voyageforum.com/voyage/france_grece_trois_norvegien_s_prennent_un_avion_pour_rodez_au_lieu_rhodes_D1255765/

dramatiques. En 2006, Un touriste français et un touriste allemand en ont fait la douloureuse expérience. Ils ont été condamnés à 18 mois de prison (en ont purgé 15) dans les terribles geôles de la prison d'Evine à Téhéran en Iran pour avoir navigué sur les eaux territoriales Iraniennes sans autorisations. Sur les cartes, cette zone était indiquée comme faisant partie du territoire des Emirats Arabes Unis⁸.

Nous pouvons mesurer concrètement le coût de la non-qualité. Intéressons-nous, comme cas d'école, au cas du traitement d'une enquête téléphonique et à l'impact financier qui découle directement d'une mauvaise qualité des données.

Supposons qu'une entreprise veut faire l'envoi d'une brochure annonçant un nouveau service à l'ensemble de ses clients. Elle utilise sa base d'adresse client qui contient des enregistrements dupliqués avec un ratio de duplication 1.04. De plus sa base client contient des enregistrements multiples pour le même foyer avec un ratio de foyer 1.11

Cette entreprise envoie 50 000 brochures pour un coût total de 150 000€. Supposons qu'elle fasse un bénéfice moyen par vente de 1000€. Si sa base client a un ratio de duplication de 1 elle espère un taux de réponse de 2%. Le ratio étant à 1.04 son taux de réponse diminue pour passer à 1.92%. De plus si son ratio de foyer est à 1, elle espère un taux de conversion de 20%, taux qui chute à 18.02%. Cette diminution de la rentabilité du projet est exclusivement

due à une mauvaise qualité des données de départ.

Hypothèses		
Nombres de brochures envoyées	50000	
Coût total du programme	150 000 €	
Bénéfice moyen par vente	1 000 €	
Ratio de duplication	1	1.04
Taux de réponse	2 %	1.92 %
Ratio de foyer	1	1.11
Taux de conversion	20 %	18.02 %
Résultats		
Nombre de réponse	1000	962
Coût par réponse	150 €	156 €
Nombre d'acheteurs	200	173
Bénéfice total de la campagne	200 000 €	173 250 €
Retour sur Investissement	33.33 %	15.50 %

Fig 2 - Exemple de coût de la non qualité lors une campagne marketing

Supposant maintenant que l'entreprise obtienne en théorie 1000 réponses. Ce nombre de réponses passe mathématiquement à 962. Le coût de chaque réponse passe lui aussi de 150 € à 156 €. Le nombre d'acheteur de 200 à 173.

De façon tout aussi mathématique, le bénéfice de la campagne passe quand à lui de 200 000€ à 173 500€ soit un retour sur investissement qui diminue de 33.33% à 15.50 %, un ROI qui varie du simple au double.

Ce que nous venons de montrer s'applique de manière identique dans nos laboratoire de Sciences Humaines et Sociales (SHS) lors des enquêtes. Nous ne parlerons pas de « bénéfice » en euros, mais de « bénéfice » en termes d'augmentation de corpus de

⁸ <http://www.iran-resist.org/article1841-> carte publiée par un site dissident au régime politique en place suite à l'arrestation

données pertinent ou en termes de minimisation de coûts.

Prenons maintenant l'exemple de nos établissements d'enseignement supérieur et de recherche, lorsque nous manipulons des données, la qualité de celles-ci est souvent vue dans le sens d'assurer « la bonne information à la bonne personne au bon moment », nous ne la voyons pas toujours avec cette notion de coût. Cependant, dans le contexte économique actuel, nous devons absolument prendre en compte cette notion de coût liée à la qualité des données.

Deux exemples pour éclairer cette notion. Les établissements d'enseignement supérieur travaillent sur des dispositifs d'enquêtes sur le devenir des étudiants (insertion professionnelle post diplôme à 6, 18 et 30 mois). Pour ce faire, c'est l'adresse mail de l'étudiant qui est une des données de référence puisqu'on envoie à cette adresse une enquête sur le devenir puis le parcours professionnel. Le taux de mauvaise adresse peut être assez élevé si la qualité de cette donnée n'a pas fait l'objet d'une attention particulière. Pour assurer des taux de réponse satisfaisants, des relances sont assurées par téléphone, relances effectuées par des vacataires embauchés spécialement pour cette tâche. Le nombre de vacataires varie selon la taille des établissements mais peut atteindre en 2013, quinze personnes pendant quatre mois à raison de trois heures par jour pour une université d'environ 50 000 étudiants dont plus de 20 000 sont concernés par ces enquêtes. Les taux de réponse par enquête mail sont environ de 25 à 35 % et le taux final (après relance) avoisine les 85 % en moyenne. Certes, seule l'adresse mail lorsqu'elle est erronée, n'est

pas la raison unique qui motive la relance mais elle constitue un facteur important. Imaginons alors le ratio de coût que représente la mauvaise qualité des données si les bases ne contiennent pas les bonnes adresses mail ni les bons numéros de téléphone alors qu'on embauche des personnels pour faire de la relance sur ces mauvaises données de référence.

Analyser les taux de réponse permet d'être dans une logique de performance qui exprime que la qualité de notre analyse est liée au taux de réponses que l'on a réussi à obtenir et au nombre d'anciens étudiants que nous avons réussi à recontacter. Malheureusement nous oublions souvent que la base de ce travail réside sur l'excellence de la qualité des données et des métadonnées des bases de données.

Pour assurer la qualité de celles-ci, il est nécessaire de remonter au niveau le plus fin des organisations, c'est-à-dire les personnels des services scolarité (en central et dans les composantes) qui inscrivent un étudiant dans la base de données. Est-ce que ces personnes ont inscrit la bonne adresse, le bon numéro de téléphone. Sur le même principe se pose également la question de la mise à jour de ces informations avant que l'étudiant ne quitte l'établissement pour un stage ou une fin de scolarité. Par ailleurs, on constate une corrélation entre la taille de l'établissement et la facilité de mise à jour de ces données, plus l'établissement est petit, plus la mise à jour des données est effectuée. Plus l'établissement est grand, plus le nombre d'étudiants impactés est important et plus il est difficile d'assurer la mise à jour des informations et de ce fait garantir une bonne qualité des informations.

En effet, les dispositifs qualité de maîtrise et de recueil de l'information nécessitent la mise en place d'une démarche d'amélioration de la qualité des données basée sur les principes même de l'amélioration continue induit dans toute démarche qualité, notamment la responsabilisation et la formation des acteurs de terrain.

Dernier exemple lié aux ressources humaines ou aux bases de données de gestion financière, lors de la fusion d'établissements, on se retrouve avec des bases de données qui ont toutes été paramétrées et renseignées selon des règles différentes. L'ensemble des établissements ou composantes qui fusionnent doivent travailler sur une base de données unique. On constate alors une qualité de l'information très variable et la mise en place des nouveaux dispositifs comporte certes des coûts financiers mais également des coûts humains, des coûts liés au stress au travail, des coûts de gestion de dispositifs ou encore des coûts liés à l'accompagnement des acteurs pour gérer le changement. Ces investissements vont avoir un impact et c'est alors une vraie problématique de pilotage et de gouvernance concernant la qualité de l'information qui se base sur trois dimensions que sont les coûts, la qualité de l'information et les dispositifs d'amélioration à mettre en œuvre pour constituer le nouveau système d'informations.

Ces quelques exemples issus du monde de l'entreprise ou de notre quotidien, n'en sont pas moins l'illustration d'une réalité journalière où toutes les actions engagées dans nos entreprises, nos laboratoires de

recherches, reposent sur des données et où la plupart de ces données que nous utilisons peuvent être remise en questions quant à leur niveau de qualité.

La question reste à savoir qu'elle est l'impact pour la recherche et l'enseignement supérieur et que peut-on mettre en place pour combler ce déficit de qualité sur les données.

N'oublions pas l'adage bien connu dans le monde des systèmes d'information : « GIGO : Garbage In ; Garbage Out »⁹

2. Enjeux pour la recherche et l'enseignement supérieur

Nous pouvons distinguer quatre impacts majeurs de la qualité des données scientifiques. Ces quatre impacts sont à croiser en fonction du métier exercé

En premier lieu, il conviendra de réfléchir aux **enjeux scientifiques** liés à la qualité des données. Le chercheur a avant tout besoin d'avoir confiance dans la qualité et la pertinence des données qu'il a à analyser pour pouvoir effectuer ses recherches dans des conditions optimum. Nous verrons également que des réflexions visant le contrôle du processus de l'information sont d'ores et déjà engagées.

Dans un second temps, nous pouvons réfléchir aux **enjeux économiques**. On considèrera alors la notion de performance, cette notion fait désormais sens après avoir

⁹ Littéralement : poubelle en entrée poubelle en sortie, si une donnée est mauvaise au départ, tous les traitements effectués sur cette dernière le seront également.

utilisé celles d'efficience et d'efficacité au cours des 10 années passées. Cette notion de performance intègre une dimension économique très importante ou on reprendra une analyse de coût global des projets dans une dimension économique c'est-à-dire « quel revenu on a pour une thématique donnée ? ». Par exemple, se pose aujourd'hui la question pour un chercheur, de l'intégration de ce coût dans une vision plus macro-économique au niveau de l'économie d'un pays. Il faudra alors être très vigilant à la qualité des données utilisées car il s'agit de les mettre en perspective d'autres enjeux que ces enjeux purement économiques. Ces démarches qualifiées permettent également d'élargir le champ de reconnaissance, notamment par des organismes certificateurs et/ou accréditeurs. Cette reconnaissance permet d'assurer une certaine confiance aux organismes de financement, aux commanditaires, aux donneurs d'ordres, aux bailleurs de fonds, à la communauté scientifique. La qualité de l'information peut devenir un critère important de sélection pour bénéficier d'un financement national, européen ou international. Si aujourd'hui l'évocation d'une norme dans un appel d'offre n'est qu'évoqué, il y a fort à parier que demain ce sera un critère déterminant dans le choix du laboratoire partenaire.

Nous pouvons également évoquer les **enjeux sociétaux et environnementaux**. Dans une société de plus en plus sceptique, la responsabilité du chercheur et des laboratoires par rapport à leurs recherches vis-à-vis de la société, apparaît aujourd'hui normale. La qualité de l'information vise à

garantir un niveau de confiance vis-à-vis des générations actuelles et futures.

Ces enjeux sont à croiser en fonction de l'impact que peut avoir la mise en œuvre d'un référentiel sur la qualité des données et sur les personnels de recherche. Il est clair que pour l'opérateur de saisie qui n'utilise pas la donnée, la saisie de la donnée n'est qu'une tâche qu'il réalise. Le fait de saisir la bonne donnée n'a que peu d'impact sur son travail et il n'en perçoit pas l'impact. Il est important de sensibiliser les personnels sur les objectifs et l'utilisation qui est faite des données dont ils assurent la saisie. La compréhension et la responsabilisation renforcent la qualité des données et des métadonnées associées. Si pour l'opérateur, la mise en œuvre d'un référentiel « qualité de l'information » n'aura pour effet qu'une meilleure gestion des données internes, pour l'ingénieur, producteur et praticiens des données, l'impact sera plus important notamment par la mise à disposition de méthodes pour mesurer et décrire la qualité des données. C'est ainsi que nous voyons aujourd'hui la mise en place d'outils d'analyses croisées, de comparabilité de fiabilité, de contrôle de fiabilité de l'information produite et de vérification de la qualité de l'information. Bien évidemment ces contrôles sont intimement liés à une saisie fiable des données. Enfin, pour le chercheur, la confiance dans ses corpus de données aura un impact très fort sur le déroulement de sa recherche ainsi que sur les résultats qui vont en découler. La qualité des données est essentielle à son travail de production de données tout comme le sera la qualité des données pour le pilotage d'un établissement. Les décisions prises et basées

sur des données non fiables restent un risque élevé dans les établissements.

3. Quelques définitions

Avant d'aller plus loin dans l'exposé des facteurs définissant l'information ou dans les dimensions qui peuvent définir la qualité de l'information, il convient de définir les termes de *donnée*, *d'information*, ainsi que les termes qui gravitent autour de ce terme.

Tout d'abord, la qualité des données peut être définie comme un terme générique définissant à la fois les caractéristiques des données mais aussi l'ensemble des processus qui permettent de garantir ces caractéristiques.

Mais qu'entend-on par donnée ?

Tout commence par le *caractère*, signe élémentaire permettant de constituer des données.

La *donnée* est une description élémentaire, souvent codée, d'une chose, d'une transaction, d'un événement, etc. Les données peuvent être conservées et classées sous différentes formes : papier, numérique, alphabétique, images, sons, etc.

L'information : comme le définit Émile Littré dans le *Dictionnaire de la langue française* est au sens étymologique, ce qui donne une *forme* à l'esprit. Elle vient du verbe Latin *informare*, qui signifie "donner forme à" ou "se former une idée de". L'information désigne à la fois le message à communiquer et les symboles utilisés pour l'écrire.

La *connaissance* : c'est l'information intellectualisée, la représentation d'une perception par le mental qui la transforme en donnée.

On a donc un flux de transformation entre caractères, données, informations et connaissances

Au niveau des systèmes d'information, la notion de référentiel est essentielle. Le référentiel va permettre aux bases de données de se parler ensemble, il permet au sein d'un périmètre établi d'appeler toujours de manière identique la même donnée (un laboratoire, une composante formation, un niveau d'étude,...). Cette notion d'unicité d'information est la base même de la qualité des données qui permet de produire de la connaissance en minimisant les coûts de production de ces données.

On définira par qualité de donnée, à la fois les caractéristiques des données complètes, fiables, pertinentes et à jour mais également l'ensemble des processus qui permette de garantir ces caractéristiques. Au sein de nos dispositifs, il est bien évidemment essentiel qu'il y ait une compréhension globale par l'ensemble des populations impactés de l'importance de la qualité des données tant au niveau de l'établissement que d'une composante de recherche par exemple. Cette compréhension passe par une explication sur le rôle et l'importance de la donnée pour le processus global de pilotage. Tant que l'opérateur de saisie n'a pas compris l'enjeu de la qualité des données, il y a peu de chance que ce dernier que la pérennité de la qualité de saisie soit assurée. Après avoir pris en compte ce facteur humain qu'est le sens et le respect des

consignes, il convient de mettre en œuvre des protocoles de saisie des données pour assurer le maintien du dispositif.

Une fois ce cadre posé, nous pouvons étudier les quatre facteurs qui permettent de qualifier une information.

4. Quatre facteurs qui qualifient l'information¹⁰

Une fois les termes de caractères, données et connaissances définies, nous pouvons expliciter ce qui qualifie l'information.

Comme nous l'avons dit précédemment, la qualité des données repose à la fois sur des caractéristiques (complètes, fiables, pertinentes et à jour) mais aussi sur l'ensemble des processus qui permet de garantir ces caractéristiques. Le but est d'obtenir des données non doublonnées, sans omissions, exactes, sans variations superflues et conformes à la structure définie.

Comme dans toute démarche qualité, l'objectif de la démarche demeure la satisfaction du besoin du client. Les données sont dites de qualité si elles satisfont aux exigences de leurs utilisations. En d'autres termes, la qualité des données dépend autant de leur utilisation que de leur état. Pour certain une donnée de qualité ne comportera ni fautes d'orthographe, ni erreurs grammaticales, pour d'autre, ces mêmes erreurs permettront de qualifier l'exactitude de la restitution d'un entretien.

Pour satisfaire à l'utilisation prévue, les données doivent être exactes, opportunes, pertinentes, complètes, compréhensibles et dignes de confiance. Nous pouvons illustrer ce propos en déterminant 4 facteurs définissant cette information.

4.1 Accessibilité.

Le rôle des systèmes d'informations est prépondérant dans la gestion de l'information. Dit comme cela, ça paraît être une évidence et pourtant les rôles et les responsabilités inférant aux SI sont souvent mal considérés.

Afin de garantir des données de qualité, il faut veiller à pouvoir garantir d'une part la **disponibilité des systèmes**, c'est-à-dire garantir que les éléments considérés sont accessibles au moment voulu, d'autre part, garantir la **disponibilité des transactions**, c'est-à-dire garantir la disponibilité des connexions et des flux entre les différents systèmes permettant l'exécution de requêtes entre eux. Enfin, garantir la mise en place des **privilèges et droits d'accès** permettant aux personnes autorisées et uniquement à ces personnes une possibilité d'accès à l'information recherchée.

Si effectivement l'accessibilité est capitale, l'utilité n'en est pas moins un facteur indispensable à l'usage de l'information

4.2 Utilité

Comme nous l'avons vu en introduction, la quantité d'informations disponibles aujourd'hui est déjà colossale et croît de manière exponentielle. Une information de qualité est une information qui se veut utile, à savoir tout d'abord issue d'une donnée **non-dupliqué**. Les données sont gérées au sein de plusieurs bases de données, par

¹⁰ Livre blanc JEMM research – Des données de qualité - *Informatica*

plusieurs systèmes d'informations sous des identifiants différents, et donc sa vue n'est pas unifiée. L'**usage** des données est aussi important pour définir l'information dont on aura besoin. En effet cet usage permettra de prévoir quels supports et quels formats utiliser.

On entend par **opportunité** le fait que les données sont à jour au moment de leur utilisation. Cette propriété nécessite de travailler de façon permanente sur l'actualisation des sources, l'actualisation des cibles ainsi que sur la volatilité des éléments qui composent l'information.

Troisième facteur qui permet de définir l'information, la crédibilité.

4.3 Crédibilité

Afin d'être qualifié de crédible, les données doivent être **standardisées**, c'est-à-dire permettre aux données de rentrer dans un cadre de codification et de vérifier que les valeurs sont correctes par rapport à un intervalle de répartition ou par rapport à un domaine. L'objectif est d'éviter qu'une information comme un nom de laboratoire soit codée dans un système d'information *UMR4130* dans un autre système *A4130*, dans un troisième *LBI* et dans un quatrième *laboratoire de biologie intracellulaire*. La standardisation soit par modification de la métadonnée soit par la mise en œuvre d'EDI¹¹ a pour vocation de garantir l'unicité

¹¹ *Electronic Data Interchange* ou, en français, *Echange de Données Informatisées*.

L'EDI peut être défini comme l'échange, d'ordinateur à ordinateur, de données concernant des transactions en utilisant des réseaux et des formats normalisés.

Les informations issues du système informatique de l'émetteur transitent par l'intermédiaire de réseaux vers le système informatique du partenaire pour y être intégrées automatiquement.

de l'information. Cette propriété doit être en adéquation avec le besoin du client.

Autre fondamental, l'information doit être exacte. L'**exactitude** garantit que les données représentent la réalité, ou sont vérifiables à partir d'une source externe comme par exemple le code postal qui doit correspondre à la localité.

Les données doivent être **cohérentes**. La cohérence est la capacité pour un système à refléter sur la copie d'une donnée les modifications intervenues sur d'autres copies de cette donnée. Concrètement, si une donnée *d* est écrite sur un système *A* puis dupliqué sur un système *B* et *C*, toute modification de *d* sur *A*, *B* ou *C* fera l'objet d'une mise à jour sur les deux autres systèmes. Aujourd'hui la vérification de cohérence fait partie des fondamentaux de la gestion d'un système d'information.

Enfin, toutes les données nécessaires au besoin du client sont à sa disposition, c'est-à-dire l'ensemble des données et des métadonnées¹² toutes complétées. On parlera alors d'**intégralité des données**,

¹² *Métadonnée* signifie littéralement « des données sur des données ».

Une métadonnée est une donnée qui a pour but de décrire une autre donnée. Les métadonnées peuvent être des informations complémentaires, nécessaires à la compréhension d'une autre information ou dans le but de permettre une utilisation pertinente. Par exemple, les métadonnées d'une page web peuvent informer sur le créateur de la page, de la date de création, de la date de publication, son contenu, etc.

Des métadonnées de qualité facilitent la consultation d'informations. Elles permettent d'améliorer la pertinence des résultats affichés lorsqu'un utilisateur effectue une requête.

<http://www.graphic-evolution.fr/definition/metadonnees-64.html>

Le terme de métadonnées est utilisé pour définir l'ensemble des informations techniques et descriptives ajoutées aux documents pour mieux les qualifier.

<http://www2.cndp.fr/standards/metadonnees/general.htm>

4.4 Interprétabilité

Une donnée doit être stockée et affichée dans un format sans ambiguïté et cohérent. On peut montrer comme exemple le problème d'Interprétabilité soulevé par le format de date entre Paris et New York. Une date affiché sur un écran français 12/11/2010 (jour / mois / année) doit être affiché chez son homologue américain 11/12/2010 (mois / jour / année). Nous l'avons vu avec l'exemple d'airbus. Cette notion d'Interprétabilité peut être à l'origine de nombreux problèmes liés aux traitements des données.

Les contrôles syntaxiques permettent de vérifier que les données saisies dans une zone prévue à cet effet respectent une syntaxe prédéfinie : la syntaxe porte sur le format des données saisies, l'organisation des éléments d'un champ, les séparateurs entre éléments, etc. Par exemple, on pourra vérifier qu'un champ DATE est bien de la forme AAAA, ou que les éléments d'un champ multivalué¹³ sont séparés par un « ; », etc.

Pour garantir l'interprétabilité des données, il faut tout d'abord garantir la **syntaxe** de ces données, c'est-à-dire l'ensemble des règles qui concourent à l'écriture des données. La norme ISO2382¹⁴ à d'ailleurs pour objet de faciliter les échanges internationaux dans les systèmes de traitement de l'information. A cet effet, elle présente un ensemble bilingue de termes et de définitions ayant trait à des

notions choisies dans ce domaine, et définit les relations pouvant exister entre les différentes notions. Les définitions ont été établies de manière à éviter les particularismes propres à une langue donnée, en vue de faciliter leur transposition dans les langues autres que celles ayant servi à la rédaction initiale.

Une fois la syntaxe normalisée, on se concentrera sur la notion de sémantique. En effet, la **sémantique** joue un rôle essentiel. On s'intéressera alors au sens de la donnée en complémentarité à la syntaxe. On peut dire qu'il y a la même différence entre syntaxe et sémantique qu'entre fond et forme. Le décryptage sémantique des données permet une optimisation et une efficacité accrue des recherches d'informations opérées par un utilisateur. La valeur ajoutée repose sur un mode de requête qui écarte les informations parasites (le bruit documentaire) et réduit simultanément les silences documentaires (les informations pertinentes existantes mais non rapportées). Les contrôles de structure permettent de vérifier que le document produit respecte la structuration de l'information définie dans les normes de description (présence d'un champ défini comme obligatoire, non redondance de champs, ordre des champs, ...).

Autre élément essentiel à l'interprétabilité des données, le **contrôle des versions** d'une information, à la fois en tant que tel mais aussi dans sa diffusion. Les contrôles d'unicité permettent de s'assurer qu'une valeur permettant d'identifier de manière non équivoque un élément d'un corpus n'apparaît qu'une seule fois dans ce corpus (identifiant unique d'une notice, d'un dossier

¹³ Partie d'un enregistrement composée de valeurs multiples issues d'un même domaine.

¹⁴

http://www.iso.org/iso/fr/iso_catalogue/catalogue_tc/catalogue_detail.htm?csnumber=7238

ou d'une image, par exemple). L'identifiant doit en outre répondre à une règle de nommage, précisée dans les *Systèmes descriptifs* correspondants.

Les nouvelles technologies de l'information et de la communication comme le cloud computing¹⁵ ainsi que tous les outils de connexion à l'information en mode nomade permettent aujourd'hui d'accéder à l'information en mode partagé, à distance en conservant la notion d'unicité de l'information.

Il en est de même pour la notion d'**alias** qui permettra de ne pas dupliquer l'information, mais de créer des liens entre les fichiers, et ceux afin de garantir l'unicité de l'information.

Enfin, garantir l'interprétabilité des données consiste également à en garantir l'**origine**.

Les données doivent avoir la qualité nécessaire pour supporter le type d'utilisation pour laquelle elles ont été générées. En d'autres termes, la demande de qualité est aussi importante sur les données nécessaires à l'évaluation d'un risque que sur celles utilisées pour une opération d'enquêtes de masses.

Ces quatre facteurs posent les bases essentielles de ce que peut être l'information. Il faut maintenant se demander de quelles notions a-t-ont besoin

pour élargir notre propos à la qualité de l'information et comment l'appliquer à une organisation de production de données scientifiques.

L'établissement de ces critères qualifiant les données permettent d'établir six dimensions qui définissent la qualité des données.

5. Six dimensions définissant la qualité des données

Mentionnons tout d'abord la pertinence des données.

5.1 Pertinence

On entendra par pertinence des données, la satisfaction des besoins clients. Pour être qualifiées de pertinentes, les données doivent **éclairer les utilisateurs** sur les questions les plus importantes à leurs yeux. Cette pertinence a un **caractère subjectif**, car elle dépend avant tout des divers besoins utilisateurs et des réponses que l'on pourra apporter compte tenu des contraintes en matière de ressources.

Pour assurer cette pertinence, il faut s'appuyer sur quatre processus : un *processus de liaison* avec les clients, un *processus d'examen* des programmes de recherche, un *processus d'établissement des priorités* et enfin intégrer le *suivi de la performance*.

Suivi des besoins

L'organisme de recherche doit se doter d'un ensemble de mécanismes qui lui permette de rester au courant des besoins en matière d'informations actuels et futurs de ses principaux utilisateurs. Pour ce faire, on pourra par exemple :

¹⁵ Le cloud computing est un concept de déportation sur des serveurs distants des traitements informatiques traditionnellement localisés sur le poste utilisateur

Le concept d'informatique dans le nuage est comparable à celui de la distribution de l'énergie électrique. La puissance de calcul et de stockage de l'information est proposée à la consommation par des compagnies spécialisées. De ce fait, les entreprises n'ont plus besoin de serveurs propres, mais confient cette ressource à une entreprise qui leur garantit une puissance de calcul et de stockage à la demande.

- participer à la création et / ou participer à des comités consultatifs composés de spécialistes des principaux secteurs étudiés
- participer à la création d'observatoires
- participer à des rencontres avec les principales associations industrielles
- travailler sur les rétroactions formulées par les utilisateurs et/ou tirer des demandes de renseignements.

Ces dispositifs ont pour but de cerner les lacunes liées à la qualité de l'information (l'information nécessaire aux utilisateurs qui n'est pas disponible ou pas suffisamment fiable).

Examen du programme

Le suivi des besoins tel qu'explicité ci-dessus aura pour effet d'obtenir un retour des utilisateurs sur les programmes de recherche ainsi que sur les besoins pour les programmes futurs. Il faut cependant également examiner les programmes de manière explicite et sur une base régulière pour évaluer si ceux-ci répondent aux besoins des clients, non seulement afin de valider la pertinence des sujets traités, mais aussi pour ce qui concerne l'exactitude et la rapidité de diffusion des données produites.

Pour évaluer le programme, on va donc rassembler les données issues de la phase précédente de suivi des besoins que l'on complètera par le rapport d'un expert indépendant par exemple. Les programmes peuvent également recueillir et évaluer périodiquement les commentaires qu'ils reçoivent et rédiger un rapport énonçant les changements possibles afin de répondre au mieux aux besoins des clients.

Etablissement des priorités

Il y a fort à penser que les demandes seront souvent pour ne pas dire toujours supérieures aux ressources disponibles. Il va donc falloir comparer en faisant preuve de jugement, les divers besoins des différents groupes d'utilisateurs. Il faut cerner également les possibilités d'obtentions et d'extensions des ressources initiales qui permettraient de répondre pleinement à l'ensemble de la demande client.

Suivi de la performance

Nous pouvons citer deux principaux types de suivi de la performance en matière de pertinence.

Premièrement, un suivi par des descriptions des mécanismes particuliers utilisés qui sont appuyés aux mieux sur des mesures, à défaut sur des exemples de leur incidence qui montrent que les processus préalablement décrits sont en place.

Deuxièmement, on peut apporter la preuve de la pertinence des données par l'usage qui en est fait selon les résultats d'évaluations remplis par les clients eux-mêmes.

La seconde dimension que nous exposerons ici est l'exactitude, à savoir dans quelle mesure l'information décrit bien le phénomène qu'elle doit mesurer.

5.2 Exactitude

Habituellement, elle se définit par rapport aux estimations statistiques erronées et est traditionnellement décomposée en composante de biais¹⁶ et de variance¹⁷. On peut également la définir par rapport aux

¹⁶ Erreurs systématiques

¹⁷ Erreurs aléatoires

sources d'erreurs principales susceptibles de mener à des données inexactes¹⁸

Il faudra tout particulièrement s'assurer de l'exactitude des données aux cours des trois principales étapes du projet : la conception, la mise en œuvre et l'évaluation.

Conception

Il est assez rare que les programmes de recherche définissent des cibles en ce qui concerne l'exactitude des données. Ils indiquent plus souvent des quantités à estimer, des zones géographiques à étudier, des niveaux de détails ou de précisions voulus, des estimations, mais l'exactitude des données demeure au mieux que très vague tout comme la définition des niveaux d'erreurs non imputable à l'échantillonnage. Lors de la définition des programmes de recherche, on essaie souvent de trouver le juste milieu entre exactitude et rapidité de diffusion en fonction des contraintes de ressources. Il convient alors d'envisager dans le processus de conception les diverses options en matière de niveaux d'exactitude.

Mise en œuvre

Il faut tout d'abord des renseignements qui permettent de surveiller et de corriger en temps réel les problèmes qui peuvent survenir lors de l'acquisition des informations. L'utilisation d'un système d'information permettra au chercheur d'obtenir au bon moment les renseignements dont il a besoin afin d'adapter sa recherche ou pour corriger les situations difficiles.

Il faudra également pouvoir accéder à l'information après coup si le modèle a été exécuté tel que prévu et que certains aspects ont posés problèmes. Les leçons à tirer pour les programmes suivant permettront d'entrer dans le cercle vertueux de l'amélioration continue.

Bien entendu, l'information se rapportant directement à l'exactitude ne constitue qu'une partie infime des données que requière la gestion opérationnelle. Les renseignements liés aux choix et aux coûts du moment des opérations doivent tout autant être pris en considération pour les modèles ultérieurs.

Evaluation de l'exactitude

Cette troisième phase, probablement la plus importante du processus, est l'évaluation de l'exactitude de l'information.

Nous pourrions décomposer cette notion d'exactitude en trois composantes.

La première composante réside dans l'évaluation du **biais**. D'une façon générale, le terme « *biais* » désigne un écart systématique (non aléatoire) entre une grandeur et la prédiction de cette grandeur. C'est donc une valeur **d'erreur systématique**.

La seconde composante est la composante de « *variance* ». Habituellement, la **variance** d'erreur est définie en statistique comme la part de la variance observée (ou variance totale) qui est imputable aux fluctuations aléatoires de l'échantillonnage. Ces fluctuations concernent notamment le choix (aléatoire précisément) des composantes du dispositif de mesure utilisé, ainsi que des conditions dans lesquelles la mesure a lieu,

¹⁸ Erreur de couverture, d'échantillonnage, de non-réponse et de réponses

choix des items, des classes, des moments d'observation, des correcteurs. Soit une valeur d'**erreur aléatoire**.

La troisième composante est constituée d'**erreurs principales** susceptibles de mener à des données inexactes. On regardera alors les erreurs issues des incertitudes de couverture, d'échantillonnage, de réponses, de non réponses etc.

Les efforts à investir dans la mesure de l'exactitude sont avant tout une décision de gestion qui doit être prise compte tenu des compromis habituels lors de la conception de l'échantillon et des méthodes mises en œuvre.

5.3 Rapidité de diffusion de l'information

On entend par rapidité de diffusion de l'information, le délai entre la fin du recueil de l'information et la date à laquelle les données sont disponibles.

Comme nous l'avons vu, il existe un lien fort entre la rapidité à laquelle l'information doit être diffusée et la pertinence de l'information. En effet, cette dernière influe fortement sur la célérité avec laquelle l'information va être diffusée.

On peut également se poser la question de la durée pendant laquelle l'information est utile. Nous pouvons dire que la réponse à cette question dépend fortement du phénomène observé.

Reste à trouver le bon compromis. En effet, est-il préférable d'avoir des données plus exactes plus tard ou moins exactes plus tôt ? On admettra donc que la rapidité de diffusion n'est pas un objectif intrinsèque.

Les principales dates de rendu de production doivent être annoncées longtemps à l'avance, laissant ainsi aux chercheurs le temps de planifier leurs recherches ainsi que la publication de leurs résultats.

5.4 Accessibilité

Répondre à la question d'accessibilité revient à travailler sur quatre points essentiels.

Tout d'abord, avec quelle facilité peut-on se procurer les données ? Avec quelle facilité constate-t-on que l'information existe ? Quelle est le caractère approprié de la présentation de l'information ? Ou encore, quel est le coût d'accès à l'information.

Il faut veiller qu'aucune erreur ne se glisse en transférant l'information d'un programme de recherche vers les utilisateurs. A cette étape du processus, il conviendra de vérifier le transfert des bonnes informations dans les bases de données, de publier des tableaux dans la bonne version. Comme de telles erreurs sont susceptibles d'avoir lieu à l'étape de la livraison, les risques existent et les systèmes qualité réduisent la possibilité de telles erreurs.

Dans le même esprit, le retour des utilisateurs est essentiel. La rétroaction des utilisateurs peut provenir de statistiques automatisées sur l'utilisation des diverses composantes des systèmes, d'enquêtes sur le degré de satisfaction des utilisateurs quand à des produits, services ou systèmes de livraison ou il peut s'agir de commentaires de suggestions, de plaintes ou de témoignages d'approbations librement exprimés par les utilisateurs.

5.5 Interprétation

On entend par possibilité d'interprétation, la disponibilité de **renseignements supplémentaires** ou de **métadonnées** nécessaires à l'exploitation des données.

Une fois acquise cette notion de métadonnées, on s'intéressera également aux méthodes de collectes de données ainsi qu'aux indicateurs de l'exactitude des données.

L'information nécessaire à la compréhension des données peut se grouper en trois catégories. Tout d'abord les **concepts et les classifications** à la base des données c'est-à-dire quels sont les éléments mesurés afin que l'utilisateur puisse juger de la pertinence par rapport à son besoin, ensuite les **méthodes de collectes et de rassemblement** des données c'est-à-dire quelles méthodes j'utilise pour pouvoir juger de la pertinence des outils utilisés, enfin, les mesures de l'exactitude des données, quel est le résultat obtenu pour garantir la confiance dans les résultats. De ce fait, la description technique devient un véritable indicateur de l'exactitude des informations.

Lorsque l'on parle interprétation, on se doit également d'aborder la forme des rapports. Prétendre que l'information éclairant les données doit être compréhensible relève d'une évidence. Néanmoins, le producteur d'information doit s'efforcer de communiquer dans la langue de l'utilisateur et non pas dans un jargon interne, moyennant quoi l'utilisateur aura beaucoup de mal à interpréter les données.

5.6 Cohérence

Dernière branche pouvant qualifier la qualité de l'information, la cohérence des données

concerne la liaison avec d'autre renseignement statistiques dans un cadre analytique au fil du temps.

Cette propriété se rapporte à la cohérence entre les mêmes éléments de données se rapportant à des unités de temps différentes, à la cohérence entre divers éléments de données se rapportant à la même unité de temps ou à la cohérence internationale.

Concrètement, une terminologie est la même pour tous les programmes de recherche. Par exemple, le « niveau d'instruction » aura la même signification dans une enquête sur le recensement de la population que dans une enquête sur la scolarité. En conséquence, les quantités exprimées ont des liens connus entre elles.

On arrive à cette notion de cohérence par l'adoption de cadres nationaux voir internationaux comme par exemple les normes IFRS¹⁹ qui rendent cohérent les systèmes de comptabilité mondiaux. Tous les systèmes de classifications types s'appliquant à tous types de variables permettent de garantir cette cohérence des données. La comparaison sur le plan international se fait en vérifiant si les normes adoptées sont conformes aux normes internationales quand il y en a.

Il faut également s'assurer que le processus de mesure n'entraîne pas d'incohérence entre les sources de données même quand

¹⁹ Normes internationales d'information financière, plus connues sous leur nom anglais d'International Financial Reporting Standards
Normes comptables internationales, élaborées par le Bureau des standards comptables internationaux

les quantités à mesurer sont garanties de façons uniformes.

Pour évaluer dans quelle mesure les données sont cohérentes, on peut définir trois grands ensembles de mesures. Il s'agit en premier lieu de valider l'existence et de mesurer le niveau d'utilisation des cadres de références de variables et des systèmes de classification types. Dans une deuxième phase, de valider l'existence et de mesurer le niveau d'utilisation des méthodes et des outils pour la conception et la mise en œuvre des programmes. Et enfin de mettre en place des indicateurs pour mesurer la fréquence et l'incohérence des données publiées.

6. Indicateurs et mesures

Comme dans toute démarche qualité, la mise en place d'indicateurs est une étape essentielle pour s'engager dans le processus d'amélioration continu.

A partir de ces définitions théoriques, chaque organisme devrait s'approprier les concepts et créer sa propre définition opérationnelle en fonction des objectifs et des priorités des programmes de recherche afin de définir des indicateurs pour chacune des dimensions et vérifier par des mesures régulières leurs évolutions dans le temps. Chacune des dimensions que nous avons exposées peuvent être mesurées, soit de manière objective au travers de suivis automatiques d'indicateurs spécifiques, soit de manière subjective en recueillant la perception des utilisateurs.

Nous pouvons citer quelques indicateurs à titre d'exemple.

L'âge des données est-il conforme au besoin métier ? Se poser cette question permet de s'interroger sur le critère d'**opportunité**. Les indicateurs que nous pouvons alors mettre en œuvre pourraient être *la date de la collecte des données, la date du dernier traitement et/ou encore le contrôle de la version*.

Est-ce que toutes les données nécessaires sont disponibles ? Répondre à cette question permet de mesurer si la donnée est **complète**. Mesurons alors si l'intégralité des données optionnelles est disponible, le nombre de valeurs non renseignés, le nombre de valeurs par défaut par rapport à la moyenne.

Quelles sont les données sources des informations contradictoires ? Permet de mesurer la **cohérence** des données. Nous pouvons alors mesurer par exemple la valeur de la déviation standard²⁰ ou effectuer des vérifications de plausibilité.

Afin de vérifier l'**exactitude** des données, interrogeons-nous afin de savoir en quoi les valeurs représentent la réalité ? On pourra alors mesurer la fréquence des changements de valeurs ou encore le feed-back des clients.

Les données sont-elles compréhensibles par les utilisateurs ? Cette question est un exemple d'interrogation permettant de répondre à la notion d'**interprétabilité des**

²⁰ La déviation standard, permet d'évaluer la dispersion des mesures autour de la valeur moyenne

données. On pourra alors mesurer la valorisation des données utilisateurs

Quels sont les données saisies, stockées, ou affichées dans un format non standard ? On s'intéressera dans ce cas à la mesure de **standardisation**, de **conformité** d'une donnée. On pourra alors mettre en œuvre un *certificat de conformité* par exemple.

Quelles sont les données répétées ? Permettra de s'interroger sur le caractère de **duplication** d'une donnée. Nous pouvons mesurer alors le *nombre d'enregistrement dupliqués*.

Une fois les indicateurs définis, il faut mettre en place un système de mesure qui permet de surveiller leur évolution dans le temps. L'utilisation de ces indicateurs permettra de définir les plans d'actions nécessaires à l'amélioration des différents processus.

7. Qui est concerné ?

La qualité des données concerne tous les personnels de l'organisation. Nous l'avons déjà vu, le chercheur, l'ingénieur, comme l'utilisateur de base ont un intérêt différent, mais un intérêt tout de même à participer à la démarche de qualité des données.

Les *dirigeants*, doivent être les premiers promoteurs de la qualité de l'information. Comme dans toute les démarches qualité, la direction de l'organisme à un rôle prépondérant²¹ dans la réussite d'une démarche qualité.

La direction de l'unité doit avant tout **définir la stratégie** de l'établissement en matière de qualité des données. Elle doit **prendre les décisions** nécessaires au bon fonctionnement du système qualité de l'organisme. Enfin, elle doit assurer le **pilotage** du projet « qualité des données ».

Les *responsables opérationnels* sont avant tout les garants de la qualité de l'information à sa création.

Le responsable opérationnel a pour mission de *gérer au quotidien d'important volume* de données. *Travailler sur les interfaces* entre la direction et le métier. Pour lui, le besoin d'informations se traduit par le besoin d'outils de reporting puissants et d'éléments pertinents et synthétiques.

Les *collaborateurs opérationnels* sont créateurs et utilisateurs des données correctes, ils ont quant-à eux *besoin d'informations fiables* et *besoins d'un accès simple et rapide* à la donnée.

Enfin le *responsable du système d'information*. Il est garant de l'intégrité et de la disponibilité de l'information. Son rôle est de former, conseiller, aider et accompagner les utilisateurs à maintenir les données aux meilleurs niveaux de qualité. Il est également garant de l'infrastructure technique qui va permettre d'acquérir, stocker, modifier, restituer et sécuriser l'information.

8. Gouvernance

Dans le cadre de la démarche qualité des données, l'établissement doit définir son modèle de gouvernance, c'est-à-dire la

²¹ Cf Paragraphe 5 de la norme Iso9001

formalisation de son modèle de pilotage des technologies, des personnes et des processus.

On pourra distinguer deux instances dans le pilotage de ce projet : la direction de l'unité appuyé par les directions métiers et le comité « qualité des données »

8.1 Direction de l'unité et directions métiers

Son rôle comme on a déjà pu le voir réside dans la sponsorship du projet. Comme dans toute démarche qualité cette sponsorship est indispensable. Il faut convaincre la direction générale et les directions métiers de l'impact de la non-qualité des données et surtout démontrer que la qualité est source d'excellence pour le laboratoire. La conduite du changement et toutes ses méthodes seront des outils précieux pour convaincre les directions métiers de l'intérêt de la démarche. Elle s'assurera que la démarche est lancée.

Autre rôle de la direction générale, mettre en place un comité qualité des données, dont l'objectif sera de s'assurer que l'ensemble des projets intègrent la gestion de la qualité des données dans l'ensemble de leurs processus de recherches

8.2 Comité qualité des données

Ce comité est responsable de la qualité des données de l'entité. Ce modèle de pilotage doit comprendre une structure organisationnelle en charge de l'amélioration de la qualité de l'établissement, sous la responsabilité d'un sponsor qui doit avoir une influence sur les directions métiers.

Il définit les objectifs, les priorités. Il s'assure que tous les projets intègrent la qualité des

données dans leurs processus. Il s'assure également de la disponibilité des financements nécessaires au projet.

Ce comité se réunit régulièrement pour assurer le suivi du projet qualité des données et assurer le suivi des actions d'améliorations. Enfin, il décide des nouvelles priorités.

Ce comité est composé d'experts issus de l'ensemble des directions du laboratoire. Ces derniers sont responsables de la définition, de la surveillance des mesures et indicateurs. L'analyste Qualité des Données doit avoir un lien fort avec la direction des systèmes d'informations. Son rôle est de mettre en place les outils de contrôle de la qualité des données, définir les principaux indicateurs et mesures de la qualité des données, justifier des programmes d'amélioration à mettre en œuvre et de mesurer de façon régulière les progrès effectués.

9. Conclusion

La qualité des données peut être aujourd'hui un des enjeux majeurs des laboratoires de recherche et des établissements d'enseignement supérieur. Même si actuellement les appels d'offres de recherche ne mentionnent pas la qualité que comme un plus, et non comme une obligation, il y a fort à parier que dans les années à venir, ces démarches deviennent indispensables pour obtenir des contrats de recherche voir pour faire face à la concurrence des laboratoires étrangers.

La non-qualité des données a un coût important, tant humain que financier. Les enjeux à la fois scientifiques, économiques, sociétaux et environnementaux font que les laboratoires doivent être en mesure aujourd'hui de contrôler la qualité de leurs données.

Pour répondre à ces questions, la mise en place d'un comité de pilotage de la qualité des données veillera à traiter la pertinence, l'exactitude, la rapidité de diffusion ou encore l'accessibilité, l'interprétation ou alors la cohérence des données en liaison forte avec la direction du système d'information, direction garante du stockage, de la sécurisation ou encore de la diffusion de l'information.

L'idée de démarche et de pérennité est essentielle et caractéristique de l'approche qualité. Elle va à l'encontre d'une opération unique et isolée qui ne permet de nettoyer et d'améliorer les données que ponctuellement. Cela signifie que les objectifs, mesures et indicateurs associés doivent être portés par l'ensemble des acteurs concernés et en particulier se traduire par une implication forte de la direction.